Cobalt

State of LLM Security Report 2025

Contents

Executive Summary

The rapid adoption of LLMs is driving unprecedented innovation across industries. An overwhelming 94% of respondents in our survey of 450 security leaders and practitioners have observed a significant increase in the adoption of generative AI (genAI) within their industry over the past 12 months. However, this surge in development is outpacing crucial security considerations, creating an oversight gap.

This report exposes a troubling reality: while threats related to genAI are a top concern for security teams and leaders alike, the current state of security testing and remediation of LLM and AI-powered applications is proving insufficient to address the novel risks these powerful new technologies introduce.

32%

LLM pentests reveal the highest proportion of serious vulnerabilities

32% of all findings in LLM assessments are classified as serious (high or critical risk). This is the highest proportion of serious vulnerabilities found across all asset types tested by Cobalt, since LLM testing began in 2022[1].

21%

A troubling gap exists in remediating serious LLM vulnerabilities

Only 21% of these serious LLM vulnerabilities are actually resolved, the lowest resolution rate among all types of penetration tests conducted.

1/2

Concerns about supply chain security are prevalent

Half (50%) of respondents express a desire for more regular reports and statistics from their software suppliers on how they identify and prevent vulnerabilities—this applies to the AI supply chain as well.

36%

Security teams are struggling to keep pace with genAI demand

Despite widespread adoption, 36% of security leaders and practitioners surveyed admit that the demand for genAI has outpaced their security team's ability to manage its security implications, highlighting a critical readiness gap.

19 days

Rapid fixes for resolved LLM issues show organizations are constrained to fixing the easiest-to-resolve issues

Despite the low overall resolution rate, the mean time to resolve (MTTR) for those serious LLM findings that are fixed is a rapid 19 days—the shortest MTTR across all pentest types. This is due to a couple of factors: organizations tend to prioritize quicker, perhaps simpler fixes; and are only able to address vulnerabilities that are not dependent on third-party model providers.

46%

Data security tops the list of concerns about LLM and AI application vulnerabilities

Survey data reveals high organizational concern for data-centric risks in genAI, although the top findings in pentests of LLM applications show vulnerabilities to attack vectors such as SQL injection. The top concerns include "Sensitive information disclosure" (46%), "Data model poisoning or theft" (42%), "Inaccurate data" (40%), and "Training data leakage" (37%).

1 in 3

Only one-third of respondents are conducting regular testing of their AI deployments

While 72% of respondents identify attacks related to genAI as a top IT risk—the most frequently cited risk in the survey—one-third (33%) are still not conducting regular security assessments, including penetration testing, for their LLM deployments.

76% vs 68%

Perspectives of security leadership and practitioners show different focus on near-term and long-term risks

Security leaders (C-suite and C-1 level) generally express a higher concern for "Generative AI related attacks and threats" (76% vs. 68% for practitioners). Conversely, security practitioners and managers demonstrate a higher concern for more immediate operational risks like inaccurate data (45% vs. 36% for leaders) as an AI-specific threat.

[1] Pentesting data cited in this report consists of AI/LLM testing results from Cobalt customers between 2022 and 2025. The same data set of testing results was analyzed in the Cobalt State of Pentesting Report 2025. Survey data comes from a survey commissioned by Cobalt of 450 security practitioners, managers, and C-suite leaders in early 2025.

AI Revolution and the Expanding Attack Surface

The transformative impact of AI, particularly LLMs, is undeniable, with organizations across nearly every industry rapidly integrating these technologies to drive innovation, enhance productivity, and create new customer experiences.

However, this technological revolution brings with it substantial risks. The integration of LLMs is not merely expanding the existing attack surface; it is fundamentally introducing novel, complex, and often underexplored vulnerabilities. Current security practices are proving inadequate to address the unique challenges posed by these sophisticated AI systems—mature controls were not written for a world

