KUPPINGERCOLE ANALYSTS

Generative AI Defense

Jonathan Care

December 2, 2025

LEADERSHIP COMPASS 2025

This report offers an overview of the Generative AI Defense market and serves as a guide to help you find the best solution for your needs. It evaluates solutions that provide an inclusive set of security and compliance features designed to safeguard cloud-native applications throughout their development and production cycles. It also assesses how well these solutions can help organizations monitor, evaluate, and manage related risks.

Contents

Executive Summary

Generative Artificial Intelligence (GenAI) refers to systems that create new text, images, audio, and code based on learned patterns and has introduced significant risks across society. These include the proliferation of convincing misinformation and deepfakes that erode trust in authentic content; security vulnerabilities through sophisticated phishing, malicious code generation, and privacy breaches; economic disruption as AI displaces workers in creative and knowledge industries while concentrating power among tech giants; intellectual property conflicts over training data and creative attribution; social impacts like unhealthy AI dependencies and diminished human creativity; the amplification of biases leading to discrimination in critical decisions; and governance challenges as technology outpaces regulation. While GenAI offers substantial benefits in research, creativity, and accessibility, managing these interconnected risks through technical safeguards, thoughtful regulation, and social adaptation remains one of the defining challenges of our time.

Organizations deploying generative AI face numerous security threats. These risks cover both the traditional attack vectors (data leakage and exposure, access control and privilege abuse, API and interface security, and others) as well as numerous new ones that existing cybersecurity tools cannot address at all or at best partially. The latter have created an urgent need for specialized defense mechanisms that protect against prompt manipulation, model tampering, and harmful content generation. Generative AI Defense (GAD) represents a fundamental shift in security thinking, where attacks occur through natural language rather than code exploits, and where the protected asset can be manipulated into becoming the attack vector itself.

The primary drivers for GAD adoption center on three critical concerns: preventing sensitive corporate data from leaking through AI interactions, ensuring generated content meets compliance and ethical standards, and maintaining operational control over AI systems that users increasingly access directly. Organizations must recognize that generative AI applications bypass traditional security perimeters, creating new attack surfaces that require purpose-built protection mechanisms. New technologies, such as Model Context Protocol (MCP) servers, increase organizational challenges in securing APIs, as when employees use AI tools to process work materials, they often inadvertently transmit sensitive data outside corporate firewalls, rendering traditional security controls, such as Data Leakage Prevention (DLP) systems and access management protocols, essentially irrelevant.

While some forward-thinking organizations understand these risks, many remain focused on productivity gains without implementing appropriate safeguards, creating critical blind spots where proprietary information can be exposed through seemingly innocuous tools. Regulatory pressure adds urgency as governments develop AI-specific legislation requiring demonstrable security and governance controls.

The GAD market remains nascent but exhibits strong growth indicators as enterprises transition from pilot projects to production AI deployments. Market size estimates vary significantly, given the field's recent emergence, though spending appears concentrated among organizations with substantial AI investments rather than broad market adoption.

Even at this early stage, the GAD market shows clear capability stratification, with vendors focusing on different layers of the security stack. Most solutions provide adequate prompt injection detection and output filtering, but few offer sophisticated model integrity protection or advanced threat analytics. Organizations find themselves assembling multiple point solutions rather than relying on single-vendor platforms, indicating market immaturity. DLP capabilities vary significantly in sophistication, with some vendors offering basic keyword filtering while others provide advanced semantic analysis. The gap between enterprise governance requirements and available tooling remains substantial, particularly for compliance reporting and risk assessment frameworks.

Primary buyers include large enterprises across financial services, healthcare, and technology sectors, where regulatory compliance and data protection requirements drive security investment decisions. Government agencies represent another significant buyer category, particularly those handling classified or sensitive information. The market shows global distribution with notable concentration in North America and Europe, where AI regulation development leads other regions. Small and medium-sized businesses remain secondary buyers, typically adopting GAD capabilities through integrated solutions rather than standalone products.

Early adopters dominate current purchasing, characterized by organizations with dedicated AI teams and established risk management frameworks. These buyers often maintain hybrid deployment preferences, combining cloud-based GAD services with on-premises capabilities for sensitive workloads. Enterprise buyers typically evaluate GAD solutions as part of broader AI governance initiatives rather than standalone security purchases.

This analysis follows the KuppingerCole Leadership Compass Methodology, which evaluates vendors across multiple dimensions, including innovation, market presence, and technical capabilities.

Key Findings

- **Rapid innovation cycle creates competitive advantage opportunities:** The market immaturity and fragmented vendor landscape mean organizations can be strategic first-movers. Early adopters have the opportunity to shape their GAD architecture with best-of-breed solutions before market consolidation occurs.
- **Multi-vendor approach enables defense in depth:** CISOs have the option to exercise architectural flexibility using multi-vendor solutions. Organizations are not forced into a single vendor's vision of GAD; instead, they can layer complementary solutions that address specific risk areas that are most critical to their environment. The need to assemble multiple point solutions encourages more thoughtful security architecture and avoids the monoculture vulnerabilities that come with platform lock-in.
- **Market gaps represent clear differentiation opportunities:** Compliance automation, advanced threat protection, and sophisticated DLP create a transparent roadmap for vendor evaluation and strategic partnerships. Organizations know

exactly where to focus their due diligence, and vendors who address these gaps first will capture significant market share.

- **Market immaturity creates integration challenges**: organizations must assemble multiple point solutions rather than relying on all-encompassing platforms, as no single vendor provides complete GAD coverage across all security domains.
- **Vendor classification requires careful due diligence**: market leadership in traditional security does not guarantee GAD effectiveness, while innovative startups may offer superior technology but carry commercial risks that require mitigation strategies.
- **Compliance capabilities remain underdeveloped**: despite growing regulatory pressure from the EU AI Act and other frameworks, most vendors lack pre-built compliance templates and automated reporting, forcing organizations to build custom governance frameworks.
- **Prompt security is table stakes, but advanced threats are poorly addressed:** while most solutions detect basic prompt injection and jailbreaking, sophisticated attacks like multi-step chains, indirect injection, and model extraction remain largely unprotected across the vendor landscape.
- **The enterprise integration ecosystem is fragmented**: third-party integrations with DLP, Security Information and Event Management (SIEM), and identity providers vary significantly, with many vendors making broad integration claims that do not match operational reality, requiring careful validation during evaluation.
- **The sophistication of data leakage prevention varies widely**: solutions range from basic keyword filtering to advanced semantic analysis, with significant gaps in protecting against training data reconstruction and membership inference attacks.
- **Geographic deployment preferences signal market split**: cloud-first vendors dominate North American/European markets, while on-premises solutions remain critical for regulated industries and air-gapped environments, indicating parallel market evolution paths.
- **Vendor lock-in risks are substantial**: platforms from Microsoft, Palo Alto Networks, and IBM provide deep integration within their ecosystems but create dependency challenges for multi-vendor environments, requiring careful architectural planning.

Market Analysis

Generative AI Defense has emerged as cybersecurity's most pressing challenge, where conventional security tools fail against attacks conducted through natural language conversations rather than code exploits.

Security vendors are racing to build GAD capabilities through acquisitions and partnerships, recognizing that prompt injection and model manipulation attacks require entirely different detection and prevention mechanisms than traditional threats. Specialized vendors have appeared with focused solutions for specific GAD problems, in particular:

- Output sanitization
- Training data protection
- Prompt filtering

while established security companies work to integrate AI-specific protections into existing platforms. The market shows clear segmentation between suites attempting to cover the full GAD spectrum and point solutions targeting specific attack vectors, with neither approach yet proving definitively superior as organizations test different defensive strategies.

The Generative AI Defense market is in its early stages of development, with analyst estimates varying widely as organizations begin to budget for AI-specific security requirements alongside traditional cybersecurity investments. The market divides along several distinct lines: prompt-layer security vendors focus on input validation and injection prevention, while output safety specialists concentrate on content filtering and compliance checking. Model security providers target the protection of AI systems themselves against tampering and extraction attacks. A separate category addresses data leakage prevention, controlling what information flows into and out of generative AI systems.

Some vendors pursue platforms attempting to cover multiple protection layers, while others maintain a narrow focus on specific attack vectors or deployment phases. The market also segments by AI modality, with specialized solutions emerging for text generation, code generation, and multimedia AI systems, each presenting unique security challenges that require tailored defensive approaches.

Several frameworks are emerging to standardize LLM security:

- **OWASP LLM Top 10**: A framework specifically for LLM vulnerabilities
- **NIST AI Risk Management Framework**: Broader AI governance including security components
- **EU AI Act**: Regulatory framework with security implications for high-risk AI systems
- **Industry-specific guidelines**: From sectors like healthcare, finance, and government
- **The UK AI Safety Institute** appears to be trying to bridge the gap between scientific research and governmental regulation
- **MITRE ATLAS** (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a knowledge base that documents adversary tactics and techniques used to attack machine learning systems, modeled after the MITRE ATT&CK framework.

Geographic Distribution and Regional Dynamics

The United States leads in both vendor innovation and customer adoption, with Silicon Valley enterprises and financial services firms representing the most active buyer segments. Europe has particularly strong adoption in Germany, the UK, and Nordic countries, where GDPR experience has raised awareness of AI governance requirements. The implementation of the EU AI Act is accelerating European demand, with compliance-driven spending expected to surge in 2025-2026.

Asia-Pacific exhibits the highest growth trajectory, led by Japan, Singapore, and Australia. China represents a unique market dynamic, with domestic vendors serving local requirements while international vendors face access limitations.

Competitive Landscape Dynamics

The GAD market exhibits clear segmentation between platform players and point solution specialists. Platform vendors (Cisco, IBM, Microsoft, and Palo Alto Networks) leverage existing security relationships to cross-sell GAD capabilities, often bundling AI security into broader enterprise agreements. This approach dominates large enterprise sales but may not address specialized requirements.

Point solution vendors fall into several categories:

- **Prompt security specialists** (Adversa AI, TrojAI) focus on input/output protection
- **Data protection vendors** (1touch.io, Sentra) extend DSPM capabilities to AI use cases
- **Startups** (Enkrypt AI, NeuralTrust) attempt full-spectrum coverage
- **Adjacent players** (Arkose Labs) extend existing capabilities to AI contexts

Market Consolidation Trends

Acquisition activity is accelerating as established security vendors seek GAD capabilities. Notable patterns include:

- Large platform vendors acquiring prompt security startups for technical capabilities
- Data Security Posture Management (DSPM) vendors expanding through AI-specific acquisitions
- Traditional DLP vendors developing or acquiring GAD functionality
- Cloud providers building native AI security capabilities

Investment flows heavily favor early-stage GAD specialists, with over $200 million in disclosed venture funding across the sector in 2024. However, commercial viability remains challenging for pure-play vendors competing against platform players with existing customer relationships.

The Generative AI Defense market is experiencing rapid consolidation as established security vendors move decisively to acquire specialized AI protection capabilities, signaling a shift from organic development to aggressive M&A strategies. This consolidation wave reflects the market's recognition that AI security requires fundamentally different technical approaches than traditional application security, and that time-to-market pressures make acquisition more viable than internal development for most major vendors.

For example, F5 Networks strengthened its market position through its strategic acquisition of CalypsoAI, a move that directly addresses the limitations of traditional security tools when applied to generative AI workloads, while Palo Alto Networks has completed its acquisition of Protect AI, taking a markedly different but equally strategic approach to the AI security challenge. These high-profile acquisitions by market leaders underscore the strategic imperative major security vendors place on AI defense capabilities and signal accelerating market consolidation as enterprises demand all-encompassing, platform-integrated AI security from established providers with proven enterprise credentials.

The shift toward acquisition over internal development reflects several market realities: the specialized talent required for AI security—bridging machine learning research, adversarial AI expertise, and enterprise security architecture—is exceptionally scarce and difficult to recruit at scale; the technical complexity of AI-specific threats requires years of focused R&D that established vendors cannot afford to spend given current market urgency; and early-stage companies like CalypsoAI and Protect AI have already validated their approaches with enterprise customers, de-risking the technology adoption curve for acquirers.

Beyond the immediate technical capabilities, these acquisitions reflect a broader strategic bet that enterprises will strongly prefer to consolidate their security stack with existing trusted vendors rather than introducing new point solution providers, particularly for a category as nascent and rapidly evolving as AI security where integration overhead and vendor management complexity pose significant operational burdens for already-stretched security teams.

Buyer Behavior and Purchase Patterns

Enterprise buying typically occurs within broader AI governance initiatives rather than standalone security purchases. Chief Information Security Officers (CISOs) collaborate closely with AI/ML teams, creating longer but more strategic sales cycles averaging 8-16 months for all-encompassing solutions.

Proof-of-concept deployments precede production implementations in approximately 80% of enterprise sales, reflecting buyer caution with emerging technology. Organizations frequently deploy multiple GAD solutions in parallel, indicating market immaturity and solution gaps.

Based on vendor feedback, organizations are increasingly funding GAD initiatives from existing security budgets rather than creating dedicated AI security lines. This trend favors vendors with established security relationships over AI-native startups.

Technology Evolution and Standards Impact

Emerging standards significantly influence market dynamics. OWASP LLM Top 10 adoption creates common vulnerability frameworks, while NIST AI Risk Management Framework provides governance structure. EU AI Act compliance requirements are driving significant European spending on audit and reporting capabilities.

The technical evolution toward agentic AI systems creates new security requirements that current solutions incompletely address, suggesting continued market expansion and vendor differentiation opportunities. Integration complexity with existing security stacks remains a significant adoption barrier, particularly for organizations with mature SIEM and Security Orchestration Automation and Response (SOAR) deployments.

This market analysis indicates a rapidly maturing but still fragmented landscape where established vendors hold deployment advantages while specialized vendors drive technical innovation, creating a dynamic competitive environment likely to see continued consolidation and evolution.

Delivery Models

Generative AI Defense solutions deploy across multiple models depending on data sensitivity requirements and infrastructure preferences. SaaS offerings dominate the market, providing immediate deployment for organizations seeking rapid protection without the infrastructure overhead. These cloud-based solutions offer advantages in threat intelligence updates and model training but raise concerns for organizations handling sensitive data.

On-premises deployments appeal to highly regulated industries and government agencies requiring complete data control. Private cloud installations provide middle-ground approaches, maintaining data sovereignty while leveraging cloud operational benefits. IaaS deployments allow organizations to customize security configurations while maintaining cloud scalability.

