SECURING AI/LLMS IN 2025:

A Practical Guide To Securing & Deploying AI

Detailed product and vendor guide on securing AI
for practitioners and security leaders

March 2025

Software Analyst®
Cyber Research

Software Analyst®
Cyber Research

LinkedIn (Business)
https://www.linkedin.com/company/software-analyst/

Substack
https://softwareanalyst.substack.com/

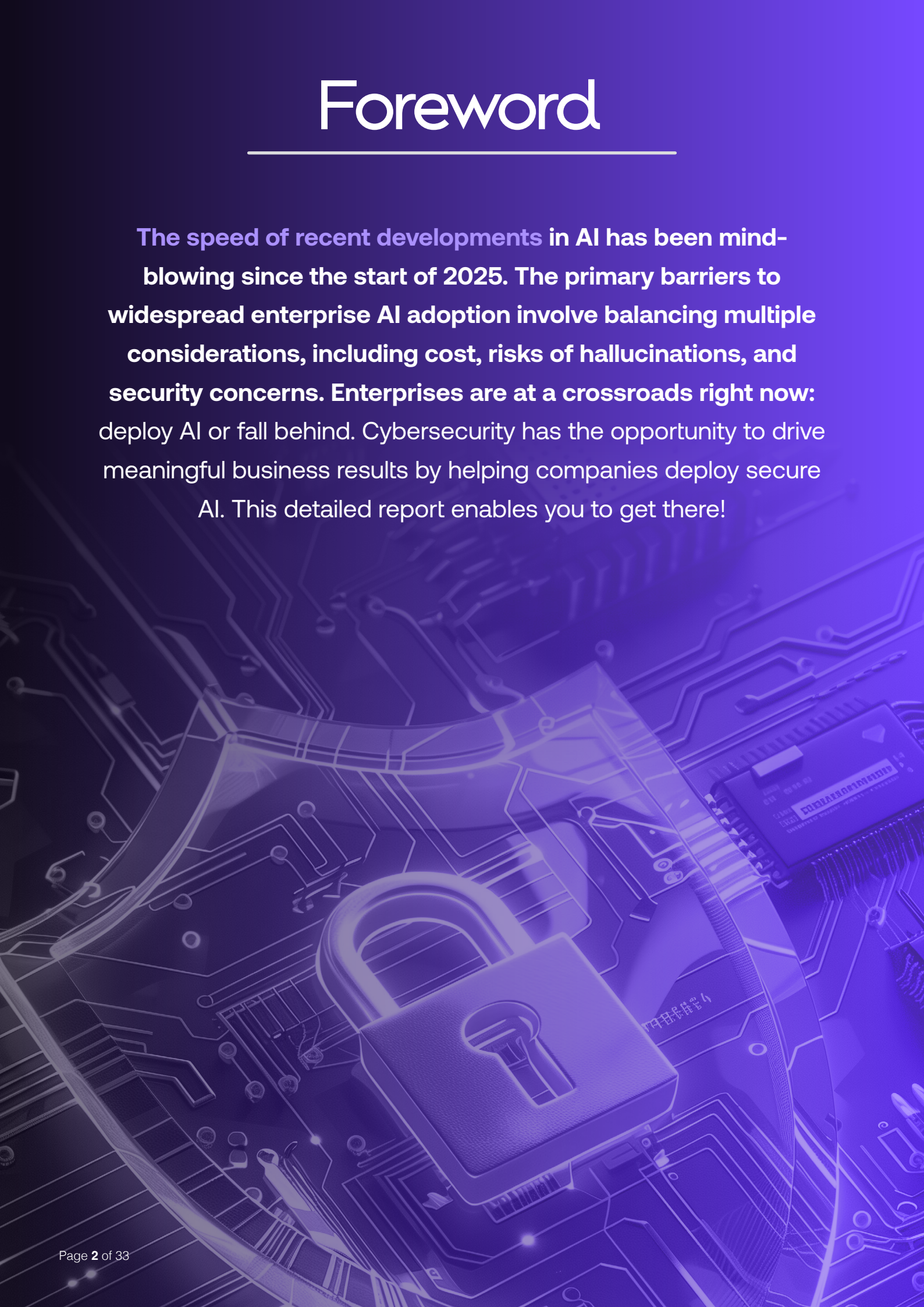
X
https://x.com/InvestiAnalyst

LinkedIn (Personal)
https://www.linkedin.com/in/francis-odum-0a8673100/

Table of Contents

Foreword

The speed of recent developments in AI has been mind-blowing since the start of 2025. The primary barriers to widespread enterprise AI adoption involve balancing multiple considerations, including cost, risks of hallucinations, and security concerns. Enterprises are at a crossroads right now: deploy AI or fall behind. Cybersecurity has the opportunity to drive meaningful business results by helping companies deploy secure AI. This detailed report enables you to get there!



Actionable Summary

1. **Outline / Key objectives of this report:** In 2024, SACR wrote a deep dive on how enterprises could secure their AI. This report is version 2, incorporating key market insights from the past 8 months. This report features many vendors that can help secure AI in the enterprise context, meaning they can help provide AI governance, secure data, and protect models. This report helps readers uncover the following:
    1. Understand the state of AI adoption and risks as of 2025.
    2. Understand market categorization and vendor focus areas as of 2025.
    3. Provide clear recommendations for securely deploying AI based on our CISO conversations.

2. **Security for AI Market solutions:** Our research shows a significant overlap among vendor capabilities. Some vendors have significantly more capabilities, while others are relatively new entrants. Across the market, three core capabilities consistently emerged:
    1. AI governance controls
    2. Runtime security for AI
    3. GenAI red teaming and penetration testing

3. **AI Developments In 2025:** Currently, the most prominent security risks for enterprises are related to DeepSeek. SACR's research indicates that enterprise AI adoption continues to accelerate, significantly driven by advancements such as DeepSeek R1. Many organizations still favor managed AI services due to their simplicity of deployment. Although open-source solutions have made notable progress, most enterprises still prefer closed-source models. Nonetheless, there's a noticeable rise in organizations self-hosting models, driven by the need for complete data privacy and full control over the model lifecycle. This report helps practitioners understand how to secure AI and the market solutions currently on the market.

4. **Recommendations for enterprises:**
    1. Data security controls: Organizations must prioritize foundational data security controls before deploying AI technologies.
    2. AI preventative security & governance controls: Organizations should implement discovery and cataloging solutions for AI, specifically tracking location, usage purpose, and responsible stakeholders within the enterprise.
    3. AI runtime security: This area presents the largest opportunity for improving enterprise AI security. We demonstrate the gaps and limitations of existing cybersecurity controls when addressing AI-specific risks and concerns.

5. **Market categorization:** Based on our extensive work, SACR concluded that most vendors fall into two primary categories for securing AI:
    1. Securing employee AI usage and enterprise agents
    2. Securing the AI product and application model lifecycle

Illustrative Security for AI Market Categorization

AI PRUDUCT SECURITY LIFE CYCLE

PROTECT AI

HIDDENLAYER

NOMA

pillar

TROJ.AI

paloalto NETWORKS

AI USAGE & ENTERPRISE CO-PILOT APPS

zenity

WITNESS AI

Prompt:

DATA SECURITY CONTROLS → **AI GOVERNANCE** → **AI RUNTIME**

Introductory Blurb

AI has brought significant change and opportunity to enterprises. In 2023, AI emerged and some organizations became early adopters. In 2024, a significant increase in enterprise AI adoption occurred, but nothing previously observed will compare to the wave coming in 2025. Many companies have been hesitant to adopt AI due to security and privacy risks. However, in 2025, organizations that have not adopted AI, will no longer be able to remain competitive if they want to keep up with competitors. Organizations must now actively pursue AI adoption.

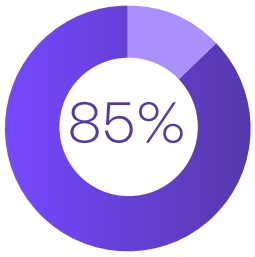
- **Enterprise Benefits of AI - Cost Savings & Spend**: For many enterprises, the return on investment from adopting AI is primarily cost savings rather than new revenue opportunities. Enterprise spend on AI is expected to increase by around 5% in 2025. To put this in perspective, IBM states that companies plan to allocate an average of 3.32% of their revenue to AI—equivalent to $33.2 million annually for a $1 billion company.
- **Challenges of AI Adoption**: Adopting AI, whether LLMs, GenAI, or AI Agents, is not without significant challenges. AI decision makers including CISOs, CIOs, VPs of Data, and VPs of Engineering cite concerns about data security, privacy, bias, and compliance. Several lawsuits have resulted in financial losses and reputation damage for companies that got AI implementations wrong. For example, an Air Canada chatbot gave a customer misleading information about bereavement fares and was later ordered to provide a refund to the customer. In another case, a South Korean startup leaked sensitive customer data in chats and was fined by the South Korean government approximately $93k. While these amounts can seem like “slap on the wrist” type of fines, it can be much worse. In February 2023, Google lost $100 billion in market value after its Bard AI chatbot shared inaccurate information.
- **One of the leading CISOs we spoke to:** “This is the first time security has become a business driver—a case where security actively supports business speed. Artificial Intelligence (AI) and Large Language Models (LLMs) are transforming industries, improving efficiency, and unlocking new business opportunities. However, rapid adoption presents significant security, compliance, and governance challenges. Organizations must balance AI innovation with risk management, ensuring regulatory alignment, customer trust, and board-level oversight. Without a structured security strategy, companies risk exposing sensitive data, enabling adversarial manipulation, and eroding stakeholder confidence. AI security is no longer solely an operational concern; it is a strategic imperative directly impacting enterprise risk, regulatory exposure, and long-term business viability.”

State of Enterprise Adoption of AI In 2025

Adoption of AI Rising into 2025 From 2024

As mentioned earlier in the report, the market for enterprises wishing to invest in AI is growing rapidly. The AI Global GenAI Security Readiness Report states that 42% of organizations are actively implementing LLMs across various functions, while another 45% are exploring AI implementation. Enterprises are adopting AI across multiple industries, with notable increases in adoption within legal, entertainment, healthcare, and financial services. According to Menlo Ventures' 2024 State of Generative AI in the Enterprise report, AI spending for these industries ranged from $100 to $500 million. The report highlights other significant enterprise GenAI use cases including code generation, search and retrieval, summarization, and chatbots.

Across enterprises, AI adoption continues to grow, with over 85% of organizations now using either managed or self-hosted AI in their cloud environments, reflecting continued but stabilizing adoption rates. Managed AI services, in particular, are currently utilized by 74% of organizations, up from 70% in the previous year.



85% of organizations are using some form of AI (either managed or self-hosted)