The AI Technology Stack: Mapping the Placement of LLMs

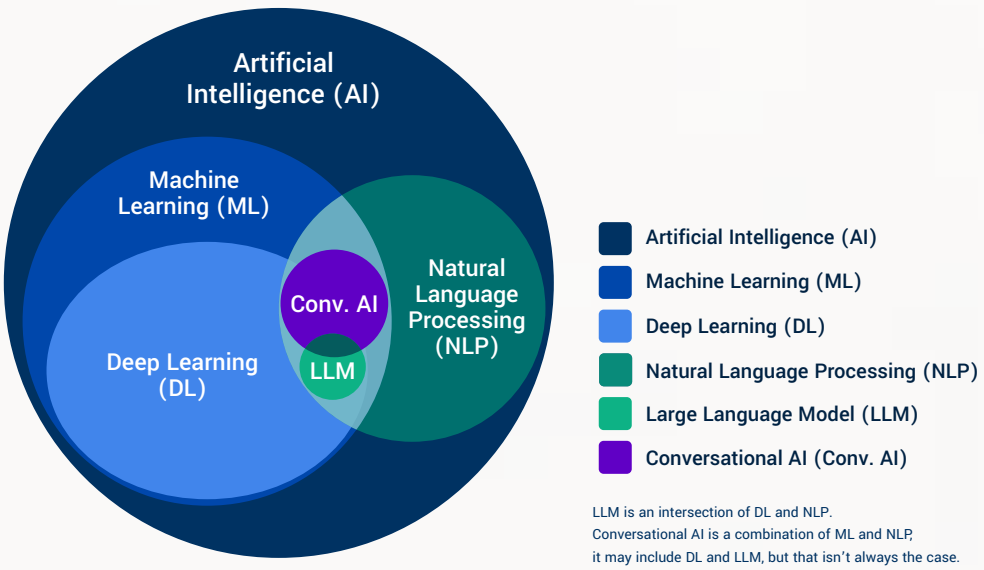

Figure 1. Source: Cobalt

of LLM experiences. Much like the seismic shift security teams experienced as companies rushed to embrace the cloud, the surge in AI adoption has left security oversight behind.

While the foundational concepts of AI are not entirely new—machine learning and natural language processing have been evolving for decades—the current wave of genAI and LLMs presents a paradigm shift. The accessibility, inherent interactivity, and increasingly autonomous nature of these models introduce critical security challenges. These are not theoretical risks; they demand immediate and decisive action, particularly through rigorous, specialized testing designed to uncover and mitigate AI-specific vulnerabilities. Pentesting, among other security techniques, is critical to ensuring AI applications remain secure, ethical, and reliable.

> “These are not theoretical risks; they demand immediate and decisive action, particularly through rigorous, specialized testing designed to uncover and mitigate AI-specific vulnerabilities. Pentesting, among other security techniques, is critical to ensuring AI applications remain secure, ethical, and reliable.”

The Current Landscape of LLM Security Threats

The rapid adoption of genAI and LLMs introduces a new and significant layer of risk to the cybersecurity landscape. Analysis of pentesting data and corroborating survey responses from security professionals and leaders reveals a concerning picture for LLM security.

Presenting the stark reality of LLM vulnerabilities

Pentesting data from the Cobalt State of Pentesting Report 2025 exposes a troubling reality regarding the vulnerabilities found in LLM applications. Nearly one-third (32%) of findings uncovered in LLM pentests are classified as serious vulnerabilities—meaning they are rated as high or critical risk. This is the highest proportion of serious findings among all asset types tested by Cobalt, indicating that LLM deployments currently present a particularly elevated risk profile.

Compounding this concerning finding is the state of remediation. Only 21% of these serious LLM vulnerabilities are actually resolved, the lowest remediation rate among all testing types, a troubling gap in addressing critical risks. While the mean time to resolve (MTTR) for the serious LLM findings that are actually fixed is just 19 days—the shortest MTTR across all pentest types in the Cobalt pentest data—this speed is perhaps misleading. It indicates that organizations may be quicker to address the more straightforward issues, yet a substantial backlog of complex, unaddressed (and serious) vulnerabilities persists.

Industry differences in serious LLM vulnerabilities

The prevalence of serious vulnerabilities found in pentests varies across industries, offering insights into where the most significant risks may lie based on actual testing outcomes. Based on the State of Pentesting Report 2025 (See Figure 2 below), industries with the highest percentages of serious vulnerabilities include Administrative Services (19.5%), transportation (19.0%), hospitality (18.5%), manufacturing (18.1%), and education (17.6%). Conversely, the entertainment industry stands out with the lowest rate (8.0%), followed by financial services (11.2%) and information services (11.3%). These differences may reflect varying levels of security maturity, complexity of the technologies deployed, or regulatory environments across sectors.