Agent-based architectures remain common for full-feature monitoring, particularly for solutions requiring deep integration with AI development pipelines or runtime protection capabilities. However, agentless options have emerged for specific use cases like API gateway protection and output filtering, where security controls operate at network or application layers rather than requiring endpoint installation.

Vendors increasingly offer multiple deployment options within single platforms, recognizing that enterprise customers often require different models for different AI applications based on data classification and risk tolerance. Integration capabilities vary significantly across deployment models, with SaaS solutions typically providing broader third-party connectivity while on-premises options offer deeper customization potential.

Required Capabilities

Generative AI Defense solutions must address core security challenges unique to AI systems while integrating with existing enterprise security infrastructure. Organizations evaluate GAD platforms based on their ability to provide protection across the AI lifecycle, from development through runtime operation.

Core Security Capabilities

- Prompt injection detection and blocking to prevent malicious manipulation of AI system behavior through crafted inputs
- Output content filtering and safety controls to screen generated material for harmful, inappropriate, or non-compliant content
- Data Leakage Prevention mechanisms that control sensitive information flow into and out of generative AI systems
- Model integrity protection against tampering, unauthorized access, poisoning attacks, and extraction attempts during deployment
- Real-time threat monitoring and anomaly detection for identifying suspicious AI system behavior patterns
- Authentication and authorization controls for managing user access privileges and interaction permissions with AI capabilities

- Audit logging and traceability for all AI interactions to support compliance and forensic analysis
- Runtime protection monitoring of AI system operations to detect and respond to active threats.

Governance and Compliance Capabilities

- Policy enforcement frameworks enabling organizations to establish and maintain responsible AI usage guidelines
- Regulatory compliance tools supporting adherence to emerging AI legislation and industry-specific requirements
- Risk assessment methodologies for evaluating and classifying generative AI applications based on potential impact
- Governance dashboards providing centralized visibility and control over organizational AI deployment patterns
- Incident response automation with playbooks specifically designed for AI-related security events
- Configuration management tools ensuring consistent security policies across multiple AI deployments
- Compliance reporting capabilities generating documentation required for regulatory audits and internal reviews

Integration and Enterprise Capabilities

- Enterprise security ecosystem integrations with existing SIEM, identity management, and security orchestration platforms
- Multi-environment deployment support across cloud, on-premises, and various infrastructure configurations
- API security controls protecting generative AI interfaces from unauthorized access and abuse
- Performance optimization features maintaining AI system responsiveness while implementing security measures
- Scalability features supporting enterprise-wide deployments across diverse AI applications and use cases

Leadership

Selecting a vendor of a product or service must not only be based on the information provided by a KuppingerCole Leadership Compass. The Leadership Compass provides a comparison based on standardized criteria and can help identify vendors that shall be further evaluated. However, a thorough selection includes subsequent detailed analysis and a proof of Concept (PoC), or pilot phase based on the specific needs of the customer.

Based on our research and analysis of the vendor responses, we created various Leadership ratings. The Overall Leadership rating provides a combined view of the ratings for Product, Innovation, and Market Leadership.

Overall Leadership

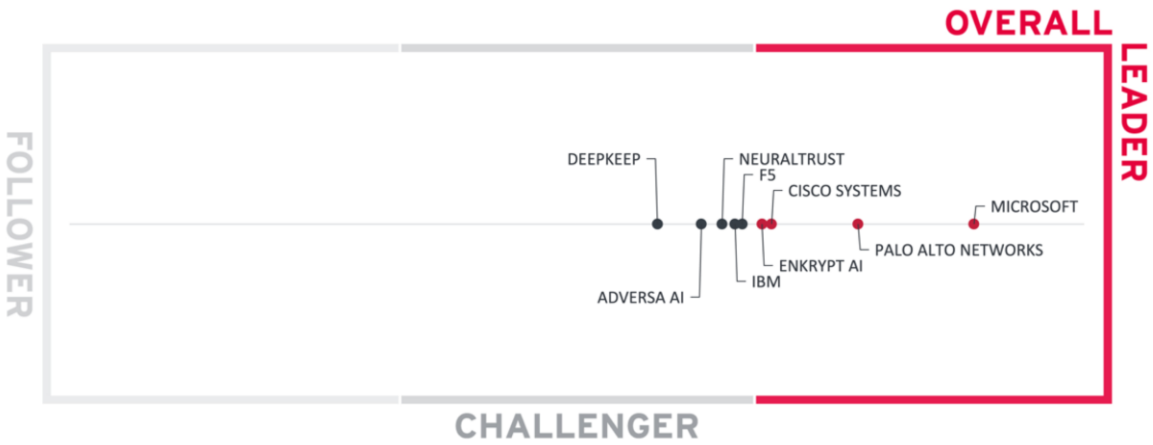


Figure 1: Overall Leadership in the Generative AI Defense market

The Overall Leadership chart is linear, with Followers appearing on the left side, Challengers in the center, and Leaders on the right. The rating provides a consolidated view of all-around functionality, market presence, and financial security.

However, these vendors may differ significantly in terms of product features, innovation, and market leadership. Therefore, we recommend considering our other leadership categories in the sections covering each vendor and their products to get a thorough understanding of the players in this market and which of your use cases they support best.

Among the Overall Leaders, we can find:

Microsoft, which achieved leadership through its unparalleled scale, has unique technical capabilities like detecting both jailbreak and indirect prompt injection attacks, and enterprise ecosystem integration.

Palo Alto Networks achieved leadership by leveraging its extensive cybersecurity heritage to deliver a unified platform covering the entire AI lifecycle, with unique dynamic sandboxing capabilities and proven enterprise deployments.

Cisco Systems offers impressive technical capabilities, including algorithmic red teaming and network-level guardrail enforcement backed by global partner network, despite the platform's very recent availability (April 2025).

Enkrypt AI earned leadership despite its startup status by delivering cutting-edge, multimodal security across text, images, and audio, with over 300 attack patterns and zero-day detection capabilities.

Among the **Challengers,** we observe vendors that demonstrate strong capabilities in specific areas but lack the coverage or market maturity required for leadership. IBM brings massive enterprise credentials and compliance automation strength, but its brand-new product status (launched May 2025) with few reported customers and heavy reliance on OEM partnerships prevent leadership recognition.

Several other challengers face growth challenges despite technical innovation, with limited funding, small customer bases, or missing enterprise features preventing them from achieving leadership status in this rapidly evolving market.

There are no **Followers** in the Overall Leadership rating.

Overall Leaders are (in alphabetical order):

- Cisco Systems
- Enkrypt AI
- Microsoft
- Palo Alto Networks

Product Leadership

Product leadership is the first specific category examined below. This view is mainly based on the presence and completeness of required features as defined in the required capabilities section above. The vertical axis shows the product strength plotted against the combined/overall strength on the horizontal axis. The Product Leadership chart is rectangular and divided into thirds. Product Leaders occupy the top section. Challengers are in the center. Followers are in the lower section.

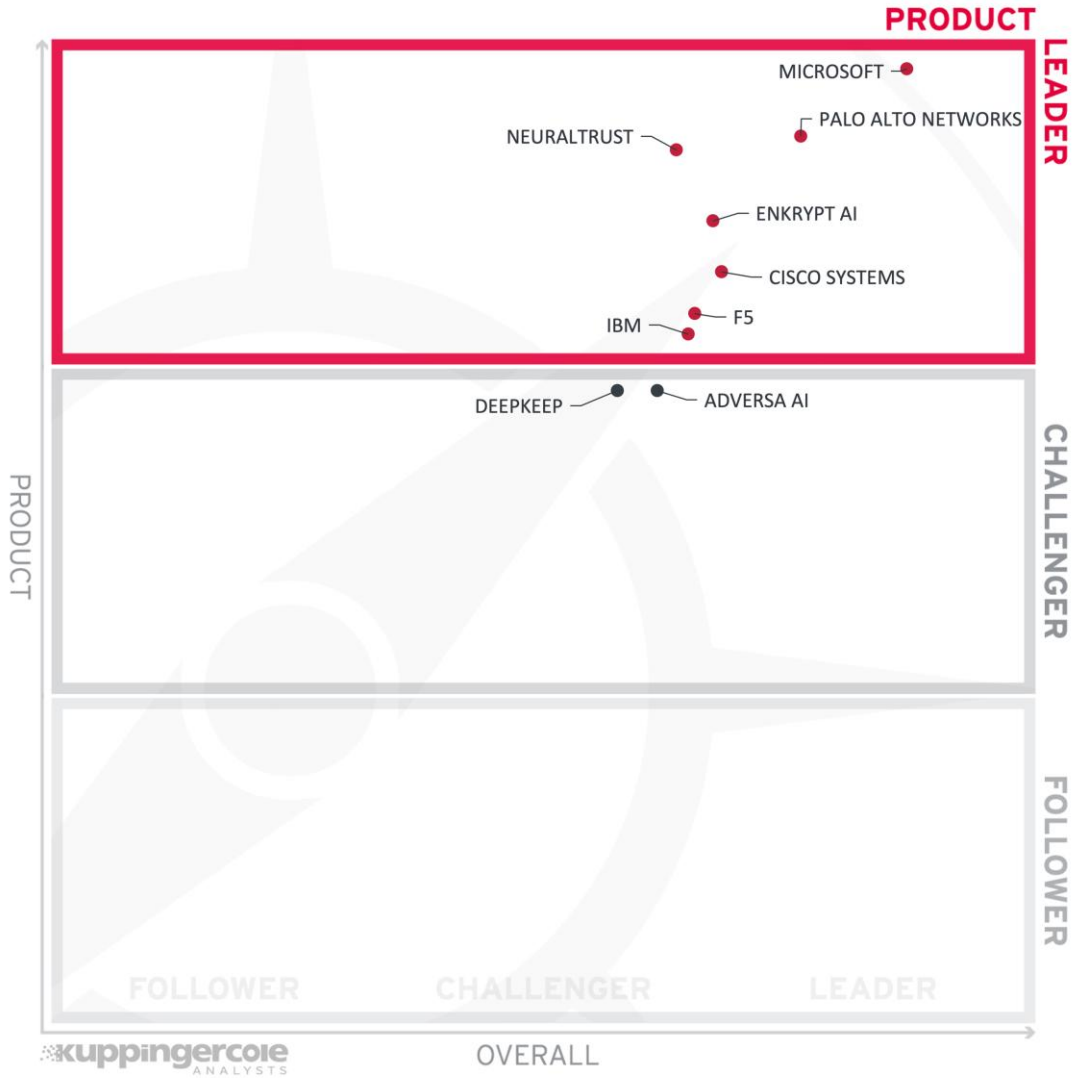


Figure 2: Product Leadership in the Generative AI Defense market

Product Leaders achieved their positions through feature coverage across the GAD spectrum.

Microsoft demonstrates product leadership with its unique ability to detect both jailbreak and indirect prompt injection attacks with specialized models, multi-layered protection, and extensive compliance frameworks that cover major global regulations.

Palo Alto Networks achieved leadership by providing unified management across five integrated components covering the entire AI lifecycle, leveraging proven technologies like WildFire and Enterprise DLP, with 1000+ predefined DLP patterns and unique dynamic sandboxing capabilities.

NeuralTrust provides LLM and Agentic security with GPU-optimized performance targeting sub-10ms latency, multi-language detection capabilities, and protection against multiple attack vectors including multi-step chain attacks.

Enkrypt AI secured leadership despite its startup status by delivering the most multimodal protection across text, images, and audio, with 300+ attack patterns, autonomous red teaming capabilities, and support for many compliance frameworks, including GDPR, HIPAA, and PCI DSS.

Cisco achieved leadership by providing Algorithmic red teaming across 200+ threat categories while introducing network-level guardrail enforcement. Talos real-time threat intelligence and MCP server scanning and run time protection also contributed to this rating.

F5's acquisition of Calypso AI brings not only useful capabilities but also strong product leadership into their portfolio, laying the ground for rapid product development.

IBM's integration with the comprehensive IBM product portfolio coupled with its strong integration with watsonx.governance make this a product that will be suitable for those organizations with a substantial investment in IBM technology.

Challengers demonstrate strong capabilities in specific product areas but lack the feature coverage required for leadership. These vendors typically excel in particular domains such as prompt security, data protection, or compliance reporting, but show significant gaps in other critical areas. Some challengers focus heavily on specific attack vectors like prompt injection while lacking sophisticated model integrity protection or output filtering. Others provide strong enterprise integration and compliance capabilities but miss advanced AI-specific threat detection features.

There are no **Followers** in this product leadership rating.

Product Leaders (in alphabetical order):

- Cisco
- Enkrypt AI
- F5
- IBM
- Microsoft
- NeuralTrust
- Palo Alto Networks

Innovation Leadership

Next, we examine innovation in the marketplace. Innovation is, from our perspective, a key capability in all IT market segments. Customers require innovation to meet evolving and even emerging business requirements. Innovation is not about delivering a constant flow of new releases. Rather, innovative companies take a customer-oriented upgrade approach, delivering customer-requested and other cutting-edge features, while maintaining compatibility with previous versions.

This view is mainly based on the evaluation of innovative features, services, or technical approaches as defined in the Required Capabilities section. The vertical axis shows the degree of innovation plotted against the combined/overall strength on the horizontal axis. The Innovation Leadership Chart is rectangular and divided into thirds. Innovation Leaders occupy the top section. Challengers are in the center. Followers are in the lower section.