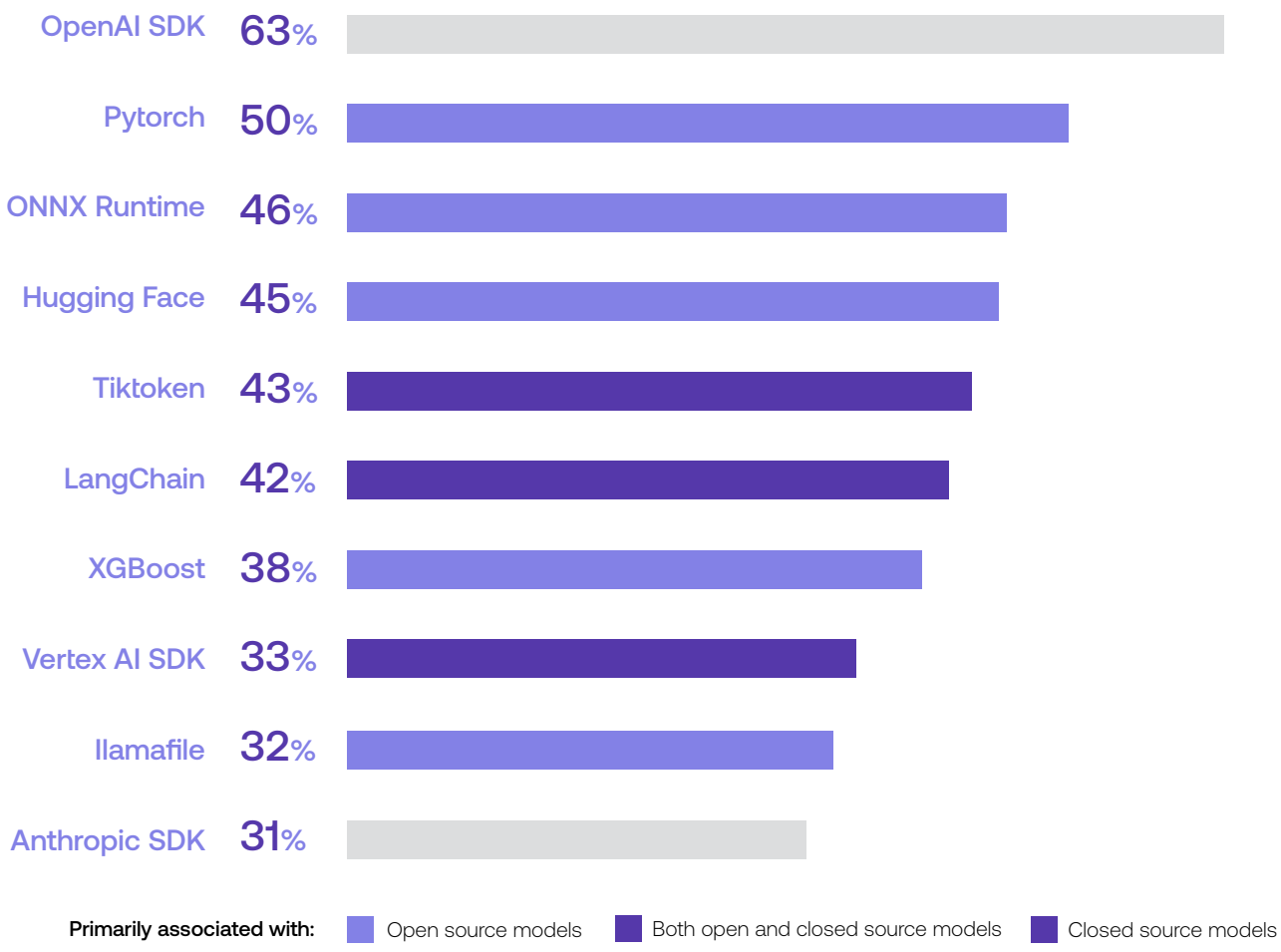
Source: State of AI Wiz Research

The Debate Over Open vs Closed Source Models

Enterprises started off using mostly closed-sourced models but are increasingly moving toward open-source options. A report by Andreessen Horowitz titled 16 Changes to the Way Enterprises Are Building and Buying Generative AI predicts that 41% of interviewed enterprises will increase their use of open-source models in their business in place of closed models. The report also predicts an additional 41% will switch from closed to open models if open-source models match closed-model performance.

Despite increased competition, OpenAI and Azure OpenAI SDKs continue to dominate, with more than half of organizations deploying them in their environments. PyTorch, ONNX Runtime, Hugging Face, and Tiktoken round out the top five AI tools used in the cloud.

AI hosted technologies by percentage of organizations

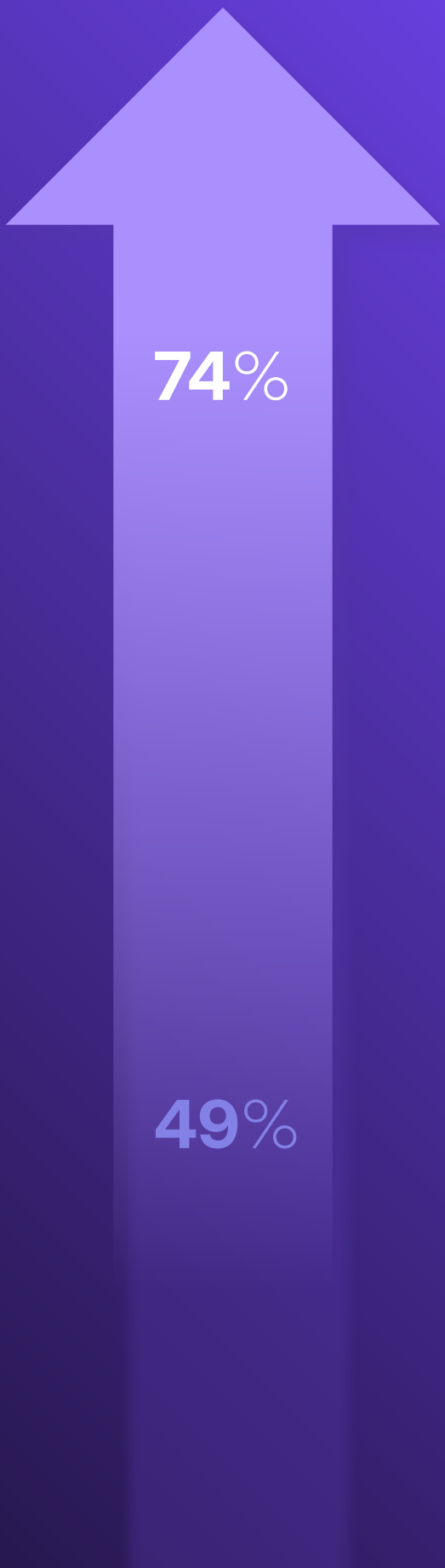


Source: State of AI Wiz Research

The advantages of closed-source models explains why it remains the top choice. These benefits include vendor support, predictable release cycles, increased security due to dedicated teams' restricted codebase access, and integration with vendor-provided infrastructure.

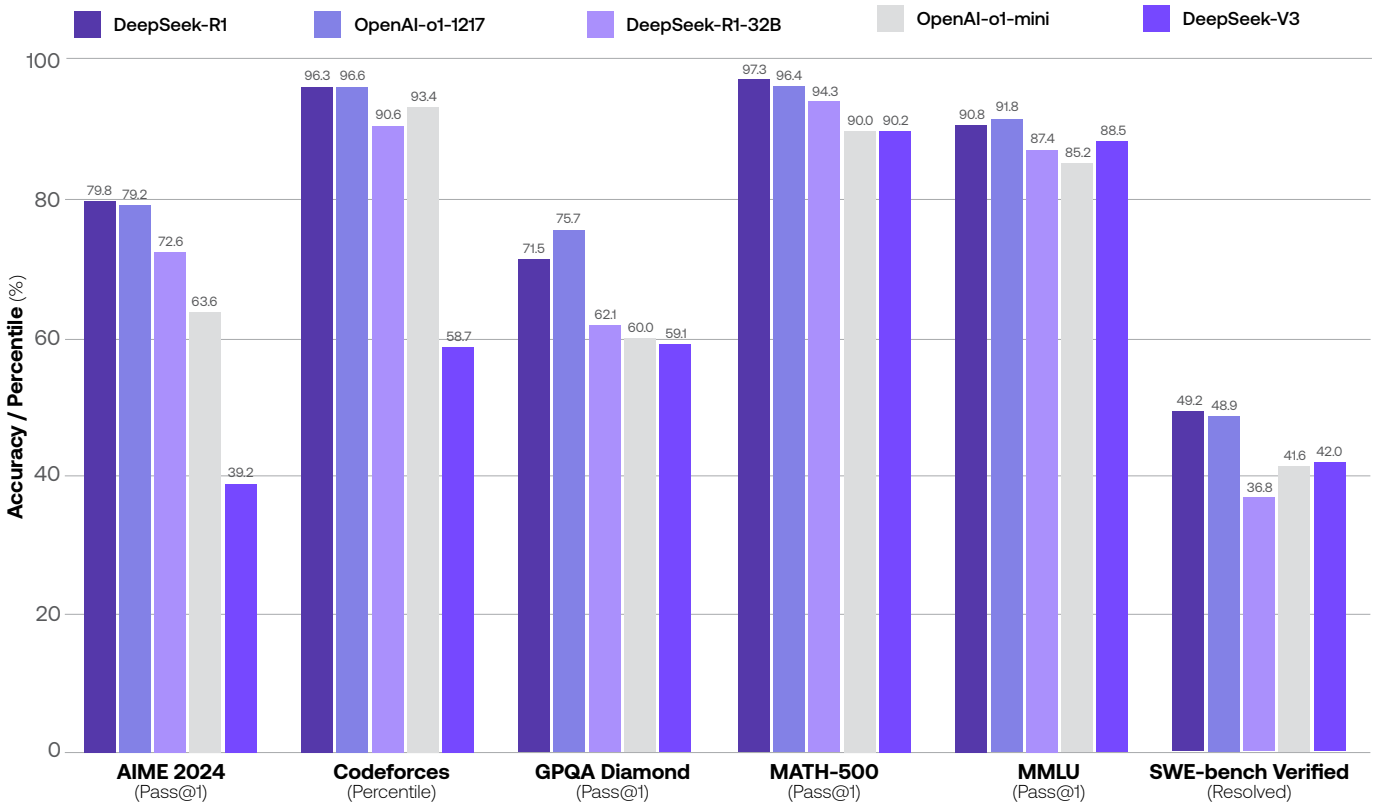
Open Source Continues To See Momentum

The advantages of open-source models include lower costs due to the absence of licensing fees, rapid innovation from a global community, customization, and transparency, potentially driving increased adoption among enterprises. Many of the most popular AI technologies are either open-source or tightly linked to open-source ecosystems. BERT remains the most widely used self-hosted model (rising to 74% adoption from 49%) among self-hosting organizations, while Mistral AI and Alibaba's Qwen2 models have newly secured a place among the top models.



Open-source momentum fueled by DeepSeek's cost savings

The interest in open source models has further intensified following disruptions caused by DeepSeek. DeepSeek achieved similar results as OpenAI's O1. According to DeepSeek, DeepSeek-R1 scored 97.3% on the MATH-500 benchmark, slightly outperforming o1's at 96.4%. On Codeforces, a competitive coding benchmark, DeepSeek-R1 earns a rating of 2029 (96.3% percentile), while o1 scores slightly higher at 2061 (96.6% percentile).

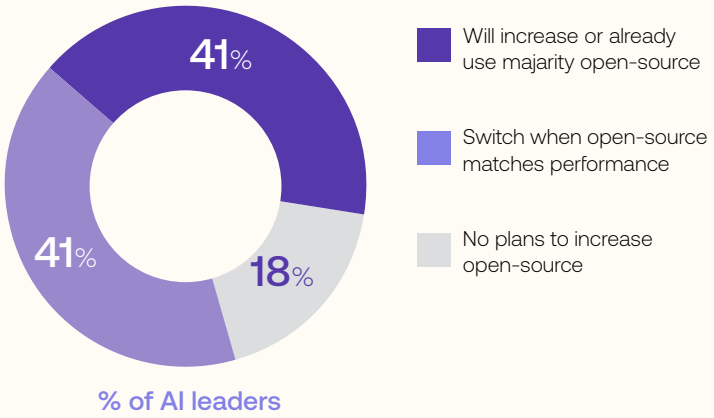


While there are numerous security concerns surrounding DeepSeek-r1, its performance demonstrates that open-source models are highly competitive with closed-source models in terms of both performance and cost. Enterprises are unlikely to overlook the cost-saving opportunities presented by open-source models. Therefore, AI security vendors have a strategic opportunity to capitalize on this by addressing and mitigating security concerns associated with open-source model adoption.

Open vs closed source models predictions

Andreessen Horowitz predicts an even divide emerging in the use of open and closed models. This represents a significant shift from 2023 when market share was 80%–90% closed-source. Teams choosing closed-source typically adopt licensed models, paying fees and establishing agreements that restrict model providers from using ingested company data for training. Teams choosing open-source models must emphasize AI model lifecycle security, ensuring models produce expected outputs and are resistant to jailbreaks.