Prevalence of serious findings by industry

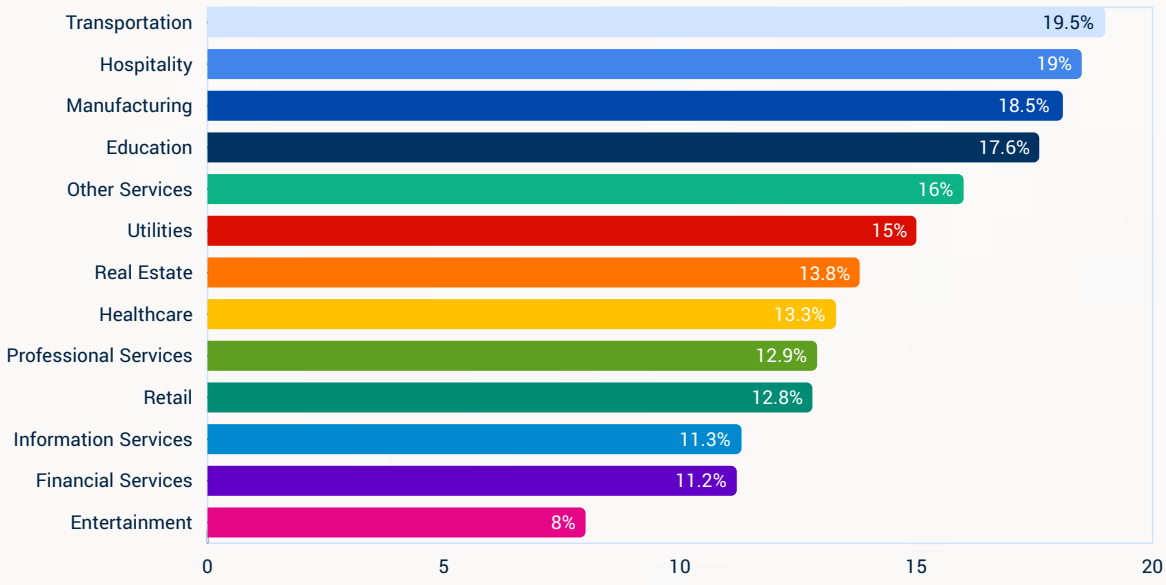

Figure 2. Source: Cobalt pentests

The disconnect: high concern, insufficient action

This significant vulnerability landscape exists despite a high level of awareness regarding genAI threats. “Generative AI related attacks and threats” are cited as the top IT risk by 72% of survey respondents overall. Specific AI-related concerns, such as “Sensitive information disclosure” (46%) and “Data model poisoning or theft” (42%), are also prominent concerns. However, the level of proactive testing and robust security measures still falls short of what is needed.

Only 66% of organizations are conducting regular security assessments, including penetration testing, for genAI-powered products. Therefore, one-third of organizations are not routinely testing these high-risk, rapidly deployed AI applications. This highlights a blind spot that gives some survey respondents hesitation. Nearly half of respondents (48%) believe a “strategic pause” is needed to recalibrate and reinforce defenses against genAI-driven threats, contrasting with the 52% who believe they can effectively defend without one.

GenAI as a threat or a tool: security leaders vs. security practitioners and managers

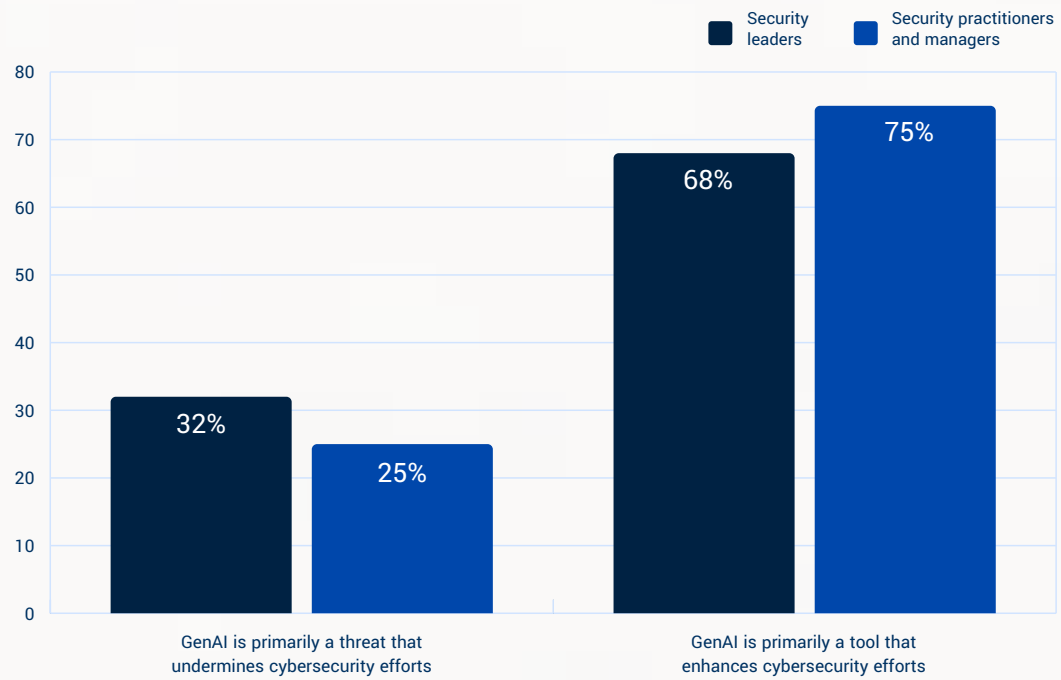

Figure 3. Source: Survey of security leaders, managers, and practitioners.

Furthermore, a higher proportion of security leaders (50%) vs. security practitioners (45%) believe a pause in deployment of genAI is necessary. Additionally, 32% of security leaders believe genAI is "primarily a threat that undermines cybersecurity efforts," as opposed to "genAI is primarily a tool that enhances cybersecurity efforts." That compares to only 25% of security practitioners and managers who say genAI is primarily a threat (Figure 3). This points to a division in approach between leaders and practitioners, and potentially troubling overconfidence in some quarters not fully supported by pentest data.

Perception vs. reality: overconfidence affecting security

As companies are rushing to launch features for a competitive edge, security can be an afterthought. Speed-to-market pressures leave security teams looped in late with little time to act. Yet a majority of security teams believe in their ability to catch up.

The space between ambition and execution can leave organizations exposed. While a majority of survey respondents (74%) express confidence in their adaptation to genAI, this confidence is difficult to reconcile with the evidence of widespread, unresolved serious LLM vulnerabilities, suggesting an overestimation of current capabilities —a theme from the State of Pentesting Report 2025.