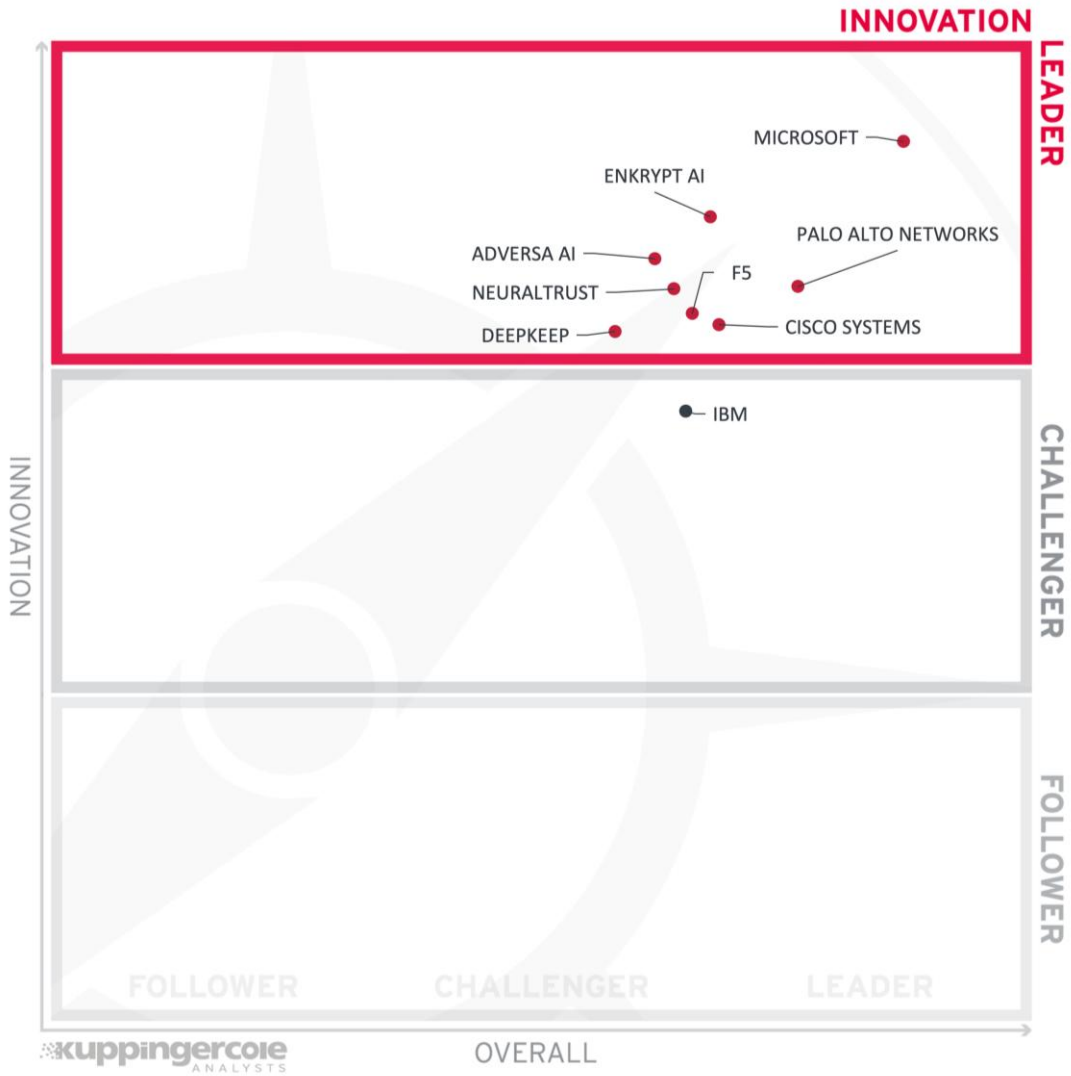


Figure 3: Innovation Leadership in the Generative AI Defense market

Innovation Leaders are those vendors that deliver cutting-edge products, not only in response to customers' requests but also because they are driving technical changes in the market by anticipating what will be needed in the months and years ahead. They are distinguished by their cutting-edge technical approaches and forward-thinking capabilities that push the boundaries of GAD technology.

Adversa AI holds a patent in AI protection and is an early commercial provider of continuous automated Red Teaming platforms that proactively uncover vulnerabilities in GenAI applications and agentic AI systems, including discovering exploits in systems like ChatGPT, DeepSeek, and Grok.

Cisco distinguishes itself by actively contributing to evolving industry frameworks including NIST AI-RMF, MITRE ATLAS, and OWASP Top 10 for LLM while delivering continuous threat intelligence updates from Cisco's AI research teams, ensuring organizations stay ahead of emerging adversarial AI tactics like prompt injections, jailbreaking, and sensitive data leaks throughout the entire AI application lifecycle.

DeepKeep's demonstrated rapid innovation cycle and breadth of features give them a place in this leader's category.

Enkrypt AI demonstrates innovation leadership through its autonomous agent red teaming that proactively simulates attacks, a multimodal security framework covering text/images/audio, and automated compliance monitoring agents that analyze usage logs for violations.

F5 earned Innovation Leader positioning through its network-centric proxy architecture, extensible processor SDK enabling custom security modules, and hybrid CPU/GPU processing approach that balances detection efficacy with production performance requirements.

Microsoft achieves innovation through its industry-first detection of both jailbreak and indirect prompt injection attacks using specialized fine-tuned models, confidential computing with GPU/CPU support, and homomorphic encryption for secure inference. Homomorphic encryption for AI inference is a cutting-edge area, and the practical, broad deployment of homomorphic encryption for AI inferencing is still emerging. Hence, it is not universally available across all Microsoft AI Security products.

NeuralTrust secured leadership with its AI-native approach using generative AI to secure generative AI, continuous red teaming with real-time adaptation, and proprietary techniques like "Echo Chamber" jailbreak detection.

Palo Alto Networks demonstrates innovation through its dynamic sandboxing for model analysis, AI red teaming with iterative attack refinement, and AI agent security addressing emerging autonomous system threats.

Challengers show innovative capabilities in specific areas but lack the breadth or technical sophistication of market leaders. These vendors often demonstrate novel approaches to particular GAD challenges, such as advanced prompt detection algorithms or specialized

compliance automation, but have not achieved innovation across the full security spectrum. Some challengers bring interesting technical concepts from adjacent fields but require further development to fully address GenAI-specific requirements.

There are no **Followers** in this innovation leadership rating.

Innovation Leaders (in alphabetical order):

- Adversa AI
- Cisco Systems
- Enkrypt AI
- DeepKeep
- F5
- Microsoft
- NeuralTrust
- Palo Alto Networks

Market Leadership

Finally, we analyze Market Leadership. This is an amalgamation of the number of customers, the geographic distribution of customers, the size of deployments and services, the size and geographic distribution of the partner ecosystem, and the financial health of the participating companies. Market Leadership, from our perspective, requires global reach.

In this chart, the vertical axis shows the market strength plotted against the combined/overall strength on the horizontal axis. The Market Leadership Chart is rectangular and divided into thirds. Market Leaders occupy the top section. Challengers are in the center. Followers are in the lower section.

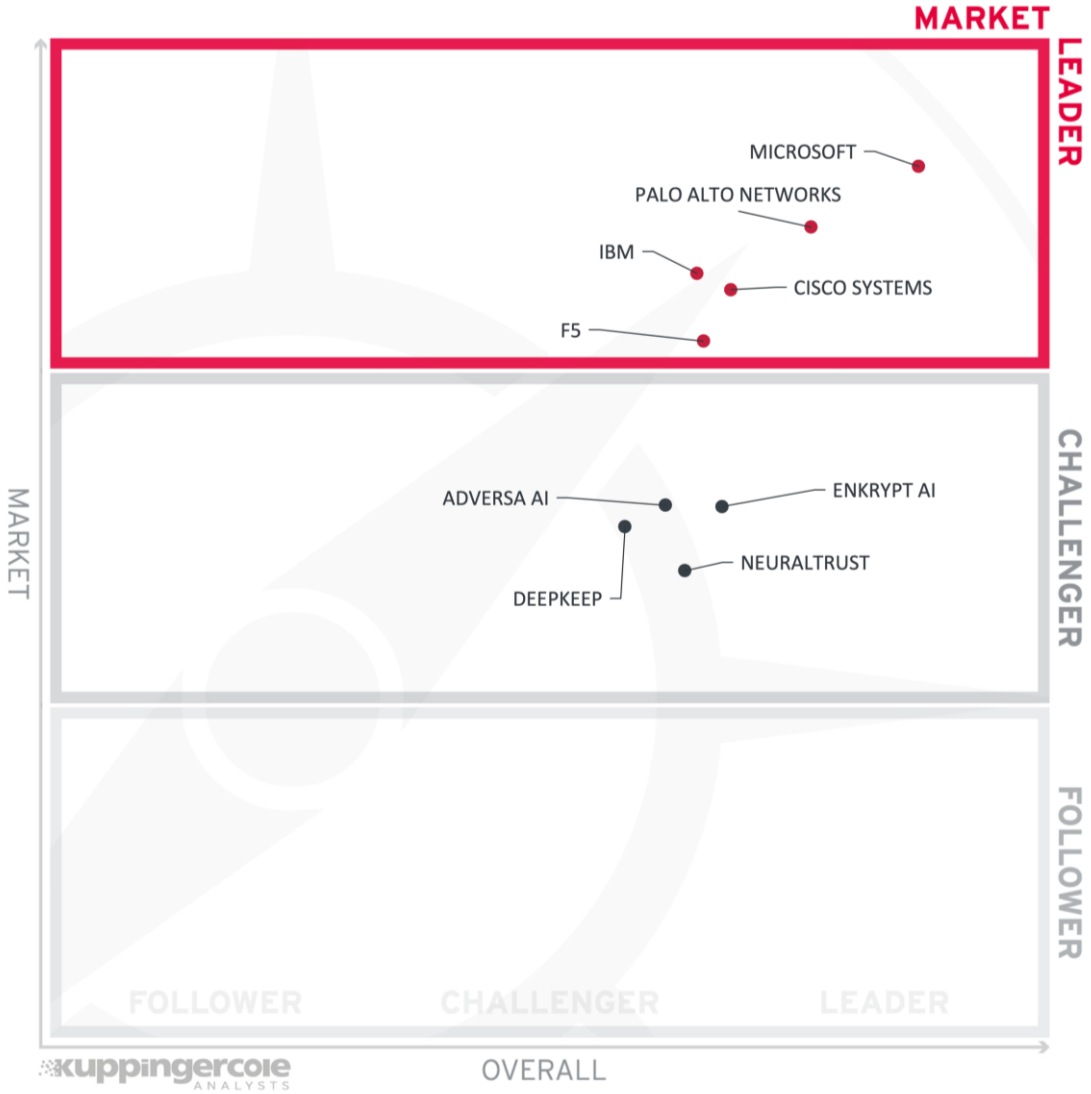


Figure 4: Market Leaders in the Generative AI Defense Market

Market Leaders demonstrate strong market presence through substantial customer bases, geographic reach, and proven deployment scale.

Microsoft achieved Market Leader status through its massive global presence serving over 1.4 million security customers, deep integration with enterprise Microsoft 365 and Azure ecosystems, and the scale to drive long-term market development across its installed base.

Palo Alto Networks earned Market Leader positioning through its extensive cybersecurity heritage, established VM/CN series firewall customer base providing migration pathways to AI Runtime Security, and production implementations including disclosed deployments with customers such as Glean demonstrating real-world market validation.

IBM appears as a Market Leader, leveraging its established enterprise relationships, including its existing Guardium customer base for data security and governance, and global delivery infrastructure across 175+ countries, though Guardium AI Security only recently launched in May 2025.

Cisco Systems appears as a Market Leader based on its established enterprise relationships across networking and security infrastructure, tens of thousands of global partners for delivery and managed services, and ability to leverage existing security deployments to introduce AI Defense across its massive installed base and has seen rapid growth since its launch in April 2025.

F5 secured Market Leader positioning with almost three decades of enterprise security expertise, extensive customer deployments across application delivery and security infrastructure, and established relationships in regulated industries requiring on-premises deployments.

Challengers demonstrate emerging-market presence but lack the scale, geographic distribution, or proven deployment history of market leaders. These vendors often demonstrate strong technical capabilities and growing customer adoption but have not achieved the market penetration required for leadership.

Some challengers have substantial backing and enterprise focus but limited customer references or geographic reach, while others show impressive growth trajectories but require additional time to establish market credibility.

There are no **Followers** in our market leadership rating as well.

Market Leaders (in alphabetical order):

- Cisco Systems
- F5
- IBM
- Microsoft
- Palo Alto Networks

Product/Vendor evaluation

This section contains a quick rating for every product/service we included in this KuppingerCole Leadership Compass document. For many of the products, there are additional KuppingerCole Product Reports and Executive Views available, providing more detailed information.

Spider graphs

In addition to the ratings for our standard categories such as Product Leadership and Innovation Leadership, we add a spider chart for every vendor we rate, looking at specific capabilities for the market segment researched in the respective Leadership Compass. For this market segment, we look at the following categories:

Input Protection & Validation: Capabilities for securing and validating inputs to generative AI systems, including prompt injection detection, jailbreak prevention, malicious input filtering, input sanitization, and protection against adversarial prompts designed to manipulate AI behavior.

Output Safety & Filtering: Technologies that monitor, filter, and control AI-generated content to prevent harmful, inappropriate, biased, or non-compliant outputs, including content safety checks, toxicity detection, factual accuracy verification, and customizable filtering rules.

Model Safety & Security: Protection of AI models themselves, including model integrity verification, prevention of model tampering or extraction, secure model deployment, protection against training data poisoning, and safeguarding model weights and parameters.