Enterprise expectations for source usage in 2024 and onward

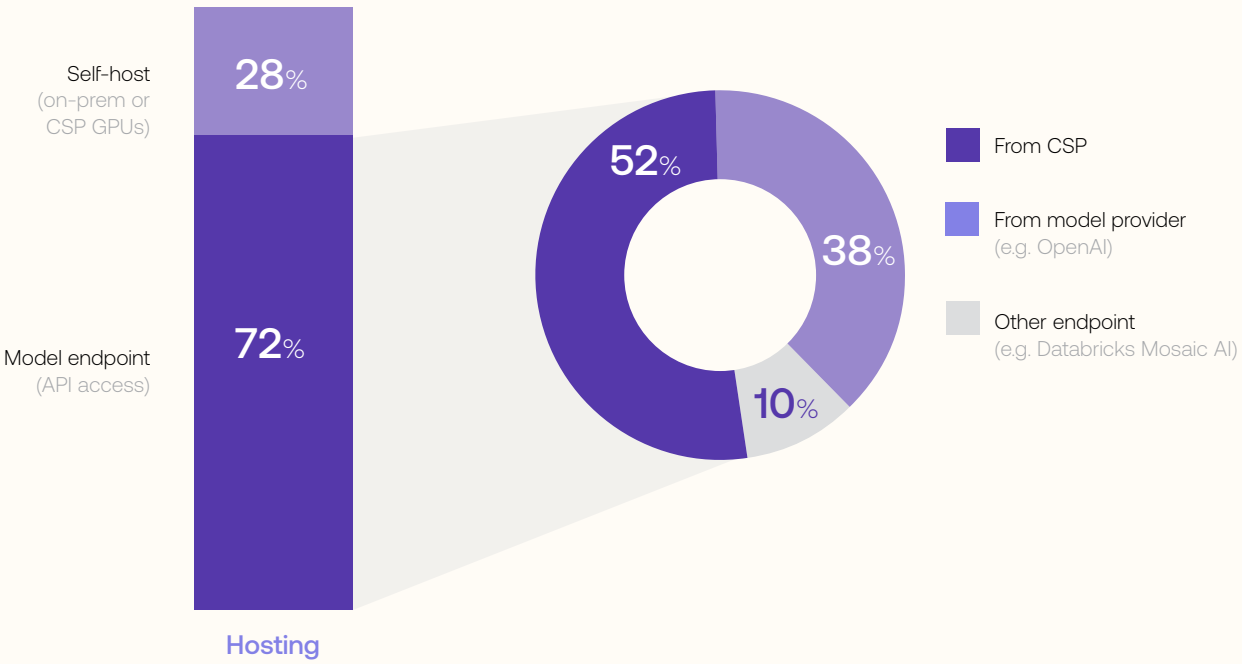


% of AI leaders

Cloud service providers (CSPs) can influence model choices

Andreessen Horowitz's report indicates a high correlation between CSP selection and preferred AI models: Azure users generally preferred OpenAI, while Amazon users preferred Anthropic or Cohere. Their chart below shows that, of the 72% of enterprises using an API to access models, over half use models hosted by their CSP. This preference likely reflects enterprises' familiarity with their CSP as a data subprocessor and acceptance of associated risks.

Where do enterprises access their large language models?



Hosting

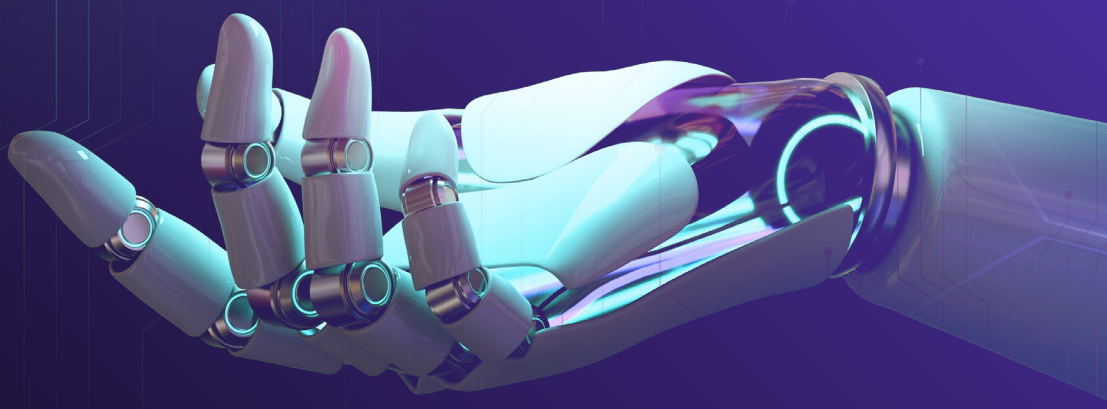
Security Implications

Organizations face numerous choices related to AI procurement and implementation. This extends to AI security procurement, where CISOs prioritize solutions which maintain data privacy by operating within existing architecture rather than relying on external third parties. Both CISOs and product executives seek configurable AI security products, allowing teams to customize detection settings for different projects. For instance, one project might require heightened sensitivity to PII leakage, while another requires enhanced monitoring of toxic responses.

The bottom-line is, AI and model security represent a rapidly evolving domain. A common theme among CISOs, CIOs, and product and engineering leaders is the demand for AI security products capable of rapid innovation to keep up with the evolving space. CISOs emphasize that actual AI threats may not become fully apparent for another 2-5 years; thus, it is critical for AI security products to continuously address emerging threats.

Managed vs. self-hosted

While many companies rely on managed AI offerings—increasing from 70% to 74% in 2025—the report highlights a dramatic rise in self-hosted AI adoption, jumping from 42% to 75% year over year. This surge stems from both AI capabilities embedded in third-party software and organizations seeking greater control over their AI deployments. However, this shift toward self-hosted AI requires robust governance to protect cloud environments.



Enterprise Risks In Developing AI In 2025

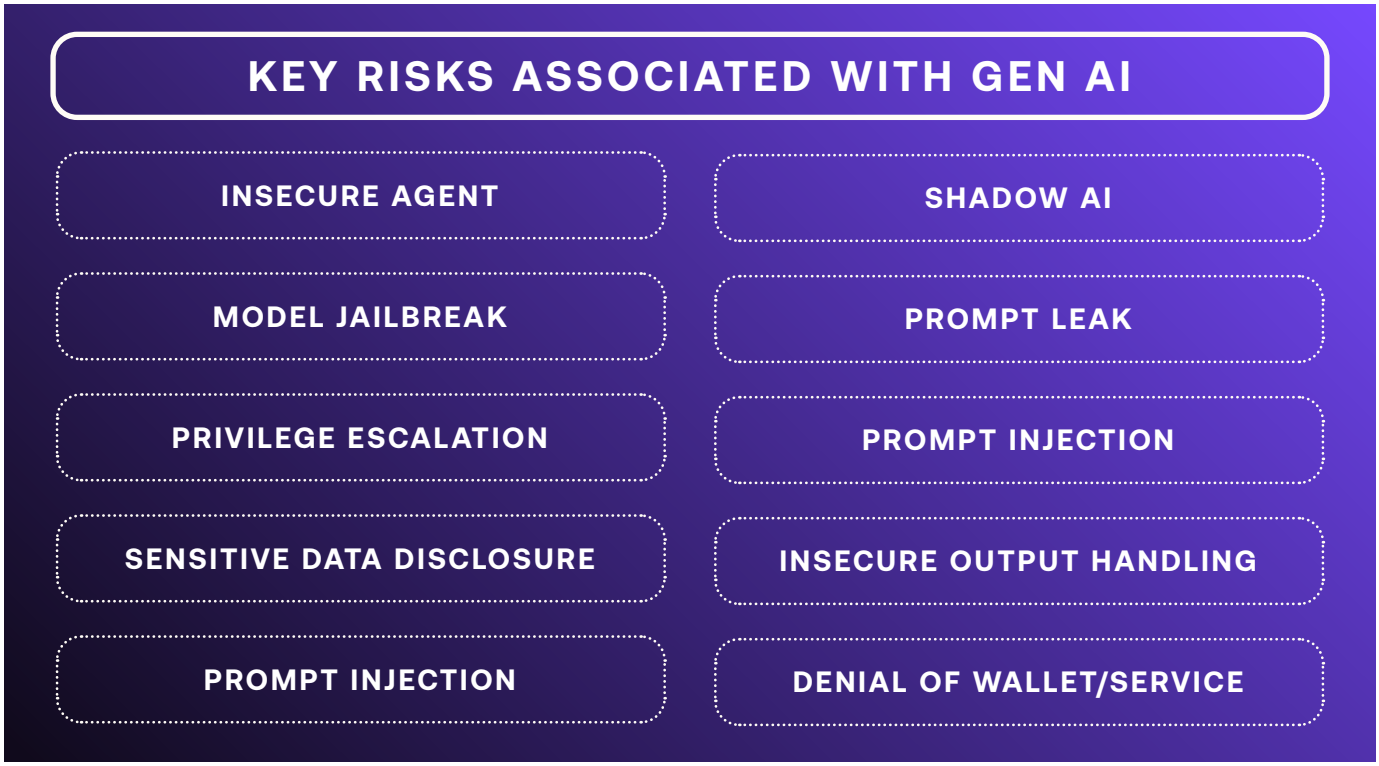
Based on our research, there is no shortage of risks that organizations are facing coming into 2025 — specially given significant progress by Chinese companies like DeepSeek. These 3 frameworks have been the most popularly used to categorize and understand risks for companies.

Industry Frameworks for AI Threats

1. **OWASP Top 10 for LLMs and Generative AI:** CISOs rely on the OWASP Top 10 framework to identify and mitigate key AI threats. This framework highlights critical risks like prompt injection (#1) and excessive agency (#6), helping teams implement proper input validation and limit LLM permissions.
2. **MITRE's ATLAS:** MITRE's Adversarial Threat Landscape for AI Systems (ATLAS) framework provides a comprehensive matrix of potential AI threats and attack methods. The framework outlines scenarios like service denial and model poisoning. Other key resources include Google's Secure AI Framework, NIST's AI Risk Management Framework, and Databricks' AI Security Framework.
3. **GenAI Attacks Matrix:** A knowledge source matrix documenting TTPs (tactics, techniques, and procedures) used to target GenAI-based systems, copilots, and agents. Inspired by frameworks like MITRE ATT&CK, this matrix assists organizations in understanding security risks applicable to M365, containers, and SaaS.

More broadly, we would categorize the risks that enterprises face into two key areas:

1. Employee Activity and safeguarding AI usage
2. Securing the lifecycle of homegrown applications



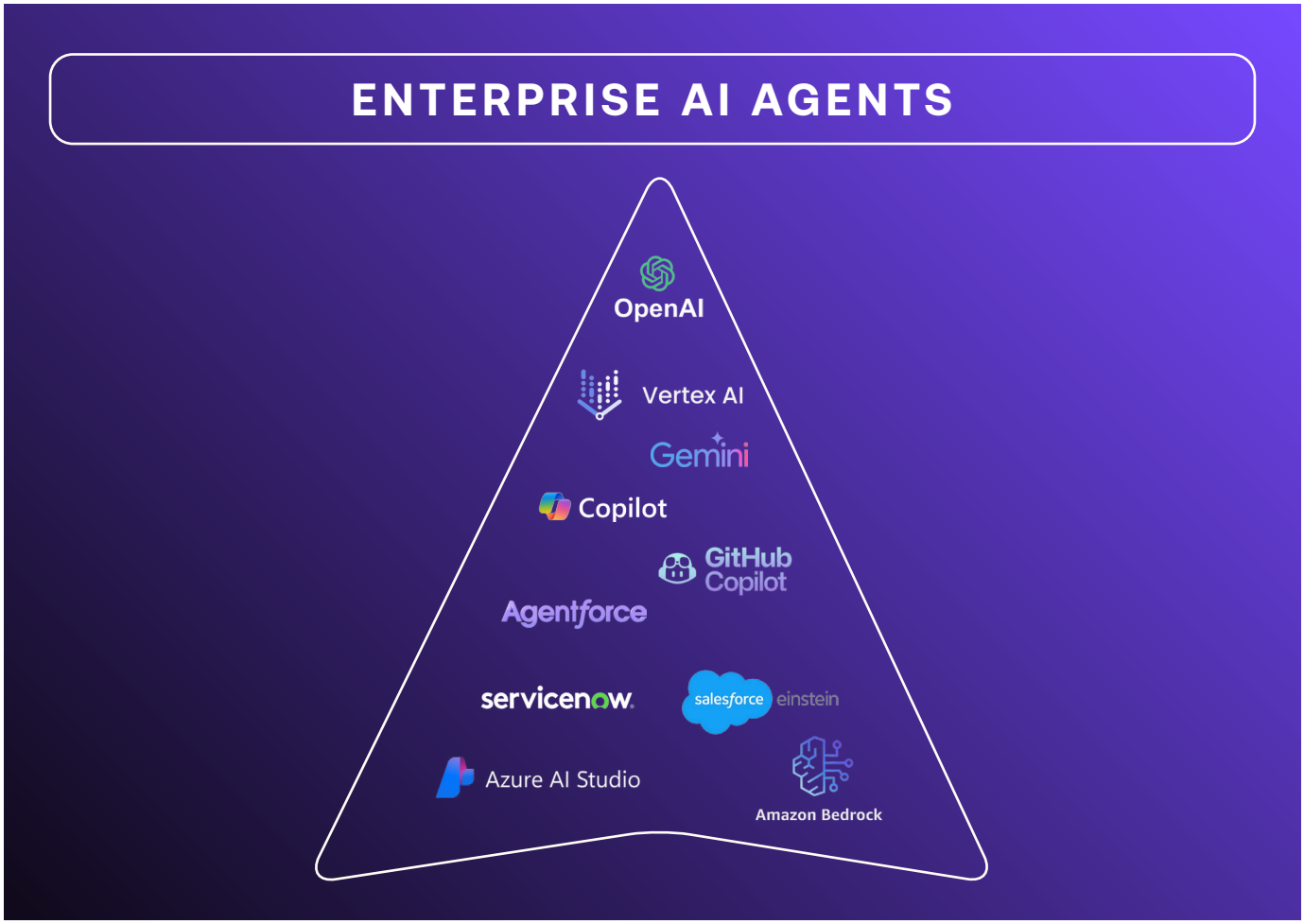
Risks Around Securing AI Usage, Activity and Guardrails

In recent months, enterprises integrating AI tools like Microsoft Copilot, Google's Gemini, ServiceNow, and Salesforce have encountered significant security challenges. A primary concern is over-permissioning, where AI assistants access more data than necessary, leading to unintended exposure of sensitive information. For instance, Microsoft Copilot's deep integration with Microsoft 365 allows it to aggregate vast amounts of organizational data, potentially creating vulnerabilities if permissions aren't carefully managed.