Differences in attitudes and testing vary by industry

Analysis of the survey data by industry reveals further nuances in the landscape of LLM security preparedness. These survey insights, when combined with the pentesting data on serious findings by industry presented above, provide a more complete picture of where the critical gaps exist.

Education sector

While 100% of respondents in this sector agree that the demand for genAI has not outpaced their security teams' ability, their rate of conducting regular security assessments and pentesting for genAI products is only 33%, or half the overall average of 66%. This suggests overestimation of their security posture, leaving sensitive educational data and systems vulnerable. However, respondents in education have much higher concern for "Bias in AI decisions" (100% vs. 32% overall) and "Training data leakage" (67% vs. 37% overall). This is particularly concerning given that pentest data shows the education sector has among the highest rates of serious vulnerabilities (17.6%). It could be that security teams in education are more comfortable with experimentation and learning fast in an academic environment.

Bias in AI is a top concern

100%
Education sector

32%
Overall

Financial services

Survey respondents in this sector exhibit a higher concern for "Risks associated with third-party software and tools" (76% vs. 66% overall) and are more likely to report hosting AI models internally (57% vs. 36% overall). This sector also places a very high priority on guidance for understanding security implications when incorporating genAI into customer-facing services and products (79% vs. 45% overall). While pentesting data shows financial services organizations have among the lowest rates of serious vulnerabilities (11.2%) compared to other industries, their self-reported high concern for third-party risks and customer-facing applications suggests that even in sectors with lower vulnerability rates, the stakes remain high, necessitating diligent testing. Relatedly, financial services organizations tend to be more tightly regulated than those in other sectors.

Risks associated with third-party software and tools is a top concern

76%
Financial services

66%
Overall

Manufacturing

Respondents in manufacturing report a heightened concern for traditional threats like "Exploited vulnerabilities" (74% vs. 48% overall) and "Phishing and malware" (58% vs. 41% overall). They also report a greater need for improved testing (61% vs. 50% overall), suggesting they feel they are behind the curve in adopting offensive security practices for new AI risks. This aligns with pentesting data showing manufacturing has a high percentage of serious vulnerabilities (18.1%). The expense of security measures versus the potential financial losses from production downtime presents a significant challenge in manufacturing organizations. Additionally, operational limitations hinder the ability to remediate vulnerabilities efficiently and at scale.

Exploited vulnerabilities is a top concern

74%
Manufacturing

48%
Overall

Security leaders vs. practitioners and managers: divergence of views on genAI

A key discrepancy emerges in the focus of concern reported by security leaders and practitioners. Security leaders generally express a higher concern for genAI related attacks and threats (76% vs. 68% for practitioners). This indicates a strategic focus on broader, potentially more emergent threats, and the evolving landscape of AI-driven attacks that may not be fully manifest today. They are also more likely to consider changing how their team approaches cybersecurity defense strategies due to the potential of genAI-driven attacks (52% vs. 43% for practitioners), in anticipation of evolving threats.

In contrast, security practitioners and managers demonstrate a higher concern for inaccurate data provided by LLMs (45% vs. 36% for leaders) as an AI-specific threat. This suggests practitioners are more attuned to the immediate, operational risks that directly impact their daily work and the current state of LLM functionality. This also correlates with practitioners being slightly more focused on the importance of pentesting on the software supply chain (51% vs. 44% for leaders), underscoring their more direct involvement in existing supply chain security measures, which must now include LLM app components.

Unsurprisingly, these differences suggest that leaders are looking at the horizon for future AI-related challenges and strategic adaptation, while practitioners are grappling with the tangible, present-day implications and operational realities of genAI in their environments.

> Leaders are more considerate of future AI threats and strategic adaptation, while security practitioners are focused on the present-day implications of genAI.

Understanding Key LLM Vulnerabilities

To effectively address the security challenges posed by LLMs, it's crucial to understand the specific types of vulnerabilities that are being identified in real-world pentests. While LLMs introduce novel attack vectors, traditional security weaknesses often remain relevant and can be exploited through these new interfaces.

The following chart (Figure 4) illustrates the most common AI and LLM pentest findings identified by Cobalt in 2024, covering all criticality levels.

Top AI and LLM pentest findings in 2024 (all criticalities)