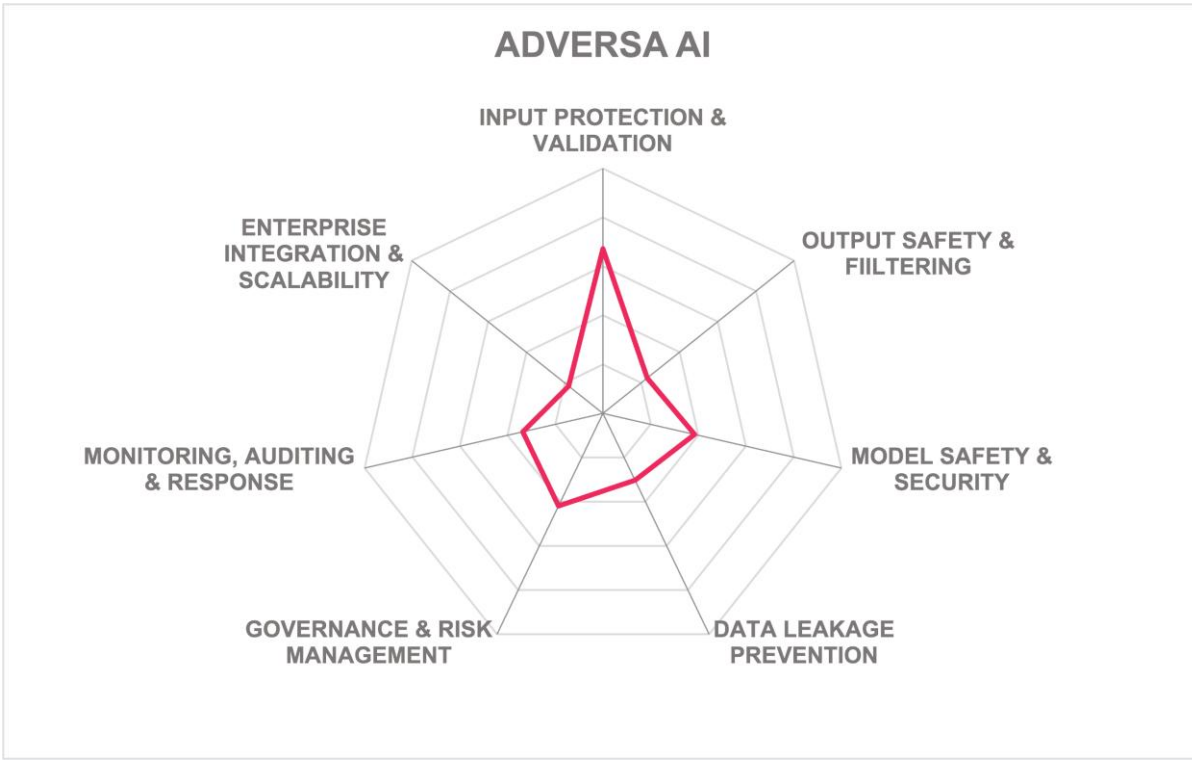
Data Leakage Prevention: Mechanisms to prevent sensitive or confidential information from being exposed through AI systems, including detection of PII/PHI in outputs, prevention of training data reconstruction, protection against membership inference attacks, and data anonymization capabilities.

Governance & Risk Management: Frameworks and tools for establishing AI governance policies, risk assessment, compliance management, regulatory adherence (GDPR, HIPAA, EU AI Act, etc.), policy enforcement, and responsible AI usage guidelines across the organization.

Monitoring, Auditing & Response: Capabilities for logging, real-time monitoring of AI interactions, audit trail generation, incident detection and response, forensic analysis, alerting mechanisms, and maintaining immutable records for compliance and security investigations.

Enterprise Integration & Scalability: Platform's ability to integrate with existing enterprise security infrastructure, scalability for large deployments, support for various deployment models, authentication systems integration, API compatibility, and performance optimization for enterprise-scale operations.

Adversa AI – AI Red Teaming Platform



Leader in

Adversa.ai was established in 2021 as an AI Security and Safety research lab and transitioned to a venture capital-backed company in 2024. The Adversa.ai Agentic Security Platform is delivered as multiple integrated products and services supporting flexible deployment models including on-premises software installation, cloud service, or managed services. Licensing is offered on annual or monthly subscription basis, or per appliance.

Adversa.ai integrates with Microsoft Active Directory (on-premises and Entra ID), LDAP directories, and IDaaS providers including Google and Keycloak. Data protection combines output monitoring and intent recognition using NLP models to identify PII, PHI, secrets, and sensitive business data through context-aware language models. SOAR/SIEM integrations use syslog and CEF message protocols.

Adversa.ai maintains a Self-Evolving Threat Intelligence Engine (SETI) collecting attacks and novel techniques from public and private sources. The AI Red Teaming engine leverages databases of known attacks, mutation algorithms, and algorithmically or AI-

generated unknown attacks, enabling continuous testing across 60+ threat and 200+ attack categories and all combinations thereof. Pattern recognition and clustering analyze prompt interactions, feeding a learning pipeline that evolves detection rules. The platform detects zero-day prompt engineering attacks through models trained on unusual attacks, continuous red teaming, and dedicated research. Pre-built defenses against common jailbreaking techniques support language-specific attack detection in multiple languages.

The platform claims the capability to detect model weight extraction via prompt engineering. Protection against adversarial attacks includes content safety with moderation, toxicity scoring, and safety classification models enforcing organizational policies. Different severity levels for content filtering alerts offer customizable thresholds. The platform supports AI agent security testing including MCP coverage.

Adversa.ai provides detailed reporting with centralized governance dashboards for AI usage visibility. Compliance reporting includes pre-built frameworks for EU AI Act, OWASP, MITRE ATLAS, NIST AI RMF, CSA, CoSAI, IEEE, and ISO 42001. The solution offers ethical AI implementation tools and risk assessment frameworks for generative AI deployments. Vulnerability triage uses CVSS-like scoring adapted for AI, factoring business risk, exploitability, model impact, and attack stability.

The platform is in process for ISO 27001 certification and holds SOC 2 Type 1 attestation. Professional services are available globally with 24×7 support in premium packages. Documentation is currently English-only. The company maintains three academic partnerships and an industry advisory board.

Adversa.ai suits organizations prioritizing AI-specific threat detection and continuous red teaming capabilities through automated attack generation and mutation testing.

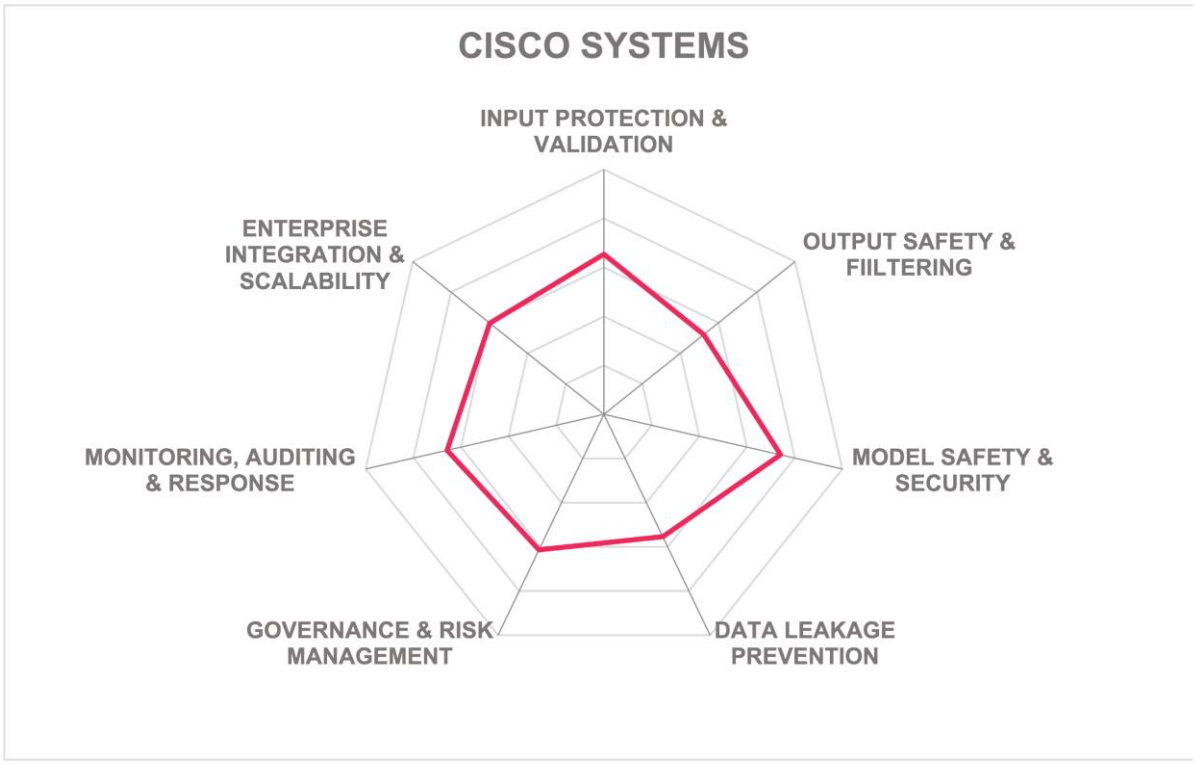
Strengths

- Early player in the GenAI security market since 2021
- Real-time detection and blocking of prompt injection, jailbreaking, and prompt chaining
- Self-Evolving Threat Intelligence Engine (SETI) for continuous threat collection
- Ability to detect zero-day prompt engineering attacks
- Advanced indirect prompt injection detection using causal trace analysis
- Red teaming covering 60+ safety threat categories and 80+ attack categories and their combinations
- Multiple attack generation approaches including AI-generated unknown attacks
- Differentiation between malicious prompts and legitimate edge cases

Challenges

- Small team may limit support and development capacity
- Limited model integrity or lifecycle security features
- Lacks support for GDPR, HIPAA, or PCI DSS
- Limited financial resources compared to larger competitors

Cisco Systems – AI Defense



Leader in

Cisco Systems, Inc. was established in 1984 and is publicly listed on NASDAQ (CSCO). Cisco AI Defense was announced in January 2025 and became generally available in April 2025. The solution is offered as a platform with subscription-based licensing per AI application per year, deployed primarily in public cloud with options for on-premises installation and managed services through Cisco partners. The platform provides deployment flexibility with consistent policies scaling across sandbox environments, hybrid cloud, and on-premises data centers, including integration with Cisco AI pods.

Cisco AI Defense is managed through Security Cloud Control with SAML and OIDC integration supporting Microsoft Active Directory, Entra ID, LDAP, and IDaaS providers. Authentication methods include token-based authentication, authenticator apps, biometrics, FIDO 2.0, passkeys, and one-time passwords. Role-based access control protects system functionality.

Built-in AI guardrails prevent sharing of PII, PHI, or PCI data with automatic detection and redaction. SIEM integration currently supports Splunk with REST API connectivity. The platform maintains a strategic partnership with NVIDIA for GPU acceleration and edge AI capabilities.

Cisco AI Defense employs algorithmic AI red teaming across 200+ threat categories. The validation engine includes multi-turn attack testing, informed by Cisco's published research demonstrating that open-source models show significant susceptibility to multi-turn attacks compared to single-prompt testing. Cisco's AI Security Research team, backed by Talos threat intelligence, maintains proprietary techniques for ingesting first-party and third-party intelligence with published research on arXiv, enabling rapid protection updates as new threats emerge. Validation and runtime results map to OWASP, MITRE ATLAS, and NIST standards.

AI supply chain security capabilities include MCP server scanning, model file component scanning, and full repository scanning. Cisco released an open source MCP scanner in October 2025. Runtime protection extends to MCP servers, with a secure MCP gateway currently in development featuring AI Defense integrated from initial design. Runtime guardrails are enforced at the network level, tailored to vulnerabilities of each model and application. The solution offers real-time detection and blocking of direct and indirect prompt injection attacks, jailbreaking attempts, and sensitive data leaking via model extraction. Content safety features detect and filter harmful instructions, adversarial attacks, and backdoors in models.

Cisco AI Defense generates detailed event logs capturing prompts, model responses, and application information. The solution supports customizable policy templates for ethical AI guidelines. ISO 27001 and SOC 2 attestations are in process. Professional services are delivered through Cisco's global partner ecosystem with 24x7 support in nine languages.

Cisco AI Defense suits large enterprises with existing Cisco infrastructure seeking established vendor partnerships for AI security, particularly organizations planning migration from cloud to on-premises AI deployments.

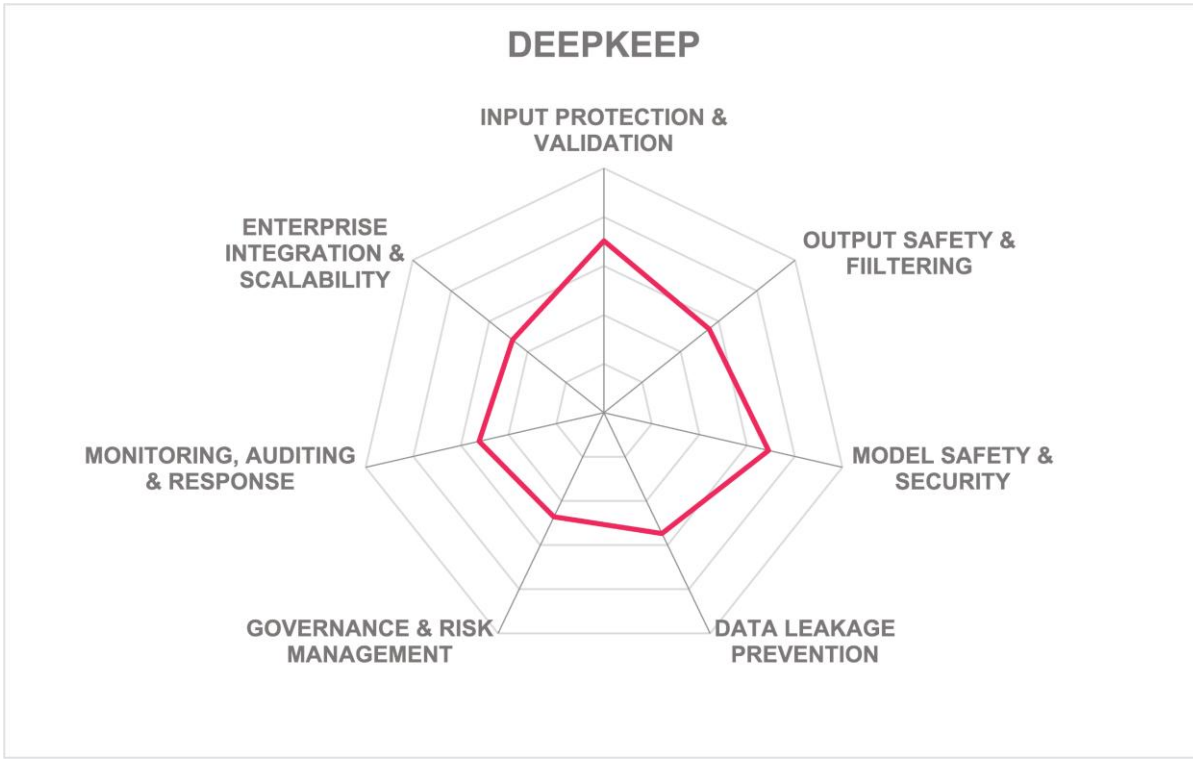
Strengths

- AI red teaming across 200+ threat categories with multi-turn attack testing
- Network-level enforcement of guardrails
- Real-time threat intelligence from Talos with proprietary ingestion enabling rapid updates
- MCP server scanning and runtime protection for agentic AI security
- Consistent policy enforcement across sandbox, hybrid cloud, and on-premises deployments
- Published security research backing detection capabilities

Challenges

- Documentation English-only
- Optimal value requires broader Cisco ecosystem adoption
- No explicit multi-modal protection
- No hallucination detection capability mentioned

DeepKeep – DeepKeep Platform



Leader in

DeepKeep was founded in 2021 and is headquartered in Tel Aviv, Israel. The company secured seed funding in May 2024 to launch its AI-Native Trust, Risk, and Security Management (TRiSM) platform. DeepKeep reports deployments with enterprises in AI computing, finance, and security industries. The solution offers custom pricing tailored to business needs. The company holds ISO 27001 certification.