The following represent prominent risks identified as of 2025, primarily focused around AI usage, data exposure, and policy violations:

Data Loss and Privacy Concerns: Organizations seek to better understand risks associated with unauthorized data exposure through AI systems. This is particularly critical for organizations in regulated sectors like finance and healthcare, where data breaches may result in significant regulatory penalties and reputational damage. In 2025, companies processing sensitive data through AI applications are actively evaluating and implementing robust safeguards to ensure compliance with HIPAA, GDPR, and other regulations.

Shadow AI Proliferation: A major challenge for enterprises in 2025 is the widespread adoption of unauthorized AI tools by employees. This issue is particularly relevant due to new Chinese AI tools employees are adopting informally. These unsanctioned AI applications, including widely-used chatbots and productivity tools, pose significant security risks as they operate outside organizational control. While LLM firewalls and other protective measures have emerged to address this issue, the rapidly expanding AI landscape makes comprehensive monitoring increasingly complex. Organizations must balance innovation with security through clearly defined AI usage policies and technical controls.



Securing AI Lifecycle of Building Homegrown Applications

Why Securing the AI Lifecycle Differs from Traditional Software Development (SDLC)

Securing the AI development lifecycle presents unique challenges that distinguish it from traditional software development. While conventional security measures like Static Application Security Testing (SAST), Dynamic Application Security Testing (DAST), and API security remain relevant, the AI pipeline introduces additional complexities and potential blind spots:

- **Expanded Attack Surface:** Data scientists often utilize environments like Jupyter notebooks or MLOps platforms operating outside traditional Continuous Integration/Continuous Deployment (CI/CD) pipelines, creating potential security blind spots often overshadowing traditional code. AI systems face specific threats, including data poisoning and adversarial attacks, where malicious actors introduce corrupt or misleading data into training datasets, compromising model integrity.
- **Much More Complex AI Lifecycle Components:** The AI development process encompasses distinct phases—data preparation, model training, deployment, and runtime monitoring. Each phase introduces unique vulnerabilities and potential misconfigurations, necessitating specialized security checkpoints.
- **Supply Chain Security and AI Bill of Materials (AI-BOM):** AI applications often rely on vast amounts of data and incorporate open-source models or datasets. Without meticulous vetting, these components can introduce vulnerabilities, leading to compromised AI systems. Given the intricate dependencies—including scripts, notebooks, data pipelines, and pre-trained models—there's a growing need for an AI Bill of Materials (AI-BOM). This full inventory aids in identifying and managing potential security risks within the AI supply chain.

Without extensively diving deep, others include monitoring RAG access, governance and compliance issues and runtime security requirements.

Illustrative breakdown of the AI development lifecycle:

The AI development lifecycle introduces distinct security challenges requiring careful consideration.

It is important to outline clearly the core steps for building and developing AI models:

1. **Data Collection & Preparation:**

   Gather, clean, and preprocess data, ensuring quality and compliance. Split data into training, validation, and test sets while applying feature engineering.

2. **Model Development & Training:**

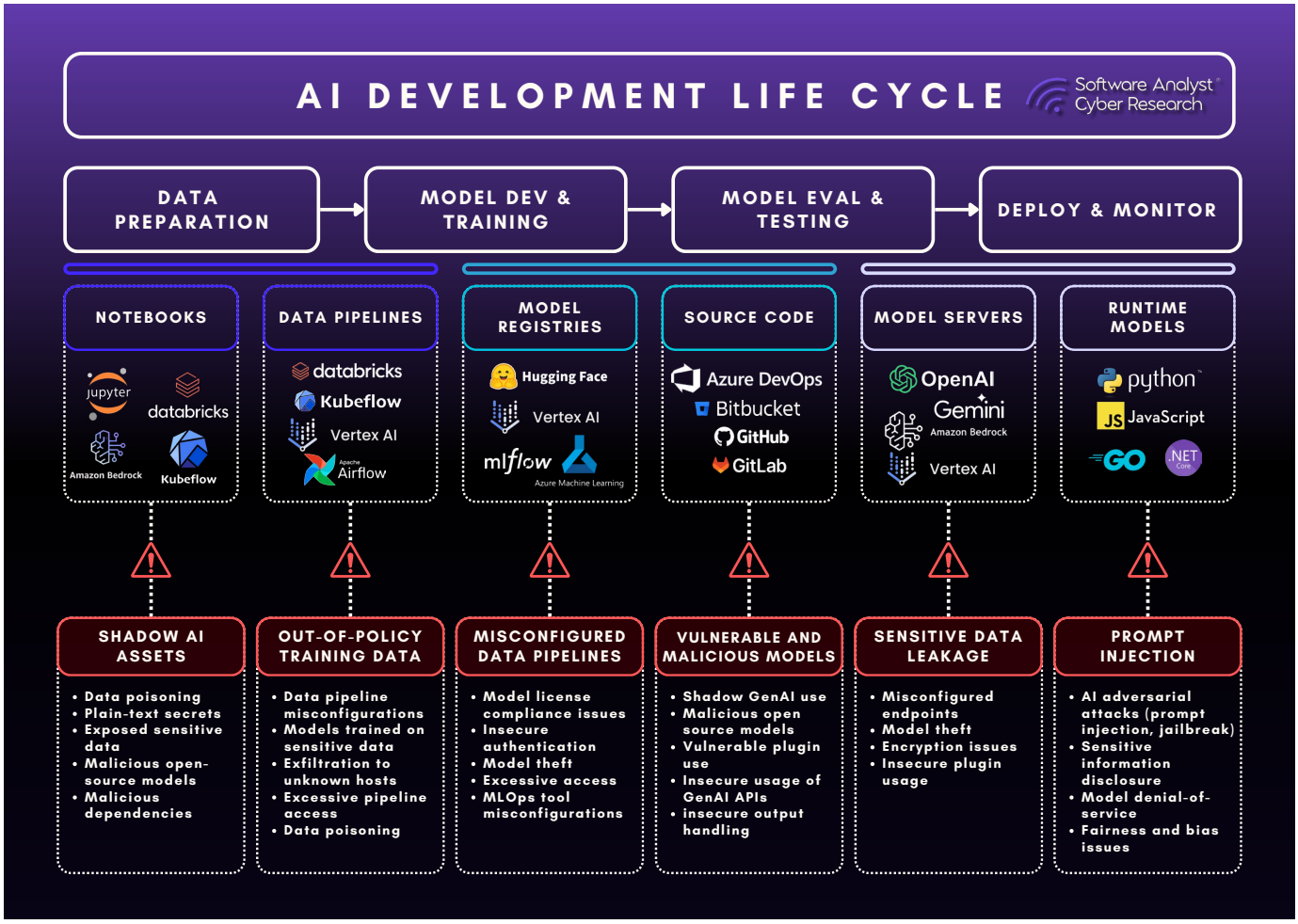
   Choose the right model type and framework, train it using optimization techniques, and fine-tune for accuracy. Address bias, fairness, and security risks like adversarial attacks.

3. **Model Evaluation & Testing:**

   Assess performance with key metrics, conduct adversarial testing, stress test for edge cases, and ensure explainability. Detect data poisoning, backdoors, and model drift.

4. **Deployment & Continuous Monitoring:**

   Deploy the model via API, cloud, or edge devices, implement real-time security monitoring, automate retraining through MLOps, and audit AI decisions for compliance and security.



Risks associated with each stage of building AI Apps

1. **Data Loss:** Enterprises face significant security risks when developing AI applications. Data leakage is particularly concerning when using models that connect to the internet or lack contractual protections against training on customer data. These risks aren't theoretical—they're happening now. David Bombal recently proved this point in a viral post where he used Wireshark to catch Deepseek sending data to China. This example illustrates why organizations must thoroughly examine model providers' privacy policies and data handling practices.

2. **Vulnerable Models:** MLSecOps is emerging as organizations recognize that AI builders operate differently from traditional software teams. Machine learning engineers and data scientists work in specialized environments like Databricks and Jupyter Notebooks rather than standard development pipelines. These environments require monitoring for misconfigurations and leaked secrets. MLSecOps includes scanning models for vulnerabilities and backdoors—analogous to scanning software dependencies for supply chain security. This practice is vital because attackers have demonstrated they can download models from repositories, modify them to transmit data to malicious servers, and redistribute them for unsuspecting users to download.

3. **Data Poisoning:** Data poisoning is another risk that can lead to backdoor insertion. Orca Security released an intentionally vulnerable AI infrastructure called AI Goat for AI security education and awareness. One vulnerability demonstrated by AI Goat is data poisoning, which involves uploading new or modifying existing data that the model uses to train

on, to change the model. Data poisoning is closely related to model poisoning, a threat allowing attackers to insert backdoors into models. Backdoors can cause deceptive behavior, which is difficult to detect. Anthropic researchers brought this to light last year, in a paper titled, "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training".

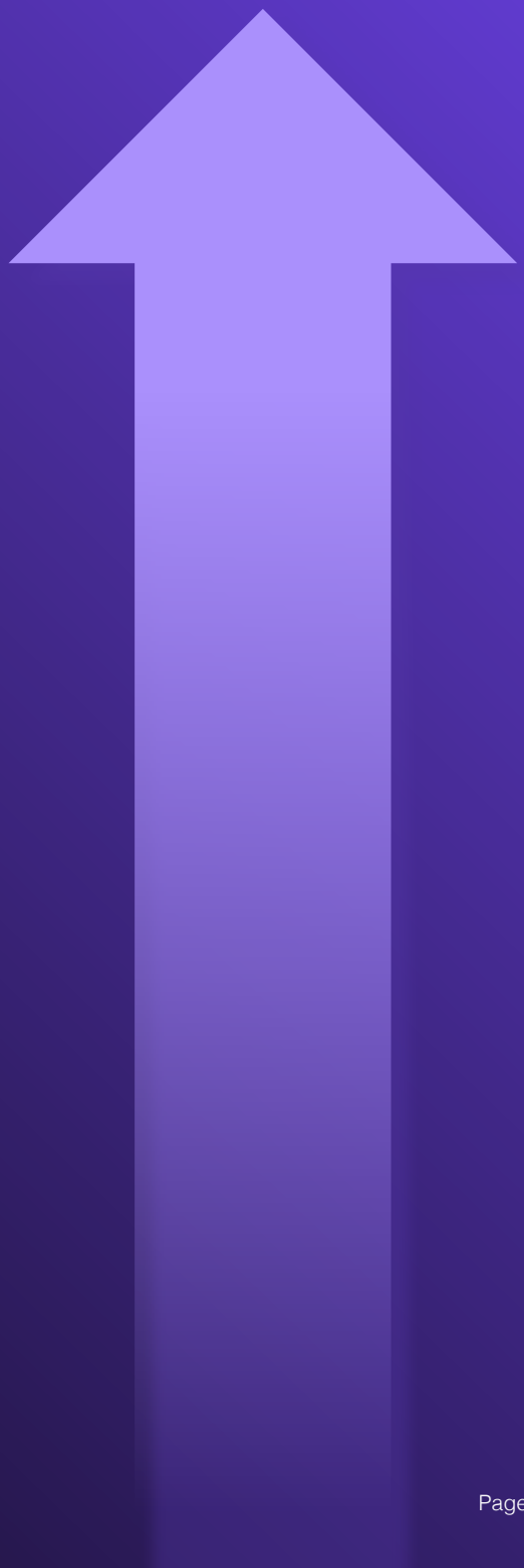
4. **Missing Machine Learning Expertise on Security Teams:** Deceptive behavior introduced through machine learning represents an important focal point for AI Security. Emmanuel Guilherme, AI/LLM Security Researcher at OWASP speaks to the complexities of machine learning and its impact on AI Security in this AI Global GenAI Security Readiness Report. Guilherme states, "The biggest obstacle to securing AI systems is the significant visibility gap, especially when using third-party vendors. Understanding the complexities of the ML flow and adversarial ML nuances adds to this challenge. Building a strong cross-functional ML security team is difficult, requiring professionals from diverse backgrounds to create comprehensive security scenarios". The complexities of machine learning and its significant role in AI application development has led to a new term, MLSecOps, which is focused on baking in security during machine learning workflows such as data pipelines and notebooks.