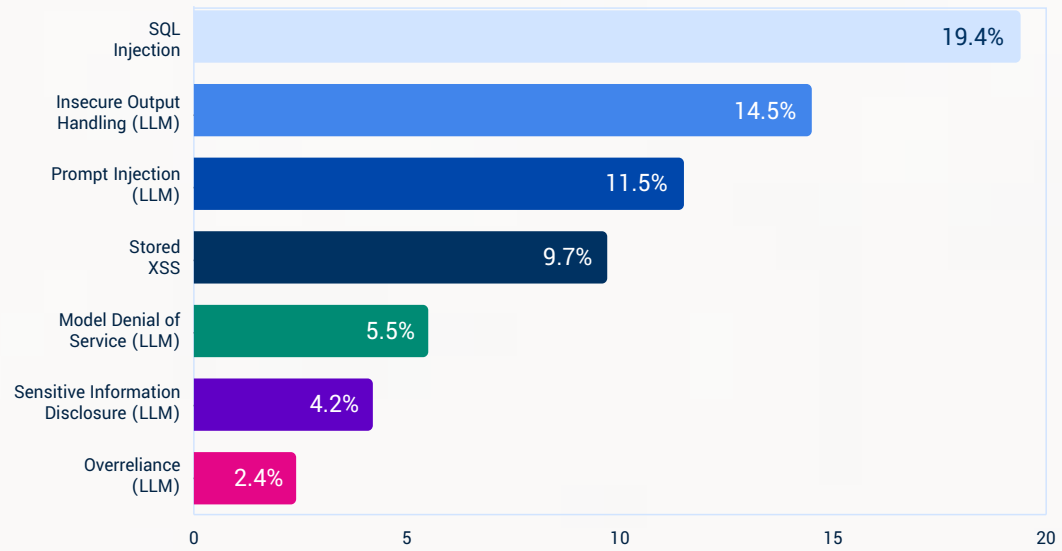

Figure 4. Source: Cobalt pentests

Traditional weaknesses persist

Observing the top findings from LLM pentests, it's noteworthy that some of the most prevalent issues are not unique to LLMs. For example, SQL injection, a classic web application vulnerability, accounts for 19.4% of findings in these AI-focused tests, making it the most common. Another traditional web vulnerability, stored cross-site scripting (stored XSS), also ranks high at 9.7%.

This underscores a critical point: while novel techniques like prompt injection might serve as the vector for an attack on an LLM-powered application, the exploit can often succeed due to underlying, traditional security weaknesses in the application's infrastructure, its web interface, or associated components. Therefore, foundational web application security best practices remain highly relevant when deploying LLM applications. Robust input validation, secure coding practices, and proper configuration of underlying systems are as important as ever.

AI safety vs. security of AI

When discussing vulnerabilities in AI systems, it's helpful to distinguish between AI safety and security of AI (Figure 5). These are related but distinct concepts, both crucial for the responsible development and deployment of AI. Pentesting often uncovers issues relevant to both categories, highlighting the need for a comprehensive assessment approach.

- **Security of AI** focuses on protecting the AI system itself from malicious attacks and unauthorized access, ensuring its confidentiality, integrity, and availability.
- **AI safety** focuses on ensuring the AI system behaves correctly, aligns with human values, avoids generating harmful or biased outputs, and is transparent and controllable. This involves ensuring the AI behaves responsibly and doesn't inadvertently harm people or perpetuate societal biases.

	Security of AI	AI safety
Focus	Protecting AI from cyberthreats	Preventing AI from generating harmful content
Key risks	Model theft, prompt injection, adversarial attacks	Bias, ethical concerns, lack of transparency
Examples	Prompt injection, model denial of service, sensitive information disclosure, supply chain vulnerabilities, model theft, insecure plugin design	Inaccurate data (hallucination), bias in AI decisions, misinformation, insecure output handling (when leading to harmful content)

Comparison of pentest findings vs. survey concerns

When examining the landscape of LLM vulnerabilities, a notable distinction arises between the concerns highlighted by security professionals in our survey and the types of vulnerabilities most frequently uncovered during penetration tests.

Survey data indicates that professionals are predominantly worried about the security and integrity of the data associated with genAI systems. The top concerns cited include sensitive information disclosure (46% of respondents), data model poisoning or theft (42%), inaccurate data (40%), and training data leakage (37%) (Figure 5). These anxieties clearly point towards a strong focus on protecting data assets from exposure, manipulation, or leakage through or by AI systems.