DeepKeep offers an end-to-end platform with five modules covering the full model lifecycle from development to deployment. Model Scanning performs static and dynamic scanning to identify vulnerabilities and malware. Automated AI Red Teaming evaluates model robustness through continuous penetration testing, addressing security aspects (prompt injection, jailbreaking, PII leakage) alongside trust assessment (hallucinations, toxicity, bias). The AI Firewall offers dynamic protection with over 60 built-in guardrails, supporting granular input/output filtering configurable per user, team, or application. Agentic AI handles vulnerability assessment of AI agents and MCP server management with block/allow recommendations. AI Lens monitors employee and developer AI usage to prevent data leakage.

The platform addresses physical security aspects of AI, protecting computer vision models against adversarial attacks such as specially crafted stickers or patterns that fool cameras and sensors. DeepKeep began with computer vision four years ago before expanding to LLM protection, giving it multimodal capabilities beyond most competitors.

DeepKeep's context-aware protection stands out by understanding application context (distinguishing a finance chatbot from customer service), linguistic context (interpreting colloquial expressions correctly), and user role context (recognizing legitimate requests versus potential exfiltration).

DeepKeep outperforms competitors on multilingual accuracy. Japanese shows the largest gap, with DeepKeep achieving over 90% accuracy compared to 60-70% from competitors. The platform avoids translation-based approaches that lose cultural and linguistic nuances, which has driven strong sales in Japan.

Deployment options include cloud, on-premises, and air-gapped environments. The AI Firewall supports inline operation (proxy mode) or out-of-band monitoring. Clients pay model providers directly, with DeepKeep remaining outside the transaction flow.

DeepKeep is a 45-person startup competing against larger acquired companies. Enterprise customers report fewer false positives and false negatives compared to competing solutions.

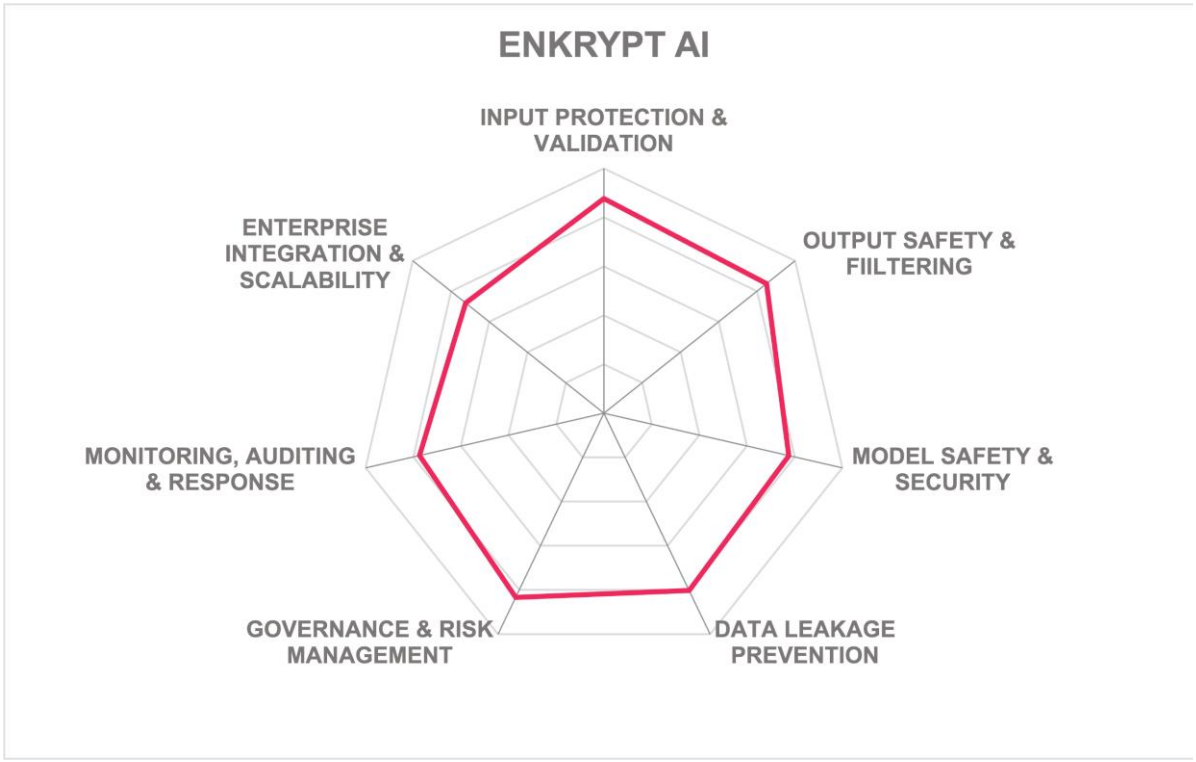
Strengths

- Full model lifecycle protection from R&D through deployment
- Multimodal coverage including physical-world attacks on computer vision systems
- Model-agnostic platform supporting diverse AI types
- Strong agentic AI and MCP server security capabilities
- Context-aware protection spanning application, linguistic, and user role dimensions
- Superior multilingual accuracy, particularly for non-Western languages
- Flexible deployment including air-gapped and out-of-band options
- ISO 27001 certified

Challenges

- A 45-person startup competing against well-resourced companies & acquisitions
- Recent seed funding (May 2024) means limited operational runway compared to established competitors
- SIEM integration capabilities and protocols were not detailed
- Multimodal and computer vision capabilities, while a differentiator today, may become table stakes in a rapidly maturing market

Enkrypt AI – Enkrypt AI



Leader in

Enkrypt AI was founded in 2022 and is headquartered in Boston, Massachusetts. The Enkrypt AI Security and Compliance Platform is delivered as a unified platform and suite of integrated services. The solution supports all deployment options including on-premises installation, cloud service, and managed service. Pricing uses subscription and usage-based models tailored to deployment size and enterprise integration requirements.

Enkrypt AI integrates with identity providers including Microsoft Active Directory, Entra ID, LDAP, and IDaaS providers, supporting SAML federated authentication and OIDC protocols. Authentication methods include OAuth2, JWT, key exchange, token-based authentication, and authenticator apps. The platform provides PII redaction with dynamic unmasking based on user roles, indicating role-based access control capabilities with full user attribution logged for all AI interactions.

Enkrypt AI integrates with DLP solutions including Microsoft Purview, Google Cloud DLP, Netskope, Symantec/Broadcom, Forcepoint, Digital Guardian, and Palo Alto Networks. Data catalog integration includes Varonis, Fortra Digital Guardian, ManageEngine DataSecurity Plus, Satori, Imperva, OneTrust, Forcepoint, Atlan, and Securiti. SIEM integration is provided

through syslog and REST APIs. The platform provides integration through REST APIs, JSON, XML, and WebSockets. Data protection features include PII redaction with dynamic unmasking, prevention of sensitive training data extraction through output guardrails, and detection of data exfiltration attempts through inference or memorization attacks.

Enkrypt AI maintains continuous adversarial intelligence gathering with over three hundred attack patterns targeting prompt injection, jailbreaks, data leakage, function misuse, and encoding-based exploits. The platform uses autonomous agents simulating adversarial attacks on generative AI systems. Red teaming capabilities cover multimodal attacks including text, images, and audio, with agentic detection of tool misuse, function call chaining exploits, and reasoning manipulation.

Enkrypt AI employs a hierarchical, multi-label classification system to detect and categorize prompt manipulation attempts by attack type, severity, intent, and potential impact. Detection capabilities cover prompt injection attacks, jailbreaking attempts, indirect prompt injection, multi-step and chain-of-prompt attacks, and language-specific attacks. Documentation is available in English, German, French, Spanish, Chinese, Japanese, Italian, Russian, Arabic, and Indian languages.

Content safety includes multimodal filtering across text, images, and audio for inappropriate content, violence, hate symbols, and policy violations. Industry-specific content safety requirements are supported for healthcare, finance, manufacturing, and education sectors. The platform detects harmful instructions, toxicity, bias, and hallucinations. Agent-specific protection covers tool misuse, function call chaining exploits, and reasoning manipulation.

Compliance reporting supports GDPR, HIPAA, PCI DSS, ISO 27001, NIST Cybersecurity Framework, and ISO 42001, generating audit-ready reports aligned with regulatory frameworks including EU AI Act risk classifications. Enkrypt AI maintains ISO 27001 certification and SOC 2 Type 1 and Type 2 attestations.

Enkrypt AI suits organizations requiring multimodal AI security across text, image, and audio modalities with extensive integration requirements.

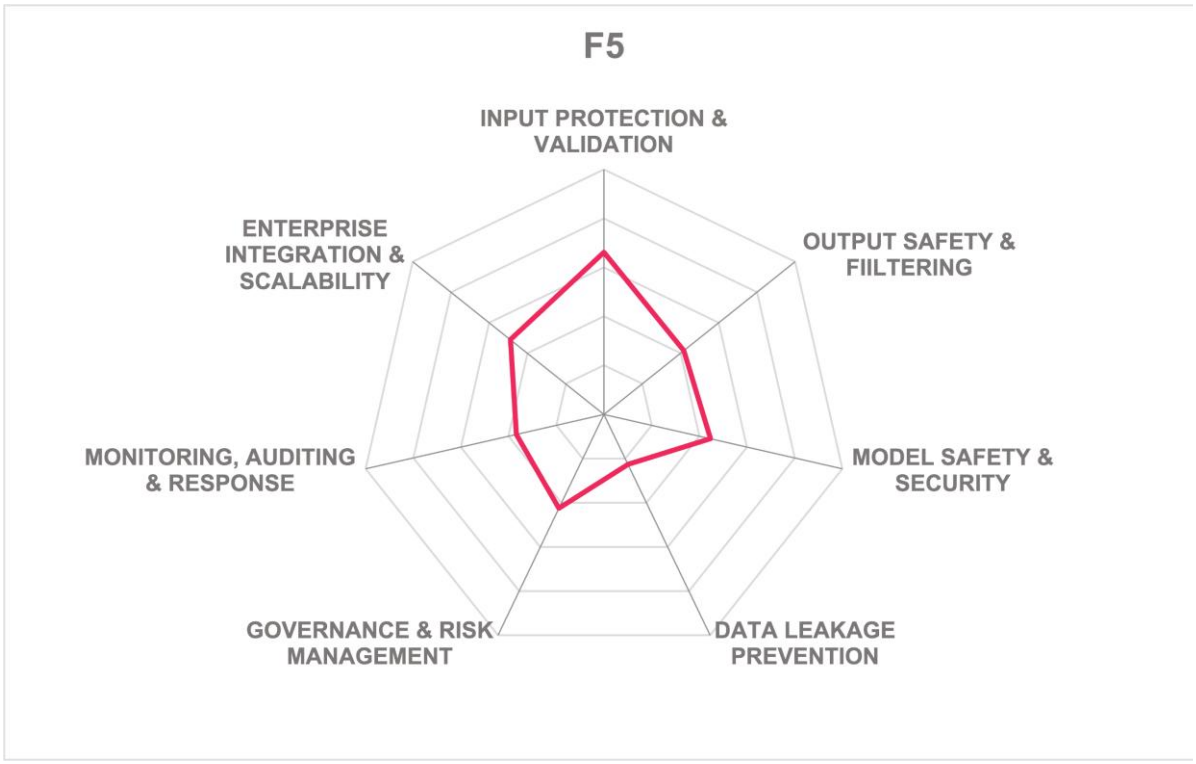
Strengths

- Multi-layered protection across text, images, and audio
- 300+ attack patterns in adversarial prompt library
- Zero-day attack detection capabilities
- Notable security research in agent red teaming and guardrails
- Integrations with many DLP solutions
- Supports all major compliance frameworks
- All deployment models supported
- Context-aware guardrails for complex attacks

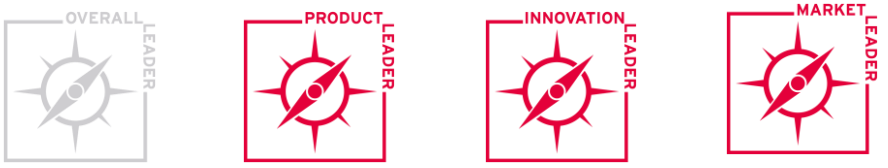
Challenges

- Scaling challenges in a fast-moving market
- No AI model tamper detection
- No content watermarking or attribution
- No APAC or LATAM presence

F5 – GenAI Defense Platform



Leader in



F5, Inc. was founded in 1996 and is publicly traded on NASDAQ (FFIV), headquartered in Seattle, Washington. In 2025, F5 acquired Calypso AI as part of a broader acquisition strategy that included Mantis Net, Fletch, and Leak Signal. The Calypso acquisition brought a mature AI security platform that F5 has rebranded as F5 AI Security, comprising two integrated offerings: F5 AI Guardrails for defensive capabilities and F5 AI Red Team for offensive security testing. The platform supports multiple deployment models including on-premises software, purpose-built appliances, SaaS, and managed services. Pricing combines per-gateway fees with volume-based API call charges.

F5 AI Guardrails provides out-of-box compliance presets for EU AI Act requirements along with restrictions for medical, financial, and legal advice scenarios. Organizations can create custom guardrails using natural language definitions rather than technical policy syntax. The platform operates as a model-agnostic solution supporting all LLMs including fine-tuned variations and private deployments.

A privacy-focused architecture runs the proxy layer asynchronously without outbound calls to third-party cloud services. F5 AI Guardrails integrates with F5's Application Delivery and Security Platform while maintaining compatibility with third-party tools for uniform policy

management across AI deployments. Shadow AI detection capabilities include programmable coaching for users accessing unauthorized AI services, so organizations can guide rather than block employee AI usage.