5. **Prompt Attacks & Jailbreaks:** At AWS re:Inforce 2024, Stephen Schmidt, Chief Security Officer at Amazon, highlighted that AI applications require continuous testing. LLMs change as users interact with them, not only when new versions are released. Traditional security approaches focusing solely on shifting scanning left—evaluating code pre-release—will not suffice. AI applications require continuous security scanning. During runtime, AI applications may encounter prompt attacks such as direct prompt injections, indirect prompt injections, and jailbreaks. These attacks can lead to unauthorized access, data loss, or harmful outputs.

Existing Cybersecurity Controls & Limitations

Organizations currently employ three primary methods to secure AI systems. There is no one-size-fits-all approach. Typically, organizations use 2/3 of the methodologies below:

1. Model providers and self-hosted solutions
2. Incumbent security controls and vendors
3. Implement pure-play AI security vendors



Model Providers and Self-hosted Solutions

The first includes AI vendors like OpenAI, Mistral, and Meta. These companies embed security, data privacy, and responsible AI protections into their models during training and inference. However, their primary focus remains on building the most advanced and efficient AI models rather than prioritizing security. Consequently, many corporations seek an independent security layer capable of protecting all models across various environments.

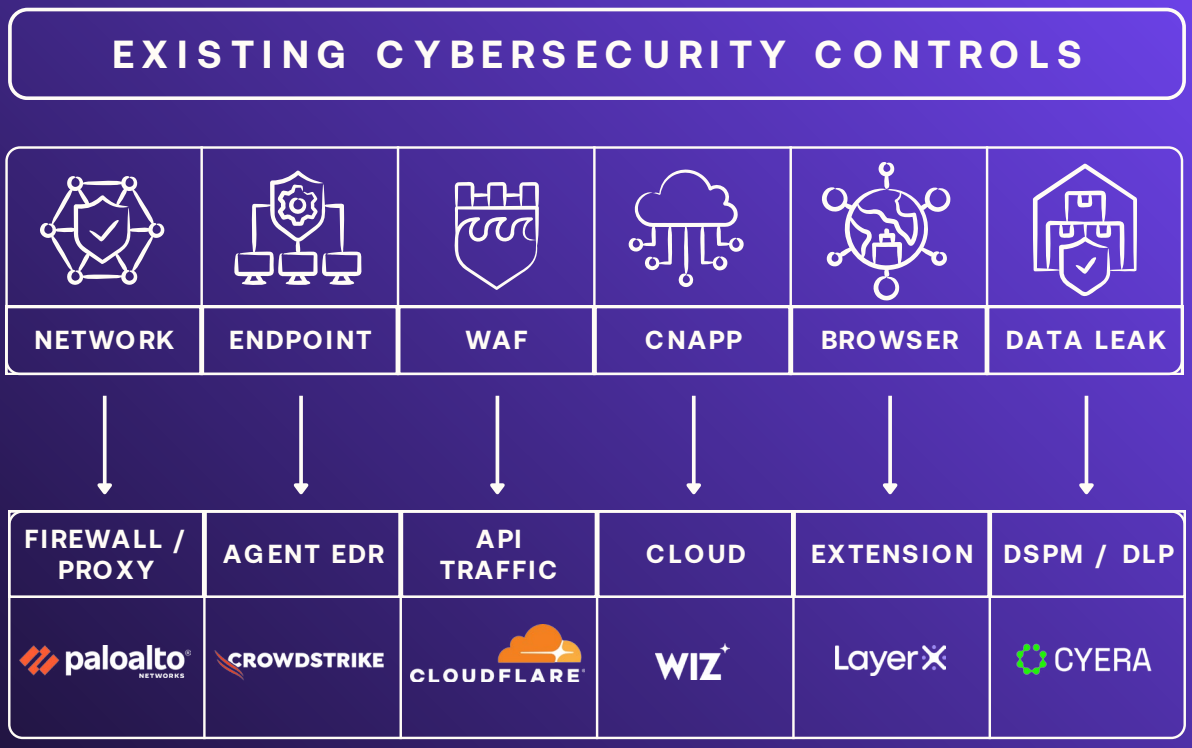
- **Self-hosted models:** As discussed earlier, some organizations choose self-hosted models primarily for enhanced security and data privacy, as sensitive information never leaves their controlled infrastructure. This approach has proven effective in industries with strict regulatory requirements like finance, healthcare, and government. By self-hosting, companies maintain full control over model updates, fine-tuning, and security protocols, reducing reliance on external providers that may introduce vulnerabilities.
- **Third-party models:** These pose security risks because data must be transmitted to external servers, increasing exposure to breaches, policy changes, and vendor lock-in. Vendors fully using these cloud-based models tend to deploy additional independent security measures.

Incumbent Cybersecurity Controls

First of all, many companies already had some level of cybersecurity controls for protecting data, network and endpoints. Our research found that enterprises lacking dedicated AI security solutions typically rely on one or more of the existing controls below. However, as detailed below, each has strengths and limitations. It's essential to recognize both the strengths and limitations of these controls in the context of AI security.

1. Network Security

- **Firewalls / SASE**: Traditional network security measures, including firewalls and intrusion detection systems, are essential **to protect** AI API endpoints and data traffic from unauthorized access and lateral movement. Some organizations implement AI-specific firewall rules to throttle high-risk API queries, such as rate-limiting LLM API requests to prevent model extraction. However, these measures often lack the capability to detect malicious AI prompts or protect against AI model manipulation, like backdoor insertions. Notably, vendors like Witness AI integrate with network security providers such as Palo Alto Networks to enforce AI-specific policies.
- **Web Application Firewalls (WAFs)** can also rate-limit AI inference requests to prevent model scraping and mitigate automated prompt injection attacks. However, they lack AI-specific threat detection capabilities.



Software Analyst Cyber Research

- **Cloud Access Security Brokers (CASBs)** provide visibility and control over AI-related data flowing through cloud environments, making them highly effective for securing AI SaaS applications and preventing shadow AI deployments. However, CASBs do not protect AI models themselves, as they primarily focus on access governance and cloud data security.

2. Data Loss Prevention (DLP) & Data Security Posture Management (DSPM)

DLP and DSPM solutions are designed to protect sensitive information from unauthorized access or exfiltration. In AI systems, these tools help prevent accidental or intentional leaks of personally identifiable information (PII) or proprietary business data. However, they often fall short in addressing AI-specific threats, such as model exfiltration via API scraping or **analyzing** the intricacies of AI model logic. Additionally, without AI-aware policies, these tools may not effectively mitigate risks unique to large language models (LLMs), like prompt leakage.

3. Cloud Security (CNAPP & CSPM)

Since AI models are frequently deployed on cloud platforms like AWS, GCP, or Azure, Cloud-Native Application Protection Platforms (CNAPP) and Cloud Security Posture Management (CSPM) tools identify misconfigurations. Many vendors have begun extending their capabilities to detect vulnerabilities specifically for cloud-centric AI models.

4. Endpoint Detection and Response (EDR):

EDR solutions are effective in detecting malware and traditional endpoint threats. However, they do not typically monitor AI model threats, such as unauthorized inference attempts or API scraping activities aimed at extracting model functionalities.

5. Application Security Scanners (SAST/ DAST):

Static and Dynamic Application Security Testing tools are proficient at assessing traditional application vulnerabilities. Yet, they struggle to evaluate AI model behaviours or their resilience against adversarial attacks, though we're seeing new vendors add controls.

6. Browser Security:

While browser security measures can control the usage of AI chatbots, they do not contribute to securing proprietary AI models themselves. These controls are limited to user interactions and do not extend to the underlying AI infrastructure.

These established security vendors, originally focused on traditional areas, are beginning to detect some AI traffic and adjacencies. Some have started incorporating AI security features into their offerings. However, innovation in this area remains slow, and they have yet to deliver groundbreaking AI security capabilities. Additionally, because they cover a broad spectrum of security areas—including cloud, network, and endpoint security—AI security often becomes just one aspect of their portfolio rather than a dedicated focus.

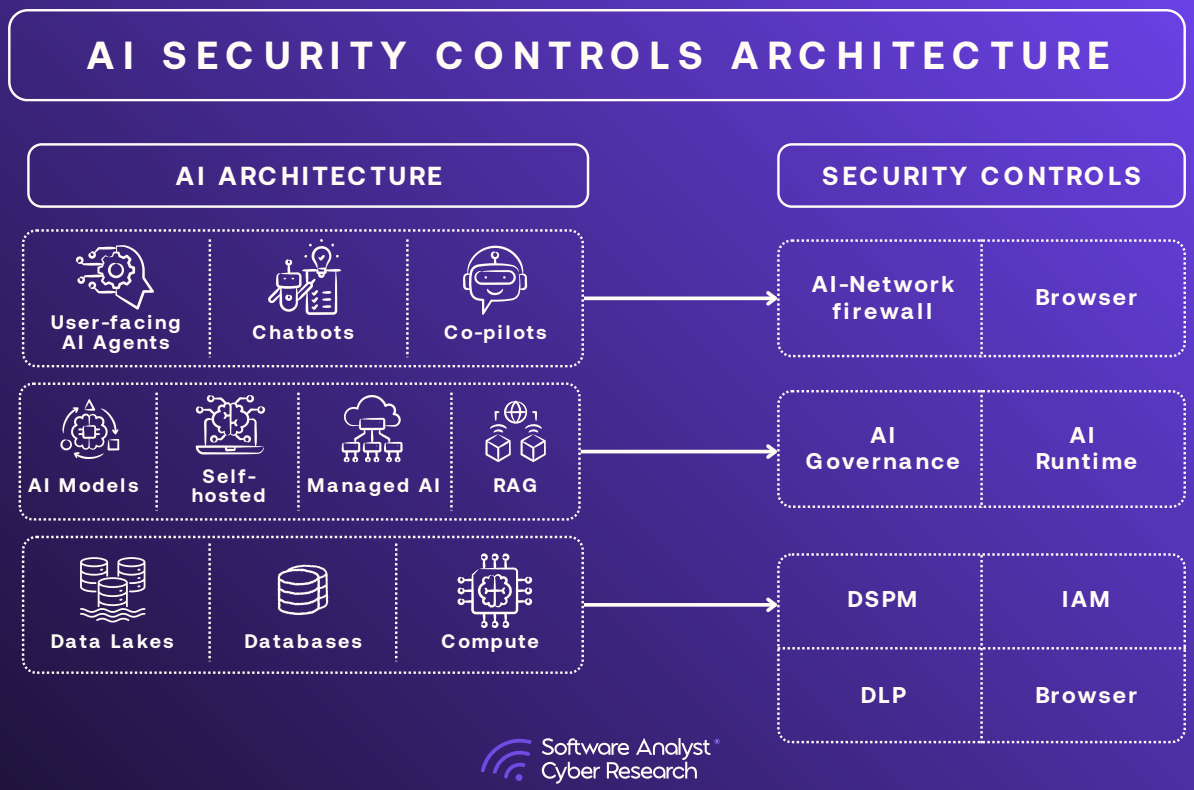
Emerging Pure-Play Solutions

This third category comprises new entrants which are purpose-built to secure AI. Most of these companies focus on niche challenges, including securing AI use, protecting against prompt injection attacks, and identifying vulnerabilities in open-source LLMs. It is evident from the analysis, so far, that AI security is sorely needed, representing a large domain and array of problem sets in a rapidly evolving space.