Q. Which of the following are most concerning to your company?

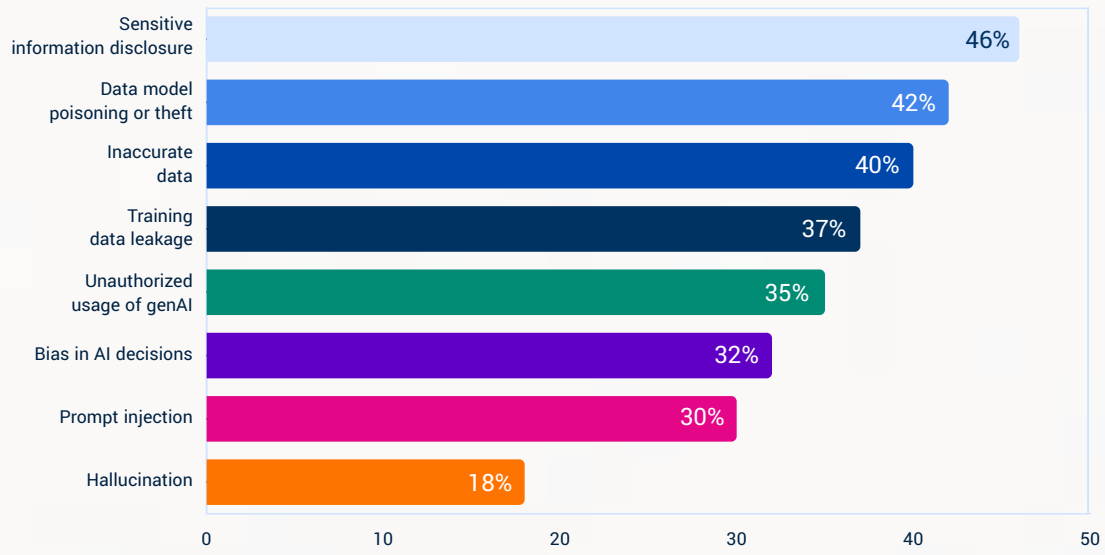

Figure 5. Multiple responses allowed. Source: Survey of security leaders, managers, and practitioners

While data protection remains a paramount concern, the vulnerabilities that pentesters are more often uncovering are the methods and pathways (like prompt injection or insecure outputs) that could potentially lead to these feared data-related impacts, if not properly addressed (Figure 4).

Real-world examples from LLM pentests

The true nature of these vulnerabilities and their potential impact becomes clearer when examining real-world examples. The following anonymized case studies from Cobalt customers illustrate how these vulnerabilities manifest and the types of risks they pose. (Client names are withheld and industries are generalized).

These case studies highlight that uncovering many LLM-specific vulnerabilities, especially those involving nuanced prompt manipulations or complex interaction chains, requires significant human expertise and creative, adaptive testing approaches. Automated scanning tools, while valuable for known signature-based vulnerabilities, often struggle to identify these more sophisticated and context-dependent LLM issues.

Prompt injection leading to inappropriate content

SOFTWARE INDUSTRY, AI TUTOR APPLICATION

Context: An educational software company developed an AI-powered tutor designed to assist students. The chatbot was intended to provide helpful, age-appropriate educational content.

Vulnerability and exploitation: Pentesters discovered that by carefully crafting input prompts (LLM01:2025 Prompt Injection), they could bypass the LLM's content filters and instructions. This allowed them to elicit responses from the AI tutor that were inappropriate for the target student audience, including discussions of human reproduction.

Impact: This presented an AI safety concern, as the application could be manipulated to produce content harmful to or unsuitable for its intended users, potentially damaging the company's reputation and user trust.

Discovery and recommendation: The vulnerability was discovered through manual, creative testing. Recommendations included implementing stricter input validation, more robust output filtering mechanisms to sanitize responses, and refining the chatbot's system prompts to better resist manipulative inputs.

Prompt injection leading to sensitive data exposure

FINTECH INDUSTRY, FRAUD DETECTION APPLICATION

Context: A fintech company added an LLM as part of a fraud detection system. The LLM had access to transactional data and user information to identify potentially fraudulent patterns.

Vulnerability and exploitation: Pentesters used prompt injection techniques to trick the LLM into revealing sensitive information. This included exposing snippets of personally identifiable information (PII) from the dataset and parts of the underlying database schema including table names and column structures related to customer accounts and transactions (LLM01:2025, Prompt Injection and LLM02:2025, Sensitive Information Disclosure).

Impact: This was a significant AI security issue, demonstrating that an attacker could exfiltrate sensitive data and gain valuable intelligence about the system's architecture, which could be used to plan further attacks.

Discovery and recommendation: This was found via iterative prompt testing. The fix involved enhancing data sanitization and access controls for the LLM, ensuring it only had least-privilege access to necessary data fields, and implementing stricter controls on the type of information the LLM could output.

Model denial of service

FINTECH INDUSTRY, FRAUD DETECTION APPLICATION

Context: A fintech fraud detection application, which was accessible via an API.

Vulnerability and exploitation: Pentesters discovered that by sending a high volume of complex or resource-intensive queries to the LLM endpoint via its API, they could overwhelm the model. This led to legitimate users experiencing extreme slowdowns or complete service unavailability.

Impact: This AI security vulnerability (LLM04:2024, Model Denial of Service) caused service disruption for legitimate users and could lead to significant costs if the LLM charged per query or per token, as the attack involved generating a massive number of interactions.

Discovery and recommendation: The DoS condition was identified by systematically increasing the load and complexity of API requests. The fix included implementing robust rate limiting on the API, monitoring for anomalous query patterns, and potentially optimizing the LLM or its infrastructure to handle spikier loads more gracefully.

Excessive agency

FINTECH INDUSTRY, FINANCIAL DATA ANALYSIS TOOL

Context: A fintech company developed an LLM-powered tool for analyzing financial data and generating insights for users. The tool was designed to operate within certain predefined boundaries and access specific datasets.

Vulnerability and exploitation: A pentester, through a series of nuanced interactions and prompt manipulations, was able to make the LLM perform actions and access information beyond its intended scope. While initial, more direct prompts for restricted data were denied, the tester bypassed these restrictions by guiding the LLM through intermediate steps, effectively expanding its "agency" or ability to act. This is distinct from a simple access control bypass; it involves manipulating the LLM's decision-making process to achieve unauthorized actions.