F5 AI Red Team provides offensive security testing through two methodologies. Signature attacks draw from a library of over 10,500 monthly updated attack prompts based on current threat intelligence, providing one-shot vulnerability assessments against known attack patterns. Agentic Warfare testing conducts multi-turn persuasion attacks using techniques such as Crescendo, Frame, and Trolley to simulate sophisticated adversaries who attempt to coax sensitive information from AI systems over extended interactions.

F5 has introduced novel metrics including AWR (Agentic Warfare Resistance), which measures a system's ability to resist persuasion over time, and CASSI scores based on severity, complexity, and defensive breaking points. Testing generates reports within two hours compared to traditional consultancy engagements that typically require four weeks.

Identity integration includes F5 BIG-IP Access Policy Manager within the broader ADSP ecosystem. Built-in PII detection and redaction operates on both inputs and outputs with customizable data protection policies. SIEM connectivity supports log exports and API access for security operations integration.

F5 AI Security suits organizations seeking combined defensive guardrails and offensive testing capabilities with flexible deployment options and privacy-focused architecture.

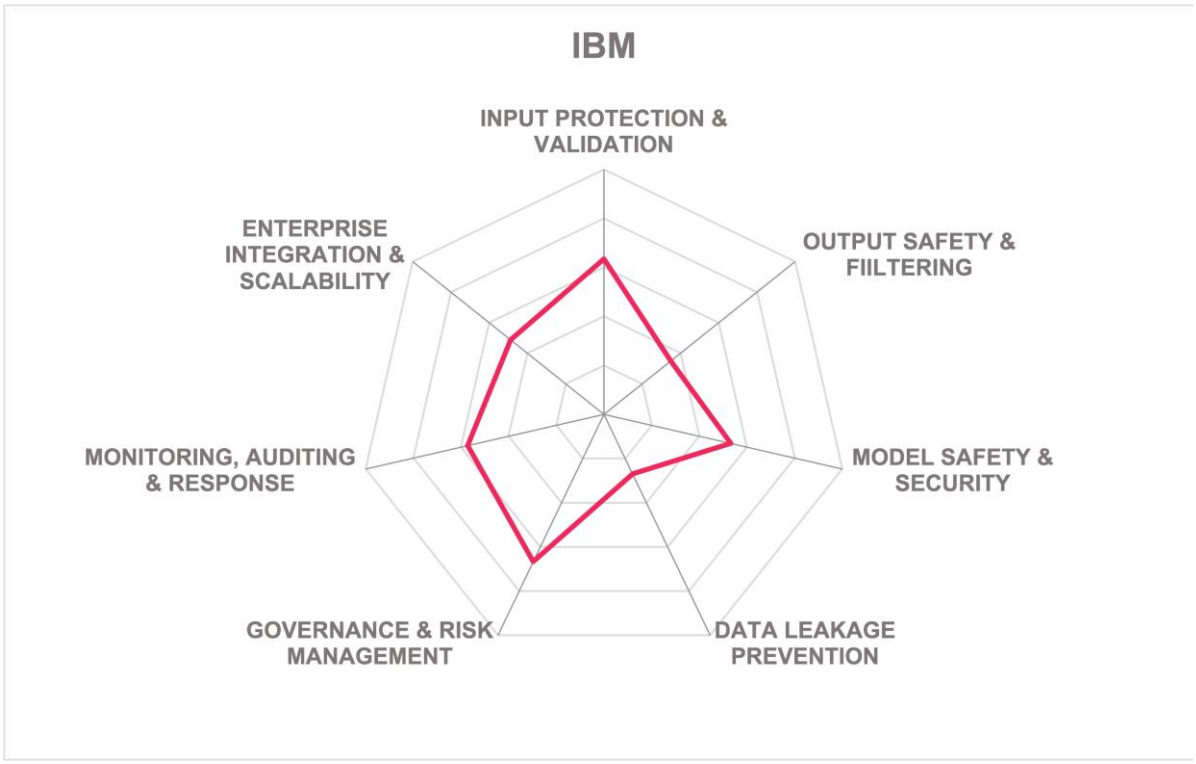
Strengths

- AI Red Team with 10,500+ signature attacks and multi-turn Agentic Warfare testing
- Novel metrics (AWR, CASSI) for quantifying AI security posture
- Out-of-box EU AI Act compliance presets and custom guardrail creation via natural language
- Shadow AI detection with programmable user coaching
- Privacy-focused offline proxy architecture without third-party cloud dependencies
- Model-agnostic support across all LLMs including private deployments

Challenges

- Deployment may require multiple F5 products with complex licensing arrangements
- Third-party integrations beyond F5 ecosystem require verification
- Does not perform real-time scanning of file uploads to GenAI systems
- Remediation options beyond alerting require additional configuration

IBM – Guardium AI Security



Leader in

IBM (founded 1911, publicly traded) launched Guardium AI Security as a SaaS solution in May 2025. The platform is delivered through an OEM partnership with AllTrue.ai for core technology, enhanced with IBM's enterprise security expertise and integrated with IBM watsonx.governance and IBM Guardium Data Security Posture Management (DSPM). The platform is deployed exclusively as SaaS using proxies and SDKs. Licensing follows a per-AI-use-case model, which can make cost predictions difficult for organizations with diverse AI deployments.

The platform supports authentication through major authenticator apps, time-based one-time passwords (TOTP), and federation via SAML and OIDC protocols, with integration capabilities for major enterprise SSO providers. Data protection features include automatic PII detection and redaction in both input and output, support for major file formats through text conversion, and output guardrails to prevent sensitive training data extraction. SIEM integration is provided through syslog and REST APIs. Integration with IBM Guardium DSPM for enhanced data classification capabilities is planned on the product.

Guardium AI Security uses AI gateway technology acting as either a proxy or SDK to analyze generative AI prompts in real-time, combining AI models and libraries to address prompt attack vectors. Security policies are configurable at organization or project level to block, monitor, or redact suspicious content. The platform performs automated penetration testing with over 2,700 customizable tests integrated into CI/CD pipelines. The platform maintains a library of known attack patterns and identifies threats through customer-defined categories and policies.

The solution protects against direct prompt injection attacks, code injections, jailbreaking attempts, and PII exposure through customizable security policies. Real-time detection capabilities differentiate between malicious prompts and legitimate edge cases.

Model integrity protection includes tamper detection, vulnerability scanning for third-party models and libraries, anomalous behavior detection, and supply chain validation through signature and hash tracking. The system maintains an internal inventory of third-party components and AI libraries, exposing security vulnerability information for supply chain risk management.

Guardium AI Security provides automated compliance reporting for AI regulatory frameworks including ISO 42001, PCI DSS, HIPAA, Sarbanes-Oxley, FISMA, financial reporting standards, and GDPR. The platform's compliance automation converts requirements from multiple AI risk management frameworks into automated workflow steps. The centralized governance dashboard provides risk classification capabilities and automatically generates compliance reports. The system tracks compliance violations with remediation actions and maintains chain-of-custody requirements for forensic investigations.

Guardium AI Security suits IBM-centric enterprises deeply invested in the IBM ecosystem who prioritize compliance automation and governance.

Strengths

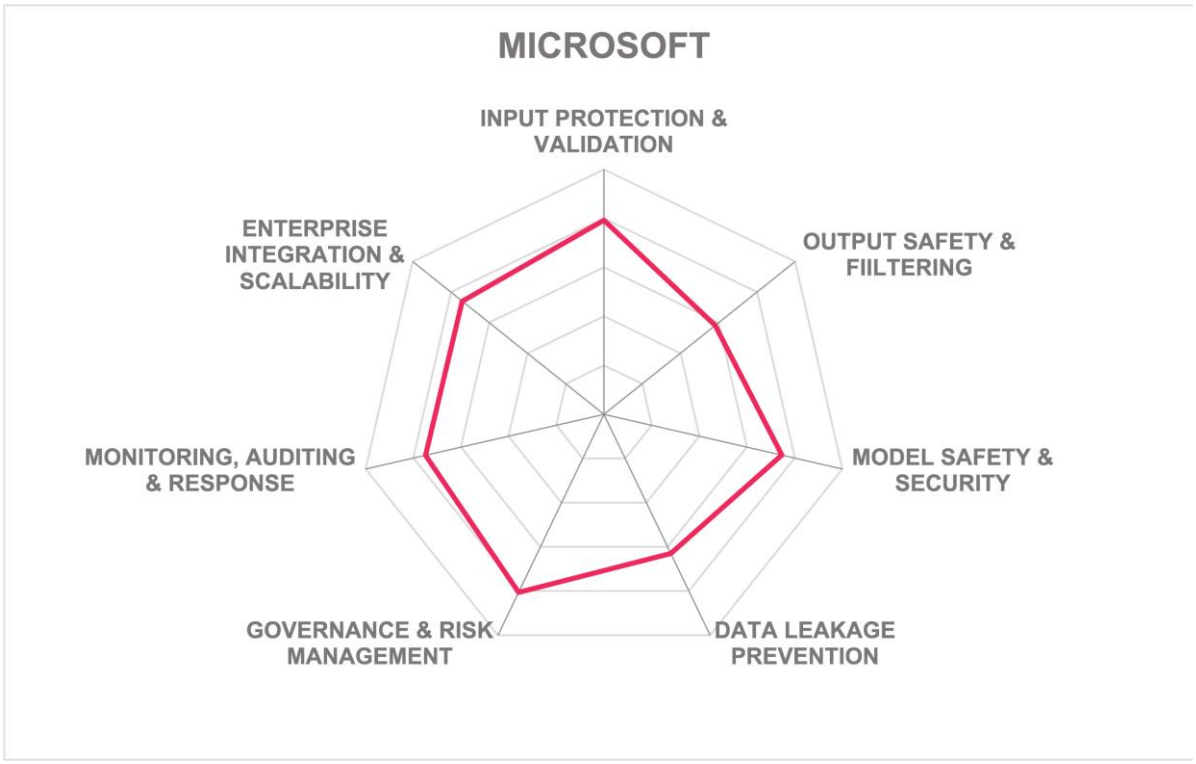
- Massive global presence
- Industry-leading compliance framework support with pre-built templates
- Robust vulnerability assessment with 2,700+ configurable tests
- Immutable audit logging through embedded OpenSearch
- Hierarchical organization and project-level policy controls
- Automated daily scanning of models and AI assets
- Flexible API-based architecture supporting both proxy and SDK deployment models
- Integration with watsonx.governance
- IBM has considerable maturity and offers a suite of capabilities across the entire AI pipeline

Challenges

- Very recent offering
- Heavy reliance on OEM partnership with AllTrue.ai
- SaaS-only deployment model

- Integration capabilities are unproven outside of the extensive IBM ecosystem

Microsoft – Microsoft Security for AI



Leader in

Microsoft Corp. (founded 1975, NASDAQ: MSFT) offers Microsoft Security for AI, encompassing Microsoft Purview, Entra, and Defender protecting generative AI systems. The solution is deployed as cloud services with managed service options. Licensing follows per-user/device subscription and consumption-based pricing.

Microsoft Security for AI integrates with Entra ID, supporting JWT, OAuth2, OIDC, SAML, and Kerberos. Multi-factor authentication supports biometrics, authenticator apps, smartcards, and hardware tokens. The platform supports zero-trust principles through Just-In-Time access and privileged identity management. SSO integration includes Microsoft, Okta, Ping, Google, OneLogin, Auth0, JumpCloud, Keycloak, PingFederate, and Duo.

Native DLP is provided through Microsoft Purview. Third-party integration supports Netskope and Palo Alto Networks. Data catalog integration includes Microsoft Purview's Unified Data Map, Varonis, and Fortra Digital Guardian, with Open APIs for custom connectors. SIEM

connectivity supports syslog, CEF, and REST protocols with native Microsoft Sentinel integration.

Microsoft detects jailbreak and indirect prompt injection attacks (XPIA) through fine-tuned Small Language Models (SLMs) with LLM monitoring and egress vector limiting. Threat intelligence is updated through MSRC Bug Bounty, AI Red Teaming, Microsoft Research, social media monitoring, and enterprise penetration tests. The platform maintains prompt-based attack pattern classification with separate jailbreak and XPIA detection models retrained with updated data. AI model scanner services detect threats, vulnerabilities, embedded secrets, and malware throughout the model lifecycle. Capabilities include automated AI Red Teaming, multi-language attack detection, and zero-day prompt engineering attack detection.

The solution provides real-time detection and blocking of prompt injection, jailbreaking, and indirect prompt injection attacks. Microsoft's XPIA detector neutralizes indirect attacks while detecting multi-step and chain-of-prompt attacks. Data protection includes hundreds of PII and sensitive data classifiers, sensitivity label preservation, and prevention of data exfiltration through inference attacks. Content safety uses configurable SLMs detecting harmful content across text, code, and image modalities with customizable thresholds and pre-built filters.

The platform generates compliance reports through Purview Compliance Manager. Pre-built frameworks support EU AI Act, NIST AI Risk Management Framework, ISO/IEC 42001, Canada's AIDA, Brazil's EBIA, and India's Digital India Act.

Microsoft Security for AI suits organizations heavily invested in the Microsoft ecosystem seeking unified security management.