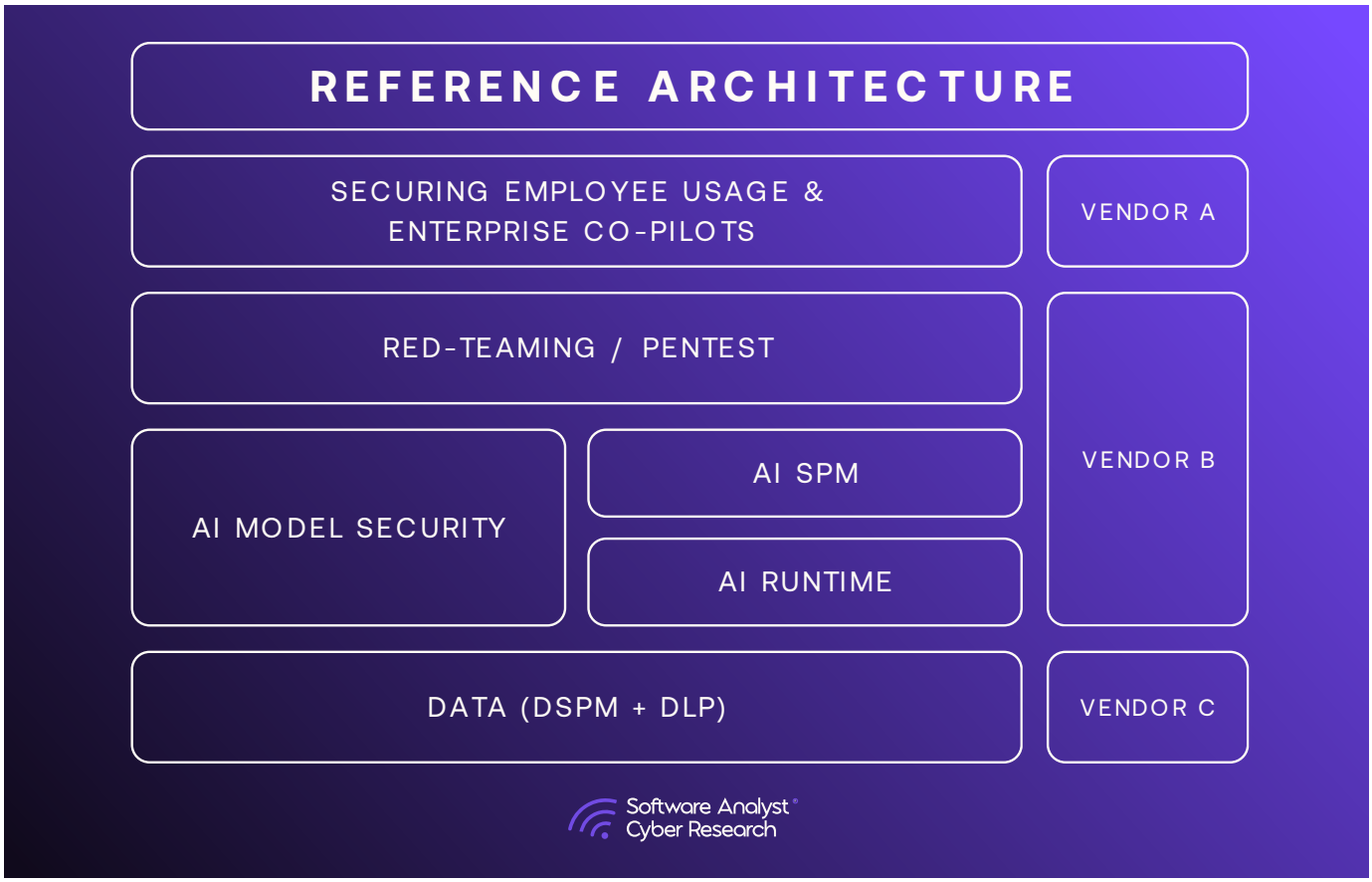
Market Solutions:
How Organizations Should Secure AI

This section is a guide to the existing market dynamics based on the market research conducted. Below is an overview of a typical enterprise architecture that requires securing, provided the organization is not fully reliant on closed-source models and vendor-hosted environments:

1. **Data & Compute Layer (AI Foundations)**: Organizations must secure databases, data lakes, or data leveraged from services like AWS S3. Some may already have a vector database. Enterprises should implement robust data governance controls.
2. **AI Model & Inference Layer (Securing AI Logic)**: Organizations typically have some combination of foundation models, fine-tuned models (open or closed-source), embeddings, inference endpoints, and agentic AI workflows, hosted either internally or on public clouds. Across these areas, enterprises must secure the entire lifecycle, from build to runtime.
3. **AI Application & User Interface Layer (Securing AI Usage)**: Organizations often use some combination of AI copilots, autonomous agents, or chatbots. Enterprises must ensure secure prompts, access controls, and implement content filtering measures.



Across these layers and depending on the enterprise's AI strategy, one or two specialized vendors should be deployed to comprehensively secure their AI stack along these areas.



SACR Recommendations for secure AI deployment

1. Data security controls
2. AI preventative security & risk controls
3. AI Runtime security

1) Data Security Controls for AI

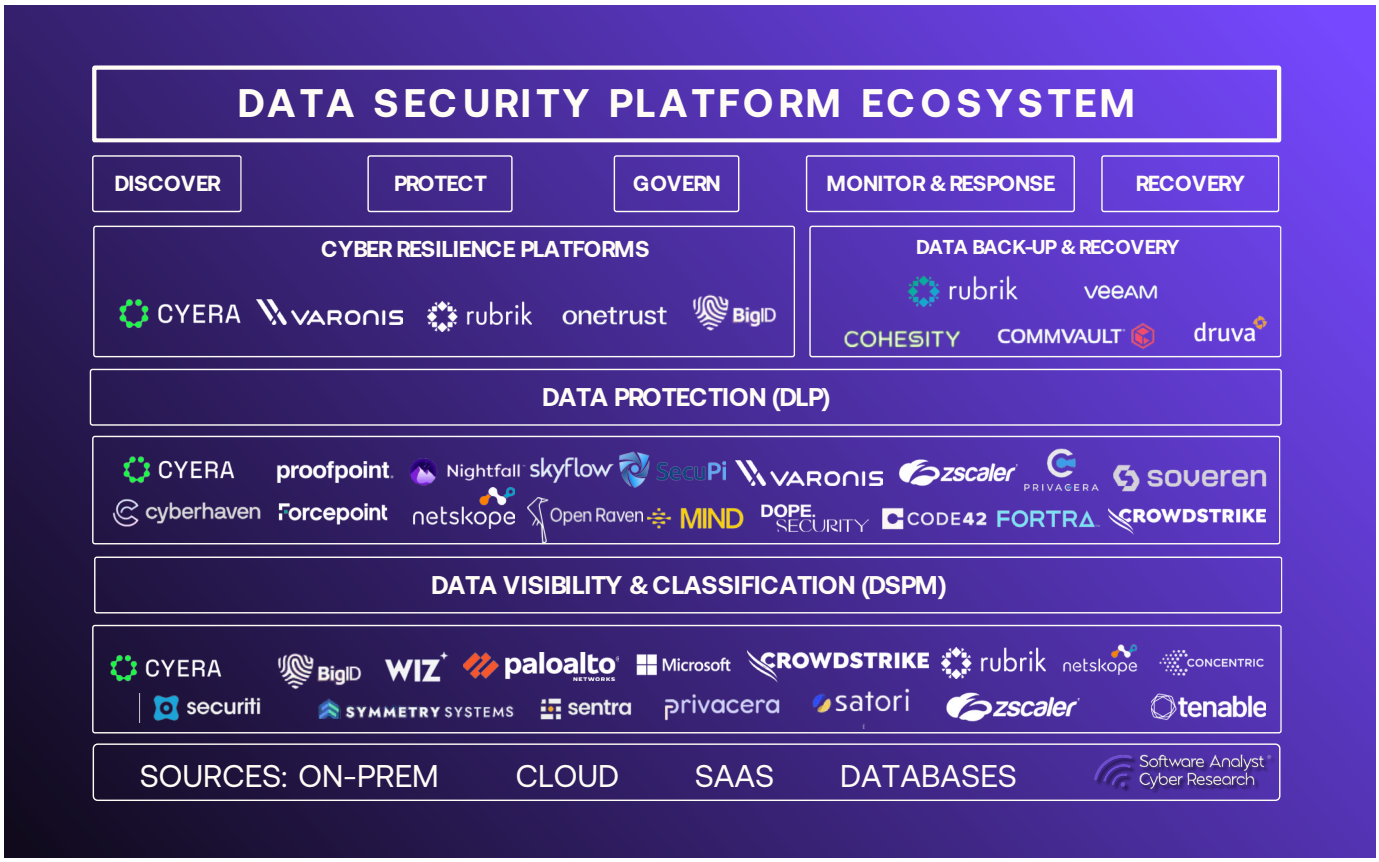
A strong AI security strategy cannot exist without robust data security controls, as AI models are only as secure and trustworthy as the data they are trained on and interact with. AI systems introduce new risks related to data access, integrity, leakage, and governance, making data security a foundational element in AI deployment. In the AI Global GenAI Security Readiness Report, surveyed respondents (CISOs, business users, security analysts, and developers) identified data privacy and security as major barriers to AI adoption. According to Adam Duman, Information and Compliance Manager at Vanta, AI has a way of making existing security issues worse. If companies are building GenAI apps or AI Agents they'll want to use their own data and will be thinking about trustworthiness, responsible outputs, and bias. For example, if a company does not have existing data governance in place such as policies on data tagging, these issues will become exaggerated with AI.

Organizations should focus on how vendors allow them to secure data at every stage before it gets leveraged by AI.

1. **DSPM (Data Security Posture Management (DSPM)**: Organizations should leverage DSPM solutions that scan their environments to give them visibility into all data stores and, most importantly, help them identify data sets for AI that must comply with GDPR, CCPA, HIPAA, and emerging AI-specific regulations, to ensure ethical and lawful deployment. The reason is that AI systems require vast amounts of structured and unstructured data, but not all data should be accessible to AI models. Hence companies should use DSPM to discover sensitive data.

2. **Role-based and attribute-based access controls (RBAC & ABAC):** Organizations should use data security vendors that have RBAC and ABAC controls to ensure that AI models and employees can only access authorized datasets.

3. **Data Loss Prevention (DLP):** Companies should prevent sensitive data leakage into generative AI models, as employees may unintentionally input regulated or confidential information.



SACR has done a full analysis and report on the data security industry, which can be read in more detail here - Data Security Platforms: The New Frontier in AI

2) AI Preventative & Governance Controls

Organizations should adopt proactive measures to protect AI. These key components include:

1. **Governance policies**: Organizations must establish robust governance and security policies for the secure deployment of AI. Governance controls should include oversight of all AI usage within the organization, both known and unknown, as well as securing the build and deployment lifecycle of AI applications.

2. **Discovery & catalogue solutions for AI within enterprises (Where, What & Who):** Organizations should deploy solutions to identify and inventory all categories of AI within their enterprise. Enterprises cannot secure what they cannot see—this fundamental cybersecurity principle equally applies to AI. Organizations must understand where AI is utilized internally: are employees interacting with ChatGPT, Claude desktop, or downloading open-source models from HuggingFace? By mapping both approved and shadow AI usage, enterprises can effectively scan models, generate AI Bills of Materials (AI-BOMs) to verify model provenance, and monitor for data loss. Organizations using open-source models should scan all models, as different models offer varying degrees of protection and guardrails. Understanding

model provenance through MLSecOps helps identify potential biases. Enterprises should also carefully review vendor privacy policies to determine where organizational data is transmitted. For instance, locally hosting DeepSeek may be preferable to mitigate data sovereignty risks associated with sending data internationally, such as to China.

3. **Leverage an AI-Security Posture Management (SPM) solution**: AI-SPM allows organizations to assess, monitor, and secure AI/LLM deployments through scanning, providing visibility, risk detection, policy enforcement, and compliance. AI-SPM protects AI models, APIs, and data flows from model-based attacks and data leakage. Leveraging elements of Cloud Security Posture Management (CSPM) and Data Security Posture Management (DSPM), AI-SPM ensures security teams have a real-time inventory of AI assets. AI-SPM should detect misconfigurations, data leakage, and risks to model integrity throughout the AI lifecycle. AI security companies in this category assist enterprises in understanding where AI is used and the connected tools.