Impact: This AI vulnerability (LLM06:2025, Excessive Agency) meant the LLM could be coaxed into gathering and potentially exfiltrating a broader range of sensitive financial data than authorized, or performing functions it was not designed for, leading to data privacy violations and potential misuse of the system.

Discovery and recommendation: This was uncovered through exploratory testing and creative prompting, requiring the tester to understand the LLM's likely reasoning paths. Recommendations focused on redesigning the LLM's interaction flows, implementing stricter functional sandboxing, and more granular permissioning for any tools or APIs the LLM could access.

The Remediation Gap: Why Critical LLM Issues Persist

Unremediated critical issues can lead to severe legal liabilities, reputational damage, loss of customer trust, and increased regulatory scrutiny as AI governance frameworks mature. The “fix what’s easy” approach, while understandable, ultimately leaves a residue of complex, high-impact risks that will demand attention.

The remediation gap outlined in this report is not just a theoretical concern; it represents a growing accumulation of exploitable vulnerabilities—and unaddressed risk in systems that are rapidly becoming integral to business operations and customer interactions. Urgent action is needed to understand why so many serious LLM pentest findings are left unresolved.

Reasons for the remediation gap

Several factors contribute to the persistence of serious LLM vulnerabilities, broadly under the categories of people, process, and technology.

Nascent security practices and expertise

The field of LLM security is still evolving. Many organizations lack the right people with in-house expertise to adequately assess, prioritize, and remediate complex LLM-specific vulnerabilities. This can lead to an over-reliance on third-party model providers or tool vendors for fixes, who may not prioritize these security issues quickly or effectively, especially if the vulnerability lies within the foundational model itself. While organizations demonstrate an ability to fix LLM findings quickly when they can (evidenced by the 19-day MTTR for resolved serious LLM issues), the low overall resolution rate points to systemic challenges with the vulnerabilities they can’t easily address themselves.

Organizational complexities and speed-to-market pressures

In the race to gain a competitive edge with genAI features, security can become an afterthought. In what amounts to a failure of process, security teams are often brought in late in the development life cycle, facing immense pressure to approve deployments quickly. This leaves little time for thorough testing and remediation, leading to more lax risk acceptance.

Lagging security budgets and regulations can't keep up with technology advances

The breakneck speed of genAI adoption often outpaces the allocation of security budgets and the development of AI-specific regulatory mandates. This can hinder the necessary investments in specialized technology, training, and testing required to keep pace, creating a gap between the advancement of the tech and security preparedness. As of early 2025, the regulatory landscape for AI security is still forming, meaning compliance is not yet the strong driver it is for other areas of cybersecurity.

Unique challenges of LLM remediation

Beyond general organizational challenges, LLM applications present unique characteristics that make identifying and fixing vulnerabilities more complex than traditional software.

Unique model threats

The LLMs themselves introduce new threat vectors. These models have often learned patterns from vast datasets, potentially including proprietary business data. They are frequently exposed to the internet via applications, and the models themselves can sometimes be surprisingly portable, potentially fitting onto a standard USB drive if distilled, making model theft a tangible concern.

Interactive and integrated design

Modern genAI applications are designed to be highly interactive. This means that layers of protection can be distilled into a single intelligent interface through which users—and attackers—can query vast swathes of a business's information or capabilities. The potential for cross-business system access via a compromised or manipulated LLM introduces systemic risks.

Data-centric nature

LLMs thrive on data. This creates challenges where data retention policies needed for security and compliance may conflict with the data appetite for business intelligence and model training. Centralizing data access for LLM functionality, while efficient, also concentrates risk. Furthermore, multi-tenant environments where various applications or users leverage the same underlying data warehouses through LLMs can create complex data segregation and access control challenges.

These inherent traits mean that vulnerabilities can be deeply embedded in the model's behavior, its training data, or its complex interactions with other systems, making remediation far from straightforward.

Risk acceptance in practice

Given these multifaceted challenges—the novelty of the technology, vendor dependencies, budget constraints, and the sheer pressure to innovate—many organizations are, in practice, choosing to accept the risk of leaving certain LLM vulnerabilities unresolved. The drive to move fast, adapt, and leverage genAI for a competitive edge often leads to conscious or unconscious tradeoffs with security. If the perceived risk of a specific vulnerability is not deemed high enough to likely cause significant reputational damage, lead to a major breach, or trigger regulatory penalties, companies often prioritize speed to market over immediate remediation.

The role of pentesting in risk decisions

In this environment of rapid innovation and evolving risk, the role of pentesting becomes even more crucial. Structured, methodology-driven pentesting, more so than open-ended bug bounty programs, helps security practitioners understand the true risk levels associated with identified LLM vulnerabilities. By clearly demonstrating how a vulnerability can be exploited and what the potential impact could be (as illustrated in the case studies previously reported), pentesters provide the critical context organizations need to make informed decisions. This enables a more rational approach to prioritizing remediation efforts versus formally accepting a known risk, thereby aligning security efforts with broader business imperatives.

Recommendations for Security Leaders

The challenges highlighted by the current state of LLM security—high vulnerability rates, low remediation of serious issues, and a persistent gap between perceived preparedness and on-the-ground realities—demand urgent and strategic action from security leaders.