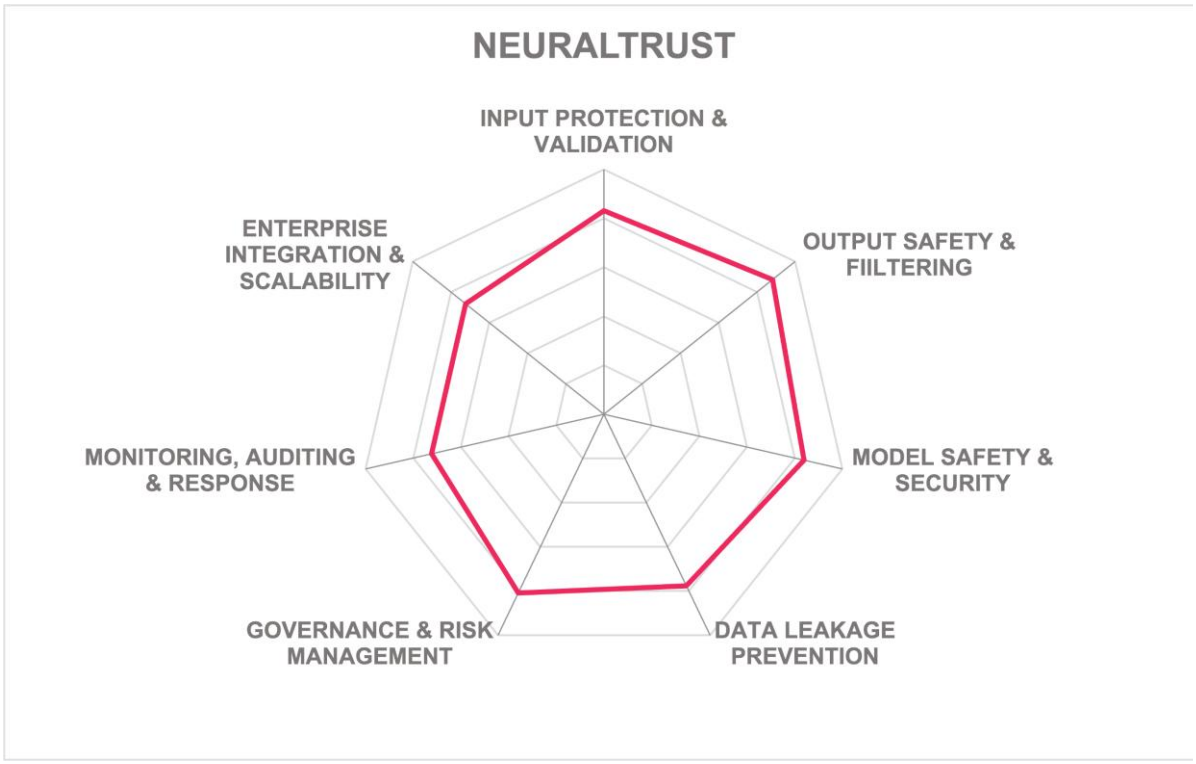
Strengths

- Massive global customer base
- Native integration across Microsoft's entire security ecosystem
- Pre-built frameworks for global regulations (EU AI Act, NIST, ISO 42001)
- Confidential computing with GPU/CPU support
- Forensic-grade immutable audit logging with 10-year retention options
- Homomorphic encryption for secure inference
- Continuous threat intelligence from multiple sources including Bug Bounty programs
- WORM-enforced audit immutability

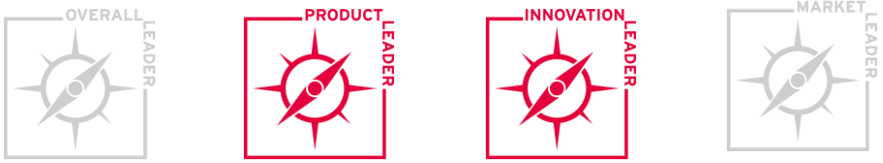
Challenges

- Complexity of integration and configuration
- Opaque cross-product dependencies
- AI explainability and governance
- Operational overhead in multi-cloud scenarios

NeuralTrust – NeuralTrust



Leader in



NeuralTrust was founded in 2022 and is a venture capital-backed startup based in Spain. The NeuralTrust platform can be deployed on-premises, in private cloud, and public cloud environments. Licensing combines fixed costs with transaction-based pricing.

NeuralTrust provides enterprise integration through OAuth2 and SSO support, with compatibility across major identity providers including Microsoft, Okta, Ping, Google, OneLogin, Auth0, JumpCloud, Keycloak, PingFederate, and Duo. Federated identity

management supports SAML, OpenID Connect, Kerberos, and Remote Authentication protocols.

The platform integrates with major DLP solutions including Microsoft Purview, Google Cloud DLP, Netskope, Symantec/Broadcom, Forcepoint, Digital Guardian, and Palo Alto Networks. SIEM integration is provided through syslog, CEF, and REST APIs.

NeuralTrust maintains an attack catalog updated based on OWASP LLM Top 10, MITRE ATLAS, and internal research. Detection capabilities span multiple languages including English, Spanish, French, German, Chinese, and Portuguese using multi-layered validation to identify sophisticated multi-turn and obfuscated attacks. The platform classifies over thirty injection types including Role-Playing Exploits, Token Smuggling, JSON Injection, Reverse Psychology, and Symbolic Encoding. The Model Scanner performs static analysis of models, datasets, and code artifacts before deployment.

NeuralTrust offers real-time prompt injection detection with claimed sub-10ms latency on GPU, detecting and blocking jailbreaking attempts, indirect prompt injection attacks, and multi-step chain-of-prompt attacks. The built-in DLP engine recognizes over forty sensitive data types including EU-specific identifiers (DNI, NIR), API keys, and financial credentials. Protection against model extraction employs behavioral monitoring, rate limiting, and output similarity checks. The platform generates compliance reports with pre-built frameworks for EU AI Act, ISO 42001, and NIST AI RMF.

NeuralTrust suits organizations requiring flexible deployment models including on-premises or private cloud options and companies with strong EU data protection requirements.

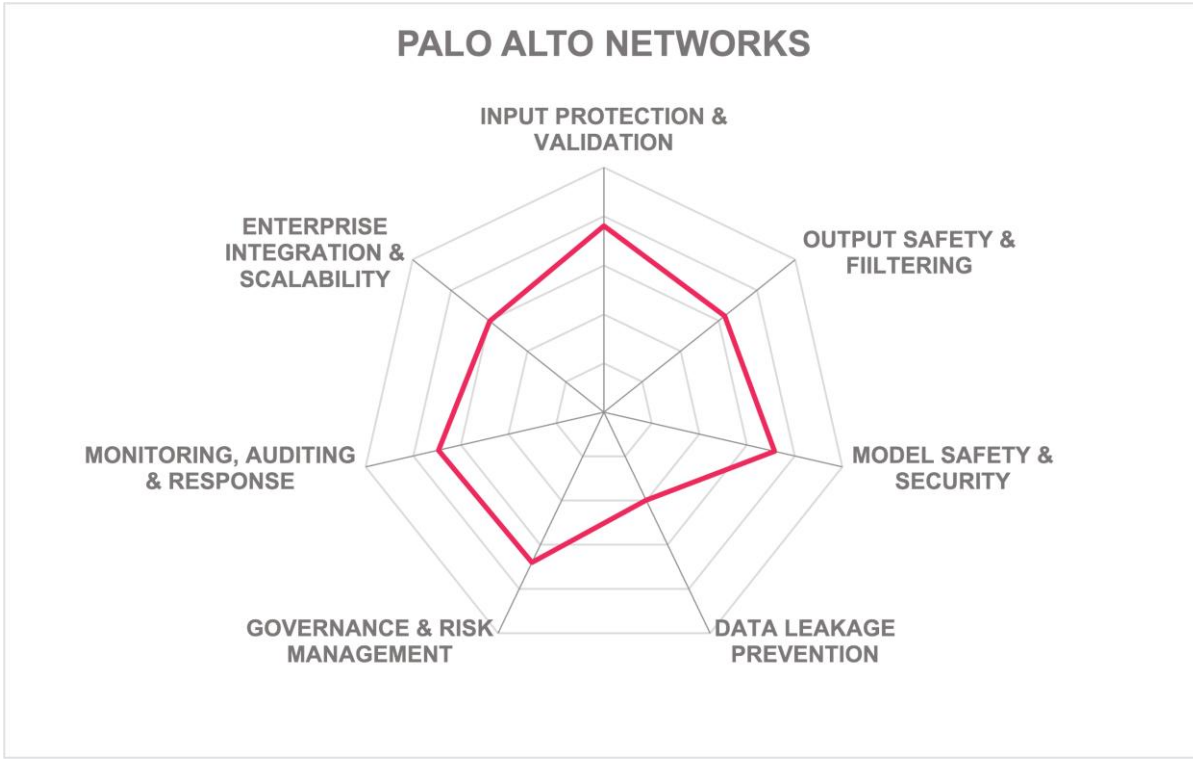
Strengths

- Industry-leading sub-10ms GPU latency
- High-accuracy detection across multiple languages
- Support for all major DLP vendors, identity providers, and SIEM platforms
- Continuously updated attack catalog based on OWASP, MITRE ATLAS, and proprietary research
- Pre-built templates for GDPR, EU AI Act, and European-specific data types
- Purpose-built GenAI security architecture (not adapted from traditional security tools)
- Automated Red Teaming and zero-day detection capabilities
- Flexible deployment across on-premises, private cloud, and public cloud environments
- Real-time detection of prompt injection, jailbreaking, indirect attacks, multi-step chains

Challenges

- Small company with multiple growth challenges
- No built-in MFA options
- Support for wide range of DLP products may present a long-term scalability challenge

Palo Alto Networks – Prisma AIRS Platform



Leader in

Palo Alto Networks, Inc. (founded 2005, NASDAQ: PANW) offers Prisma AIRS (AI Runtime Security), a platform comprising five integrated components protecting generative AI systems throughout their lifecycle. The solution launched approximately nine months ago and reports production deployments with enterprise customers. Prisma AIRS can be deployed as cloud services or software firewalls, with managed service options available.

In July 2024, Palo Alto Networks closed its acquisition of Protect AI, with integration completed in 2.5 months enabled by Protect AI's microservices architecture. Palo Alto put Protect AI's leadership in charge of the combined AI security direction, suggesting commitment to the acquired company's approach rather than just absorbing its technology. The company reports 80,000 existing customers with AI security now featuring in every executive briefing.

Prisma AIRS integrates with Strata Cloud Manager's identity framework, supporting enterprise authentication through major identity providers. The platform enforces role-based access controls and integrates with existing enterprise SSO deployments.

The platform provides native DLP functionality using Palo Alto Networks' enterprise DLP solution with 1,000+ predefined data patterns across 50+ regions, including region-specific identifiers for PII, financial data, and healthcare information. The solution integrates with SIEM platforms through Strata Logging Service using syslog, CEF, and REST protocols, with native support for Palo Alto Networks Cortex XSOAR. The AI Runtime Security firewall represents a superset of Palo Alto's VM/CN series firewalls. Prisma AIRS maintains threat intelligence updated through internal research, OWASP LLM Top 10 alignment, and threat feeds from WildFire (Palo Alto's malware analysis platform).

The AI Red Teaming module provides 1,000+ pre-built attacks plus agentic testing, executing dynamic attack simulations that interpret natural language attack goals and automatically iterate based on application responses. Detection capabilities address 28+ prompt injection attack types and 31+ jailbreak techniques across eight officially supported languages, with unofficial support claimed for 100+ additional languages. The platform detects attacks including goal hijacking, remote code execution, model DDoS attempts, system prompt leakage, role-playing exploits, token smuggling, JSON injection, reverse psychology, symbolic encoding, and adversarial suffixes.

The Model Scanner performs both static and dynamic analysis of AI models, including sandbox execution to detect obfuscated malicious code within model artifacts such as PyTorch and pickle files. Large enterprises may run 100,000+ models in production, including predictive and fraud detection models beyond generative AI. The platform has native integration with Hugging Face (one of only two companies with this integration), scanning models in CI/CD pipelines and supply chain entry points to detect unauthorized commands, file modifications, and credential theft. Palo Alto cites a fake 23andMe model, downloaded tens of thousands of times, that was designed to steal AWS credentials.

Runtime capabilities include prompt injection and jailbreak attack blocking, sensitive data recognition and redaction, malicious URL detection across seventy-four categories, malware detection using WildFire signatures, toxic content filtering, hallucination detection comparing outputs against RAG context, and natural language policy enforcement. The platform combines network firewall intercept between AI applications and LLMs with API integration to AI gateways.

Strengths

- Unified management across five integrated components through single dashboard
- Covers the entire AI lifecycle from model validation to runtime protection
- Builds on proven technologies (WildFire, Enterprise DLP, Cortex XSOAR)
- Protect AI acquisition brings specialized AI security expertise and native Hugging Face integration
- Built on infrastructure processing 30 billion+ threats and 80 million files daily
- Agent security covers ten platforms across SaaS and cloud deployments

Challenges

- No current support for AI regulations in China, Latin America, or Middle East
- Optimal value requires broader Palo Alto ecosystem adoption, limiting appeal for organizations with different security stacks
- Platform launched nine months ago; acquisition integration completed only recently

Vendors to Watch

Besides the vendors covered in detail in this document, we observe some other companies in the market that readers should be aware of. These vendors did not participate in the rating for various reasons but nevertheless offer a significant contribution to the market space.

1touch.io

1touch.io is applying its deep expertise in sensitive data discovery and contextual intelligence to address one of GenAI's most pressing security challenges: knowing what data you have before you use it to train or ground AI systems. With organizations racing to adopt large language models and generative AI applications, the company's Kontxtual platform provides the foundational visibility that security and governance teams require to prevent sensitive information from inadvertently entering AI pipelines or being exposed through model outputs.

The company's approach goes beyond traditional data discovery. Rather than simply cataloguing where data resides, 1touch.io builds contextual profiles for each sensitive data element, tracking access patterns, lineage, and exposure risk in real time. This multidimensional understanding proves essential when organizations need to determine which datasets are safe for AI training and which contain PII, intellectual property, or confidential information that must be excluded. Kontxtual's 98.6% classification accuracy, validated by the Tolly Group, reduces the false positives that typically plague data security tools and cause alert fatigue. The platform's agentless architecture scans across hybrid environments spanning cloud, SaaS, on-premises systems, and even mainframes without disrupting operations.

Why worth watching: As GenAI adoption accelerates, OWASP's 2025 Top 10 for LLM Applications elevated sensitive information disclosure to the second most critical vulnerability. 1touch.io addresses this risk at its source by providing the data intelligence that organizations need before they begin AI initiatives. Their recent 500% year-over-year growth and zero customer churn suggest that enterprises are recognizing that effective AI governance starts with understanding the sensitive data landscape, not with securing the models themselves.

Akamai

Akamai is leveraging its decades of content delivery network (CDN) and edge computing expertise to enter the Generative AI Defense market. With a massive global infrastructure spanning over 4,000 locations worldwide and a strong enterprise customer base, Akamai is uniquely positioned to provide edge-based AI security services. The company's experience in real-time threat detection and mitigation at internet scale, combined with its existing security portfolio including bot management and API protection, creates a foundation for GenAI security solutions. Akamai's edge-first approach could enable ultra-low latency AI security processing closer to end users.

Why worth watching: Akamai's unparalleled global edge infrastructure and proven ability to process massive volumes of internet traffic in real-time could revolutionize how GenAI security is delivered at scale.