4. **Simulate your defence systems for AI models (Pre-deployment and Post-deployment):** Gen-AI Penetration Testing and AI Red Teaming play a critical role in identifying vulnerabilities, testing resilience, and enhancing security across the AI model lifecycle. Their primary function is to simulate adversarial attacks and uncover security weaknesses in AI models before they are exploited in real-world scenarios. AI penetration testing and red teaming should be implemented as proactive controls, enforcing secure AI development practices, policy compliance, and risk assessments at compliance and governance levels. Organizations should also conduct red team simulations in live environments to detect model drift and operational risks.

To summarize this section, we'll use the words from one of the leading CISOs we spoke to, during our research. He phrased preventative controls as

**To address these risks, organizations must integrate Zero Trust Architecture (ZTA) for AI, ensuring that only authenticated users, applications, and workflows can interact with AI systems (Cloud Security Alliance, 2025). AI data governance and compliance frameworks must also align with evolving EU AI Act and SEC Cyber Rules to provide transparency and board-level visibility into AI risk management (Ivanti, 2025). By embedding AI security-by-design, companies can meet regulatory expectations while preserving business agility and innovation.*

For enterprises deploying AI at scale, proactive risk mitigation is essential. Implementing AI-Secure Bill of Materials (AI-SBOMs) strengthens supply chain security, while AI Red Teaming identifies vulnerabilities before exploitation (Perception Point, 2025). As AI regulations evolve, security leaders must invest in continuous monitoring, cyber risk quantification (CRQ), and alignment with industry Risk Management Frameworks (e.g., NIST, ISO/IEC, OECD) compliance to maintain resilience. The organizations that embed AI security into their core strategy—from development teams to the boardroom—will lead the next era of AI-driven business.

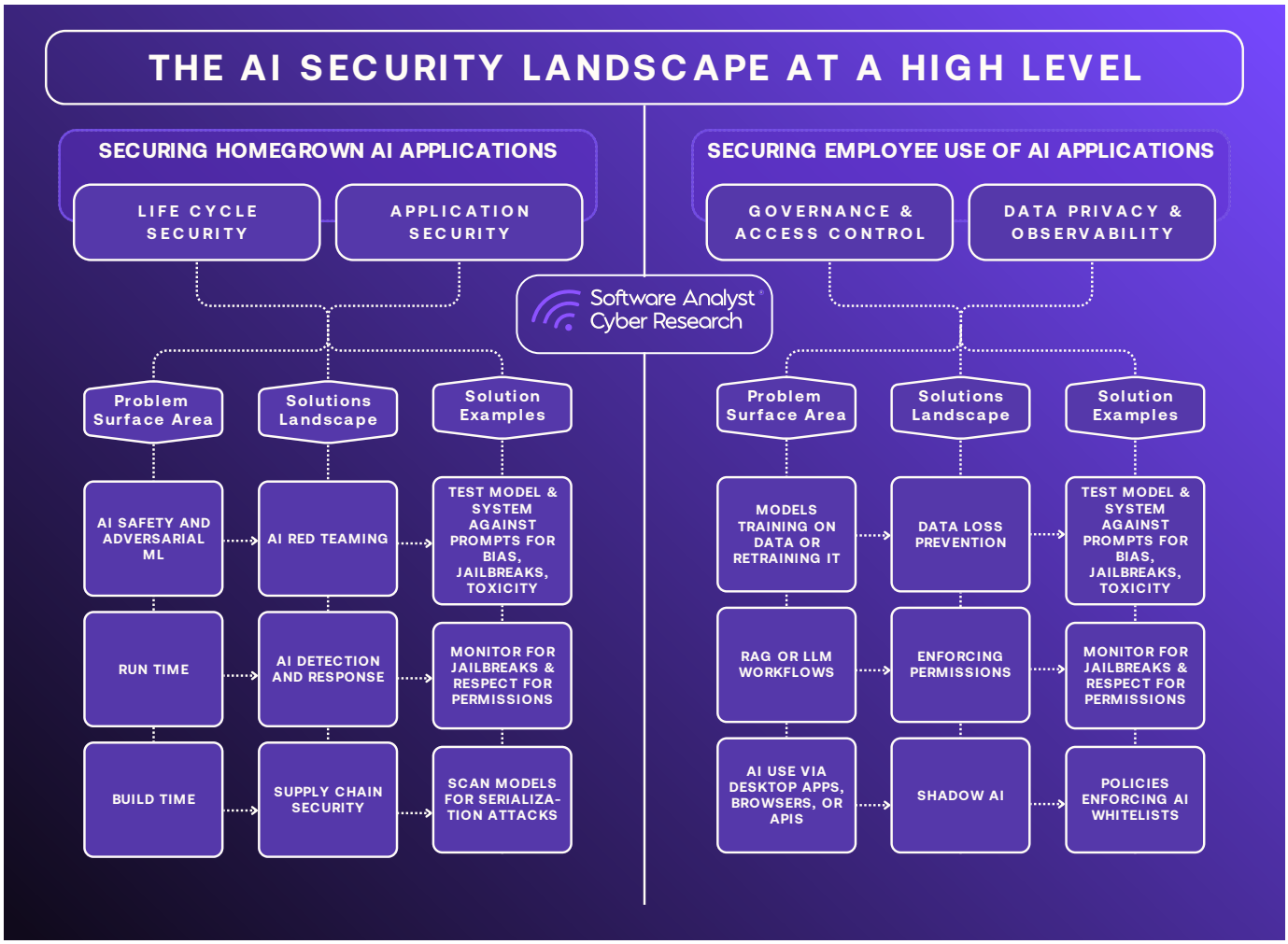
3) Runtime Security for AI

Enterprises must implement runtime security for AI models, providing real-time protection, monitoring, and response to threats during inference and active deployment. Runtime security should utilize detection and response agents, eBPF, or SDKs for real-time protection. Runtime represents a critical component of AI security. In their paper titled "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training", Anthropic researchers demonstrated how models with backdoors intentionally behave differently in training versus deployment. AI security vendors in this category identify models with hidden backdoors and monitor incoming prompt injections or jailbreak attempts. These solutions also detect repetitive API requests indicative of model theft or other suspicious activities. Vendors should additionally enforce real-time authentication for API-based AI services.

Observability is an extension of runtime security. The objective here is for enterprises to capture real-time logging and monitoring of AI activity like capturing inference requests, model responses and system logs for security analysis. It allows enterprises to track training progress and AI responses in real time. Observability serves two critical functions: it helps teams improve and debug AI chatbots, and it enables detecting harmful or unintended responses—allowing systems to throw exceptions and provide alternative responses when needed. AI Security vendors in this space help enterprises debug applications and evaluate responses effectively. The entire lifecycle from build through deployment requires continuous oversight. Due to AI's non-deterministic nature, even robust safety guardrails cannot guarantee consistent responses. These solutions should integrate with SIEM and SOAR systems for real-time incident response and analysis.

Market Landscape For Securing AI

The AI security landscape is vast, encompassing numerous challenges. The space can be divided into two primary categories:



1) Securing Employee AI Usage: Protecting How Employees Interact with AI

This first category focuses on safeguarding enterprises from risks associated with employee interaction with AI tools, such as generative AI applications (e.g., ChatGPT, Claude, Copilot). Vendors in this category primarily focus on preventing data leakage, unauthorized access, and compliance violations by monitoring AI usage, applying security policies, content moderation, and enforcing AI-specific data loss prevention (DLP).

The goal is to prevent employees from exposing sensitive information to AI models or using unauthorized AI tools.

AI security vendors in this category offer both out-of-the-box and customizable policies governing data allowed to be transmitted through AI applications. This approach, sometimes described as an “AI firewall,” allows vendors to act as a proxy on data passed to AI via APIs, browsers, or desktop tools.

These vendors also enable governance controls to ensure organizational policies are enforced on data, AI apps train on. They implement access controls (RBAC) to minimize oversharing and undersharing of information across the organization, classify for intent or catalogue how employees are using models.

2) Securing AI Models: Protecting the AI Development Lifecycle

This second category focuses on securing the AI development lifecycle, ensuring AI models are securely built, deployed, and monitored, particularly for homegrown applications. It includes protecting datasets, training pipelines, and models from tampering, adversarial attacks, and bias risks. Vendors in this space emphasize AI supply chain security, adversarial ML defences, model watermarking, and runtime monitoring to detect vulnerabilities in AI applications. These solutions are critical for organizations deploying proprietary AI models, ensuring outputs remain trustworthy, secure, and compliant with regulatory frameworks. Within this category, vendors have found success in addressing two subcategories:

1. Model lifecycle: These vendors secure AI applications from build through runtime, focusing on protecting AI models, applications, and underlying infrastructure throughout their lifecycle.
2. AI application security vendors: These players enhance traditional security approaches by addressing AI-specific concerns like non-deterministic behavior and specialized development practices, e.g. MLSecOps (Machine Learning Security Operations).

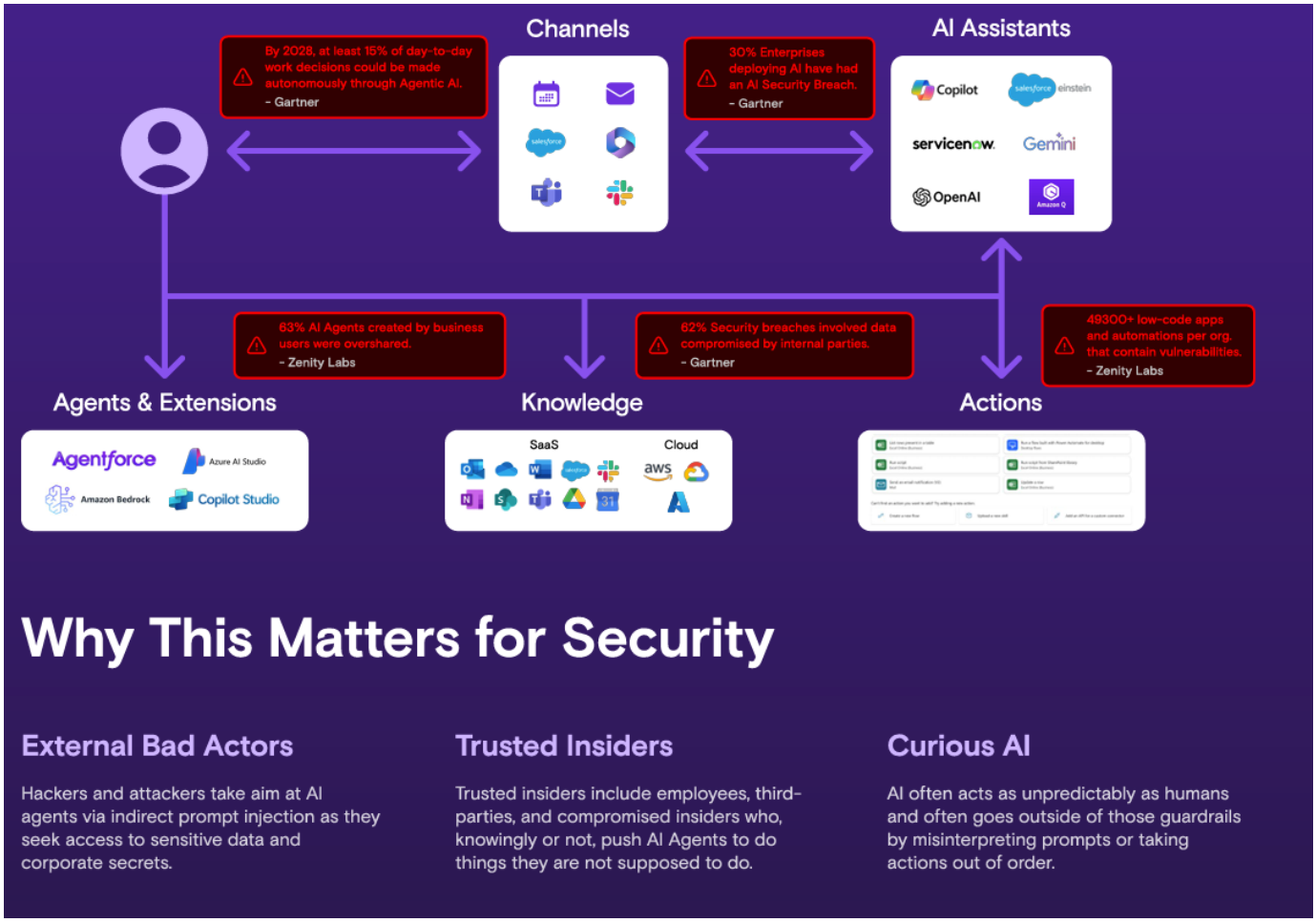

zenity

Zenity:
Securing AI Usage

Zenity

Zenity is a Series B startup based in Israel, founded by Ben Kliger and Michael Bargury in 2021. The company has experienced rapid growth and success, raising over $50 million in funding. Zenity started as a security solution for low-code/no-code development which puts them in a strategic position to be able to capitalize on AI Agents in 2025. Zenity's solution originated from its foundation in citizen development—specifically, securing no-code or low-code applications. The technology was initially built to support static analysis of business logic in no-code applications and later pivoted to address the challenges of agentic AI.