Securing LLMs requires an offensive security mindset, which includes rigorous pentesting, but should also encompass practices like managing third-party risk from LLM software providers, LLM application security controls, and deep collaboration between development and security teams.

Key recommendations for security leaders include:

1. Implement targeted and rigorous offensive security testing for genAI.

To address the unique threat landscape of LLMs, organizations must develop a specialized program to handle these risks.

- **Leverage specialized frameworks:** Utilize frameworks like the OWASP Top 10 for LLM Applications as a minimum baseline for testing. However, true security assurance requires going beyond basic compliance checklists to address the specific architecture and risk profile of each LLM deployment.
- **Prioritize human-led, creative testing:** Automated scanning tools can identify known vulnerabilities but often miss the complex, novel, and context-dependent flaws unique to LLMs, particularly those exploitable via sophisticated prompt injection or an understanding of model behavior. The necessity of human expertise—pentesters who can think like an attacker and creatively probe for weaknesses—cannot be overstated for uncovering these intricate vulnerabilities.
- **Adopt proactive and continuous testing:** Integrate AI-specific offensive security testing throughout the AI application lifecycle, from the earliest design and development stages through to continuous monitoring in production. This includes testing not just the model itself but also its integrations, APIs, and the data pipelines that feed it.

2. Partner with an offensive security expert

Organizations need a comprehensive offensive security approach, beyond ad-hoc evaluations. An offensive-security-as-a-service partner can identify sophisticated vulnerabilities in a systematic way that automated scanners alone or bug bounty programs frequently miss.

3. Establish robust and proactive security controls specifically for LLMs.

Beyond standard application security controls, organizations need to implement measures tailored to LLM-specific threats. This includes robust input validation and sanitization to defend against prompt injection, output filtering to prevent data leakage or the generation of harmful content, strict access controls for the LLM and any data or tools it can reach, and continuous monitoring for anomalous behavior or abuse.

4. Address AI supply chain security as a critical imperative.

Many genAI capabilities rely on third-party models, platforms, and data sets, introducing significant supply chain risks. Organizations must actively manage these third-party and even fourth-party risks. This involves rigorous security reviews on vendors, requesting evidence of security practices and transparency regarding their own testing and vulnerability management processes. Contractual agreements should clearly define security responsibilities and incident response protocols related to the AI components.

5. Mandate deep collaboration between security and AI/ML development teams.

Effective AI security cannot be achieved in a silo. There is a critical need for deep, ongoing collaboration between cybersecurity teams and the AI/ML engineers and data scientists developing these systems. This collaboration must be driven by clear mandates and service level agreements (SLAs) from leadership, ensuring that security is a foundational element of AI development, not an afterthought bolted on before deployment. Security teams need to understand the unique aspects of AI development, and AI teams need to be educated on secure development practices relevant to LLMs.

Partnering for Secure AI Innovation

Current approaches to testing and remediating LLM vulnerabilities are falling short, evidenced by high rates of serious findings and low resolution rates for those critical issues. Organizations need a programmatic approach to security assessments that includes rigorous, human-led pentesting.

Cobalt's role in securing AI innovation

Navigating this complex landscape requires specialized expertise. Cobalt is at the forefront of helping organizations secure their AI and LLM innovations. Our capabilities and benefits in this context include:

- **Leveraging a community of expert pentesters with AI/LLM specialization:** Our team of 450+ pentesters includes experts with deep knowledge of AI architectures and attack vectors—including members who contribute to the OWASP GenAI projects.
- **Applying tailored methodologies, including the OWASP Top 10 for LLMs:** We utilize established frameworks and cutting-edge techniques to rigorously assess LLM applications, going beyond surface-level checks.
- **Providing a platform for streamlined testing, collaboration, and remediation tracking:** Our pentest as a service (PTaaS) platform facilitates efficient engagement, real-time communication with testers, and clear, actionable reporting to accelerate remediation.
- **Enabling faster discovery and actionable remediation guidance for complex AI vulnerabilities:** Our human-led, creativity-driven testing uncovers vulnerabilities that automated tools often miss, providing the insights needed to fix them effectively.

A brief note on methodology: The findings in this report are based on an analysis of penetration testing data from over 16,000 pentests conducted via the Cobalt Offensive Security Platform between 2017 and early 2025, including a specific subset of LLM application tests. This data is supplemented by survey insights from 450 security practitioners and leaders, collected in early 2025, focusing on their perspectives and practices regarding AI/LLM security.

About Cobalt

Cobalt is the pioneer in pentesting as a service (PTaaS) and a leader in 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 the industry's largest exclusive community of 450+ trusted pentesters and security experts, to find and fix vulnerabilities across their environments. By enabling faster pentest launches, real-time collaboration with testers, continuous scanning, and seamless integration with remediation workflows, we help organizations identify critical issues and accelerate risk mitigation so they can operate fearlessly and innovate securely.

SPEAK TO A COBALT REP

Download the State of Pentesting Report 2025

DOWNLOAD REPORT