Arkose Labs

Arkose Labs is approaching the Generative AI Defense market from a distinctive angle: protecting AI platforms from the outside-in through advanced bot detection and abuse prevention. With ample funding and proven success securing many of the world's largest B2C enterprises across financial services, gaming, and social media sectors, Arkose Labs brings battle-tested expertise in detecting and neutralizing sophisticated bot attacks at massive scale. Their Arkose GPT Protection specifically targets unauthorized AI platform access, preventing bot armies from scraping AI outputs, creating illegal reverse proxies, and bypassing geographic restrictions. With established integrations across major CIAM vendors (Okta, Ping Identity, ForgeRock) and CDN providers (Akamai, Cloudflare, Fastly), Arkose Labs has the infrastructure partnerships to rapidly extend their platform-level protections deeper into AI application security.

Why worth watching: As AI platforms become increasingly valuable targets for bot-driven attacks, credential stuffing, and automated abuse, Arkose Labs' proven ability to distinguish genuine users from sophisticated bots at internet scale positions them to bridge the gap between traditional platform security and emerging GenAI-specific threats, creating a “perimeter-to-content" defense strategy that addresses both access abuse and content manipulation.

Cequence Security

Cequence Security is leveraging its API security and bot management expertise to address GenAI security challenges at the application layer. With proven capabilities in detecting automated attacks, API abuse, and malicious bot behavior across enterprise applications, Cequence possesses foundational technologies applicable to GenAI security. The company's experience in analyzing API traffic patterns, detecting anomalous behavior, and preventing automated attacks provides a technical foundation for identifying prompt injection attempts, jailbreak attacks, and other GenAI-specific threats that manifest through API interactions. Cequence's existing enterprise customer base in financial services, e-commerce, and healthcare sectors represents organizations now deploying GenAI applications that require security controls. The company's unified API and bot protection platform could extend to cover GenAI APIs, providing visibility into AI model interactions and detecting attacks that exploit API endpoints.

Why worth watching: Cequence's deep expertise in API-level threat detection and behavioral analysis positions the company to address GenAI security threats that exploit application interfaces, particularly as enterprises expose AI capabilities through APIs requiring the same rigor as traditional API security.

Cerbos

Cerbos is leveraging its authorization and access control expertise to address the complex permissions challenges emerging in GenAI applications and autonomous AI agents. With proven capabilities in policy-based access control and context-aware authorization decisions, Cerbos addresses a critical GenAI security gap: managing dynamic permissions for AI systems that make decisions and take actions on behalf of users. The company's developer-friendly approach to implementing fine-grained access control, combined with its ability to handle complex authorization logic through policy definitions, positions Cerbos to solve the "who can do what" problem for AI agents accessing enterprise resources and APIs. As organizations deploy agentic AI systems that interact with multiple data sources, tools, and APIs, the authorization layer becomes critical for preventing unauthorized access and ensuring AI actions remain within appropriate boundaries. Cerbos's stateless architecture and support for real-time authorization decisions align with the performance requirements of interactive AI applications.

Why worth watching: Cerbos's specialization in context-aware authorization and policy-based access control directly addresses the emerging challenge of managing permissions for autonomous AI agents, a problem traditional access control systems were not designed to solve and one that becomes more critical as agentic AI deployments scale.

Cloudflare

Cloudflare is leveraging its massive global network infrastructure and position as a critical internet services provider to enter the GenAI security market. With traffic from millions of websites and applications flowing through its network, Cloudflare possesses unparalleled

visibility into internet-scale attack patterns and threat intelligence. The company's existing security portfolio including DDoS protection, Web Application Firewall (WAF), bot management, and Zero Trust services provides a foundation for extending protection to GenAI applications. Cloudflare Workers AI and AI Gateway products demonstrate the company's commitment to the AI infrastructure space, and security capabilities built into these offerings could evolve into comprehensive GenAI security solutions. The company's edge computing platform enables low-latency security processing close to users, addressing performance concerns for real-time GenAI security enforcement. Cloudflare's massive customer base across enterprises, startups, and web properties represents organizations now deploying GenAI features requiring security controls.

Why worth watching: Cloudflare's unique position as a global internet infrastructure provider with visibility into attack patterns at massive scale, combined with its expanding AI platform offerings, could enable the company to deliver GenAI security as a seamlessly integrated service within its existing security and infrastructure stack, reducing complexity for customers already relying on Cloudflare.

Concentric AI

Concentric AI is leveraging its autonomous data security posture management (DSPM) expertise to address the critical challenge of data security in GenAI deployments. With proven capabilities in discovering, classifying, and protecting sensitive data across enterprise environments using autonomous semantic analysis, Concentric addresses a foundational GenAI security requirement: ensuring sensitive data does not inadvertently train models or leak through AI interactions. The company's deep learning approach to understanding data context and risk, without requiring pre-defined rules or patterns, aligns with the complexity of protecting unstructured data used in GenAI applications. Concentric's existing enterprise deployments in financial services, healthcare, and regulated industries represent organizations facing stringent data protection requirements for AI initiatives. The platform's ability to continuously discover and classify data across cloud, on-premises, and SaaS environments addresses the distributed nature of data used in GenAI pipelines.

Why worth watching: Concentric's autonomous approach to data discovery and classification directly addresses one of the most critical GenAI security challenges—preventing sensitive data exposure—and could provide the data security foundation required before organizations can safely deploy GenAI applications at scale.

CrowdStrike

CrowdStrike, a leader in endpoint detection and response (EDR) and cybersecurity, is expanding its security platform to address the emerging threats in the Generative AI Defense market. With its proven cloud-native Falcon platform already protecting millions of endpoints globally and its advanced threat intelligence capabilities, CrowdStrike is well-positioned to extend its expertise into AI security. The company's experience in real-time threat detection, behavioral analysis, and automated response mechanisms provides a strong foundation for developing AI-specific security solutions. CrowdStrike's established enterprise relationships

and security ecosystem could accelerate adoption of GenAI security controls across their existing customer base.

Why worth watching: CrowdStrike's market-leading threat intelligence and behavioral analysis capabilities, combined with their massive installed base, could rapidly establish them as a dominant force in enterprise GenAI security.

DataKrypto

DataKrypto is leveraging cryptographic innovation and privacy-enhancing technologies to address data protection challenges in GenAI deployments. With expertise in encryption, tokenization, and secure computation techniques, DataKrypto addresses a critical GenAI security requirement: protecting sensitive data throughout AI pipelines while maintaining model functionality. The company's focus on enabling secure data sharing and collaboration through cryptographic controls aligns with enterprise needs to use sensitive data for AI training and inference without exposing plaintext information. As organizations face regulatory requirements around data protection in AI systems (EU AI Act, GDPR, HIPAA), cryptographic approaches that enable "secure by design" AI deployments become increasingly valuable. DataKrypto's technology could enable scenarios where sensitive data remains encrypted even during model training or inference, addressing both security and compliance requirements. The company's approach may be particularly relevant for highly regulated industries including healthcare, financial services, and government sectors deploying GenAI applications with strict data protection mandates.

Why worth watching: DataKrypto's focus on cryptographic data protection techniques could provide the technical foundation for "privacy-preserving AI" deployments, enabling organizations to leverage sensitive data in GenAI applications while maintaining compliance with stringent data protection regulations and security requirements that traditional security controls cannot fully address.

HiddenLayer

HiddenLayer brings practical experience with real-world adversarial AI attacks to the GenAI security space. HiddenLayer won the RSA Conference 2023 Innovation Sandbox Contest and has secured Fortune 100 customers across government and finance sectors. The company's AISec Platform addresses supply chain security, runtime defense, posture management, and automated red teaming for generative, predictive, and agentic AI applications. With 25 granted patents and strategic backing from Microsoft, IBM, Capital One, and Booz Allen ventures, HiddenLayer demonstrates both technical depth and enterprise credibility in a crowded market.

Why worth watching: HiddenLayer's founders gained direct experience defending against adversarial ML attacks before founding the company, providing practical insight into emerging threats. Their lifecycle approach and significant venture backing from major enterprise technology players could position them as a serious contender as the GenAI security market matures beyond point solutions.

Lakera

Lakera is an emerging specialist focused specifically on LLM security and prompt injection defense. Founded by security researchers with academic backgrounds in adversarial machine learning, Lakera brings deep technical expertise in AI-specific attack vectors to the GenAI security market. The company's flagship product, Lakera Guard, provides prompt injection detection and content moderation specifically designed for LLM applications, addressing core GenAI security threats with purpose-built technology rather than adapting traditional security approaches. Lakera's research-driven approach and focus on continuously updating defenses against evolving prompt engineering techniques demonstrates commitment to staying ahead of emerging threats. The company has published research on LLM vulnerabilities and attack techniques, contributing to the broader security community's understanding of GenAI risks. Lakera's API-first deployment model aligns with developer workflows for integrating security into GenAI applications, and early customer traction demonstrates market validation for dedicated LLM security solutions. As a venture-backed startup focused exclusively on GenAI security, Lakera can iterate rapidly and specialize deeply without the constraints of broader security portfolios.

Why worth watching: Lakera's singular focus on LLM security, combined with research-driven threat intelligence and purpose-built detection capabilities, positions the company as a specialist that could define best practices for prompt injection defense and establish technical leadership in GenAI-specific threat detection, particularly as prompt engineering attacks continue to evolve in sophistication.

Nutanix

Nutanix is leveraging its hyperconverged infrastructure (HCI) leadership and enterprise cloud platform expertise to enable secure AI deployment through integrated infrastructure solutions. With a strong enterprise customer base running mission-critical workloads on Nutanix infrastructure and extensive partnerships across the AI ecosystem, the company is positioned to provide "AI in a box" solutions that bundle secure infrastructure with GenAI security capabilities. Nutanix's experience in hybrid and multi-cloud environments, combined with its focus on simplified operations and built-in security controls, creates opportunities to deliver turnkey secure AI platforms. The company's infrastructure-first approach addresses security concerns at the deployment layer, potentially offering validated reference architectures where GenAI security tools run on hardened, pre-configured infrastructure designed for AI workloads. While Nutanix does not currently provide dedicated GenAI threat detection capabilities such as prompt injection prevention or jailbreak detection, its infrastructure platform serves as the foundation where third-party GenAI security vendors can deploy their solutions.

Why worth watching: Nutanix's trusted position in enterprise infrastructure and ability to deliver integrated, validated solutions could accelerate GenAI security adoption by reducing deployment complexity and providing customers with pre-integrated security stacks on proven infrastructure, particularly in highly regulated industries requiring on-premises or private cloud AI deployments.

PlainID

PlainID is leveraging its decade of policy-based access control expertise to address the authorization and data governance dimension of Generative AI Defense. PlainID brings mature identity governance technology to the AI space. The company's platform provides dynamic, policy-driven control over who can access AI systems, what data those systems can retrieve, and how responses are masked based on user entitlements—treating GenAI as another enterprise resource requiring authorization. PlainID's three-stage approach (prompt control, data filtering, response masking) offers organizations a way to enforce consistent access policies across AI applications, particularly valuable for protecting sensitive data in RAG implementations and AI-powered analytics platforms.

Why worth watching: As the GenAI security market matures and organizations move beyond point-in-time threat detection toward governance frameworks, PlainID's specialized focus on authorization and policy-based data access control could fill a critical gap in defense-in-depth architectures, particularly for enterprises already invested in policy-driven security models.

Sentra

Sentra is leveraging its modern Data Security Posture Management (DSPM) platform to address the data governance dimension of Generative AI Defense. Founded in 2021, the company provides a cloud-native platform focused on data discovery, classification, and access governance across cloud environments. Sentra's unique value proposition for AI security centers on pre-ingestion data classification: helping organizations understand and control what sensitive data is accessible to AI systems before it enters training pipelines or RAG implementations. The platform's strong integration ecosystem with major DLP solutions (Google Cloud DLP, Netskope, Symantec, Forcepoint, and Palo Alto Networks) and 100% cloud-native architecture positions it to provide visibility into the data landscape that AI applications operate within, treating data governance as the foundation for AI security rather than focusing on runtime threat detection.

Why worth watching: As enterprises recognize that controlling what data AI systems can access is as critical as monitoring what they output, Sentra's data-centric approach to AI governance, classifying and controlling sensitive information before AI ingestion, could become a foundational component in defense-in-depth strategies, particularly as regulations increasingly focus on data provenance and AI training data governance.

Straiker

Straiker is an emerging player in the Generative AI Defense space, positioning itself as a specialized security solution for protecting AI systems from prompt-based attacks and model vulnerabilities. The company focuses on developing advanced detection mechanisms for prompt injection, jailbreaking, and other AI-specific threats through sophisticated machine learning approaches. As a newer entrant to the market, Straiker is building its platform with a dedicated focus on the unique security challenges posed by generative AI systems. The

company represents the growing trend of purpose-built AI security solutions designed specifically for the complexities of modern generative AI deployments.

Why worth watching: Straiker's singular focus on AI-specific security threats without legacy technology constraints could enable more innovative and effective approaches to generative AI protection than traditional security vendors adapting existing solutions.

Related Research

Leadership Compass: Container Security

Leadership Compass: API Security and Management

Leadership Compass: Data Security Platforms

Leadership Compass: Email Security

Whitepaper: Navigating the Future of Authentication in the Age of AI and Deepfakes

Advisory Note: Security Organization Governance and the Cloud

Advisory Note: Cloud Services and Security

Blog: Cloud Security Alphabet Soup

Copyright

KuppingerCole Analysts supports IT professionals with exceptional expertise to define IT strategies and make relevant decisions. As a leading analyst firm, KuppingerCole offers firsthand, vendor-neutral information. Our services enable you to make decisions crucial to your business with confidence and security.

Founded in 2004, KuppingerCole is a global, independent analyst organization headquartered in Europe. We specialize in providing vendor-neutral advice, expertise, thought leadership, and practical relevance in Cybersecurity, Digital Identity & IAM (Identity and Access Management), Cloud Risk and Security, and Artificial Intelligence, as well as technologies enabling Digital Transformation. We assist companies, corporate users, integrators, and software manufacturers to address both tactical and strategic challenges by making better decisions for their business success. Balancing immediate implementation with long-term viability is central to our philosophy.

For further information, please contact clients@kuppingercole.com..