This evolution means that the product is well-suited to analyze and secure processes that are abstracted from traditional code, which many existing security tools cannot adequately scan. Zenity provides build and runtime solutions for enterprises building AI Agents, notably those made via low-code/no-code in platforms like Microsoft 365 Copilot, Salesforce and ServiceNow. They have one of the most robust solutions for securing enterprise and custom agents at both build to runtime addressing risks from when they are created to operations.



Security for AI agents is essential, especially when using low-code/no-code frameworks like Crew AI or LangChain's LangGraph. These frameworks simplify development by abstracting functionality, but this can lead to security risks. For example, agents might have unintended database query permissions. Zenity addresses this by scanning abstracted code for vulnerabilities, mapping them to OWASP Top Ten lists for low-code/no-code and GenAI Attacks Matrix. Recognizing that low-code/no-code users often lack IT or security expertise, Zenity's platform offers auto-remediation and burn down campaigns to help security teams govern and secure these AI agents.

Key product capabilities:

1. **Zenity AI-SPM / Prevention:** Zenity leverages static analysis on no-code environments. By constructing a graph of application logic (often resembling an automation flow), Zenity can identify misconfigurations and vulnerabilities in AI-driven processes. This approach differentiates it from typical static scanning tools that focus on code-level issues like buffer overflows.

   Another key feature of Zenity is the AI-SPM piece which helps enterprises understand where agents are in their enterprise and what they are connected to. These connections are shown in the platform using a security graph which makes it easy to tell what agents are connected to. Zenity provides an agentic app inventory that helps enforce permission controls, prevent oversharing, and continuously track ownership and versions. It can also monitor and alert teams to unused and orphaned resources to ensure they comply with security guidelines. Zenity's platform allows security teams to quickly understand which agents not only have the highest security risk, but also have high business criticality scores as well. This helps teams prioritize vulnerabilities and also take into account business risk.

2. **Zenity Detection and Response:** Zenity can detect direct and indirect prompt injection attacks, least privilege violations, hidden instructions and provide automated responses as needed. Zenity can also set agentic policies that help tell agents what they are allowed and not allowed to do. Zenity also provides continuous observability for AI Agents which enables direct and prompt injection attack detection, least privilege violations. It provides playbooks for teams to use or customize in case of violations such as an agent accepting an unauthenticated chat. This is helpful for responding to events in a runtime context. Zenity's runtime solution detects direct and indirect prompt injection attacks, least privilege violations, hidden instructions, and provides automated responses. In order to do this, Zenity monitors several agentic components.

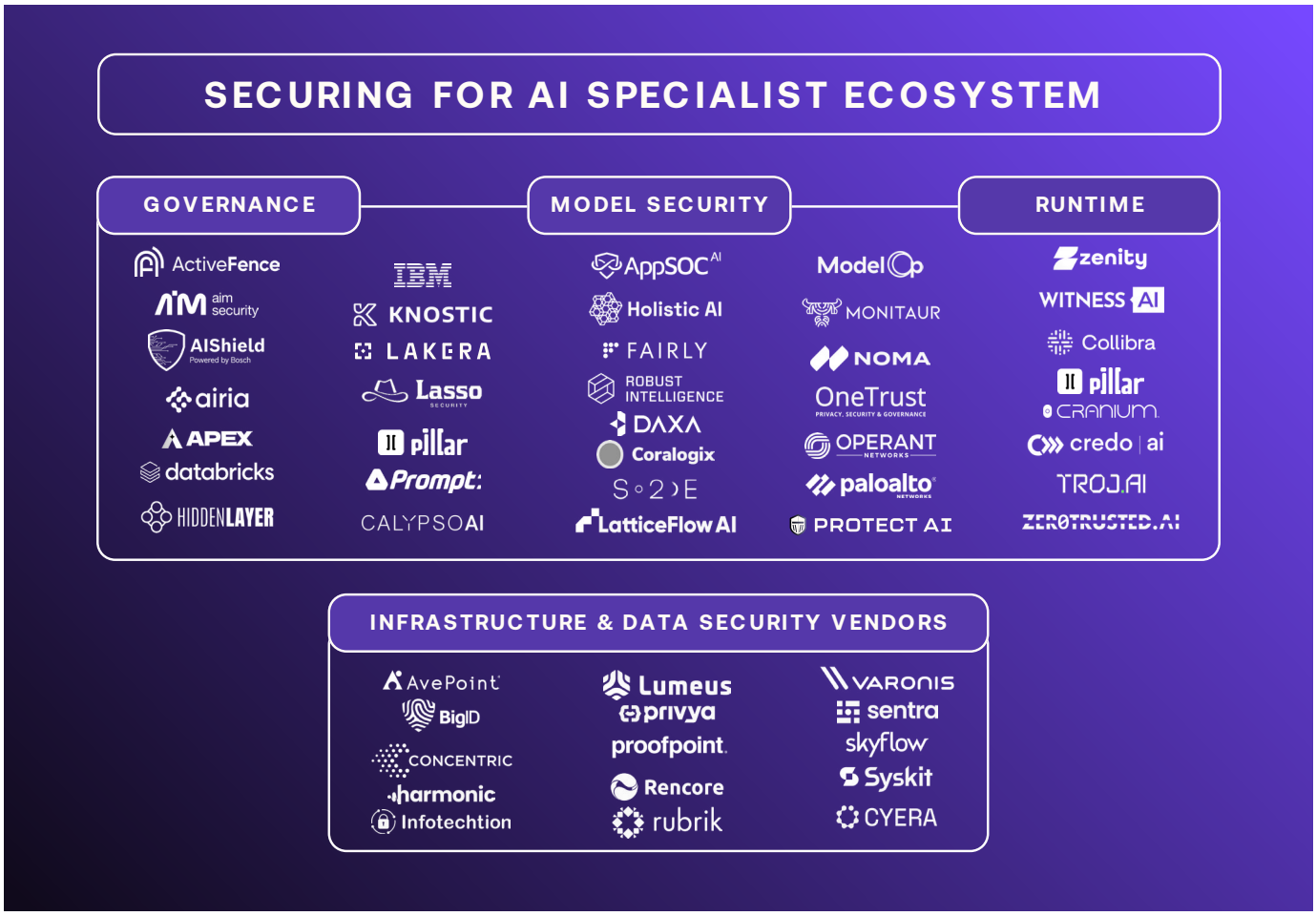
3. **Zenity Monitoring & Profiling:** Zenity monitors agent build profiles to observe vulnerabilities like over-shared agents, data leakage, and anomalous agent behavior. It also monitors multi-layer Agent actions to track the actions taken by the agent behind the scenes to facilitate the task. It will also track runtime communications which when combined with a log of agent actions provides powerful debugging and auditing capabilities. To aid in debugging, they provide visual flow diagrams. Zenity's context aware platform combined with its runtime capabilities makes it a promising choice for low-code/no-code agents. For example, Zenity would be able to detect anomalies such as unusually long emails combined with suspicious external URL requests. This agent in this example may be allowed to send email but if there are patterns that suggest this agent is exfiltrating data then that should be alerted, as a result Zenity can stop it from happening.

As organizations move from experimentation to full deployment, security teams are increasingly being tasked with integrating and managing these agentic systems. This creates a critical need for solutions that can provide deep visibility and control over AI-driven processes. Zenity's foundation in citizen development and no-code application security will help in addressing the unique challenges of securing AI agents for enterprises.

Appendix

Larger Market Map Categorizations and Illustrations

We display Gen AI attack surface and vendors across the lifecycle



Commentary on the state of the industry

The proliferation of over 50 vendors in the AI security market presents both opportunities and challenges for enterprises. On the positive side, this abundance of solutions offers organizations diverse options to address specific security needs, from model protection to runtime security and compliance monitoring. The competitive landscape drives innovation, pushing vendors to continuously improve their offerings and develop specialized solutions for emerging threats.

However, this is a nightmare for CISOs. This market fragmentation also creates significant complexities. Organizations face the challenging task of evaluating and integrating multiple solutions, potentially leading to security stack bloat and increased operational overhead. The overlap in features across vendors can cause confusion during the selection process, while integration between different security tools may not always be seamless. Additionally, the rapid emergence of

new vendors raises questions about their long-term viability and support capabilities, making it crucial for enterprises to carefully assess vendor stability alongside technical capabilities.

Furthermore, the crowded market may lead to consolidation over time, potentially disrupting existing security implementations when smaller vendors are acquired or cease operations. Organizations must therefore balance the benefits of specialized solutions against the risks of vendor volatility, while maintaining a coherent security strategy that can adapt to market changes.

What is the end state?

AI threats are evolving rapidly alongside technological advances. Microsoft's research on red teaming 100 generative AI products revealed that manually crafted jailbreaks spread quickly online, providing attackers with easy-to-use and cost-effective attack vectors compared to more complex methods. In February 2025, the Time Bandit Jailbreak attack emerged as a significant threat, affecting even OpenAI's 4o model. This exploit uses timeline confusion and procedural ambiguity to bypass safeguards, enabling users to extract potentially harmful content from the model. As AI capabilities expand, particularly with autonomous agents and browser automation, security risks grow proportionally. Microsoft aptly notes in their research: "Any tool can be used for good or ill... The more powerful the tool, the greater the benefit or damage it can cause." This principle drives their development of PyRIT, their open-source AI red teaming tool.

While it's expected that established security companies will capitalize on the growing need for AI Security solutions, there is still room for startups to enter the competition. Whether the innovative AI Security solutions of tomorrow come from established players or new startups, it's clear the winners will be those willing to innovate continuously.

Future Predictions

- **For enterprises relying on Managed AI Services** (e.g., OpenAI, Anthropic, Gemini), the primary security focus is on governing AI access, preventing data leaks, and monitoring employee interactions to ensure compliance.
- **For enterprises developing AI in-house** (e.g., Llama, Mistral, self-hosted AI), security priorities shift toward protecting data pipelines, maintaining model integrity, and securing AI infrastructure against adversarial threats.
- **Private and secure tunnels:** In our discussions with Zain, he expressed cautious optimism regarding an emerging area: companies focused on creating private and secure tunnels between AI model providers and enterprise users. This concept resembles the next evolution of Zscaler's historical role in securing traditional enterprise networks, now specifically tailored to model providers and developers. Additionally, as inference costs continue to decline exponentially, significant disruption is anticipated in legacy, compute- and data-intensive areas of cybersecurity, such as SIEM and Observability.

In conclusion, these examples and case studies emphasize that AI security is not theoretical—real incidents have occurred, ranging from accidental data leaks to exploitable vulnerabilities in AI systems. Fortunately, when properly applied, AI security solutions can mitigate these risks. We have also seen a broad range of vendors and solutions emerge to address these challenges.

To effectively combat these threats, organizations must implement comprehensive security measures that span the entire AI lifecycle—from data collection and model training to deployment and monitoring—ensuring protection at every stage.

Looking ahead, as organizations continue to share lessons learned and as industry standards evolve, AI security solutions are expected to become even more effective. The market is moving toward a future where AI can be deployed with confidence, with security, privacy, and trust embedded from the start rather than treated as an afterthought. At SACR, we'll continue to cover these developments in the coming months.

Acknowledgements

Research Collaborators

Allie Howe is the Founder of Growth Cyber. Growth Cyber helps companies build secure and compliant AI. Allie also serves as a contributor to the OWASP Agentic Security Initiative, whose goal is to provide actionable guidance to the public on securing agentic AI.

Perspectives

Zain Rizavi, an investor at Ridge Ventures, for sharing his perspectives

CISOs and security leaders from a handful of companies

17+ vendors who actively contributed insights throughout the development of this research

Research partners

This report was developed in collaboration with nine representative vendors: Palo Alto Networks, Protect AI, HiddenLayer, Noma, Pillar, TrojAI, Prompt, Zenity, and Witness AI.

Main Author - Francis is the founder and CEO of the Software Analyst Cyber Research

Software Analyst Cybersecurity Research aims to deliver the best research across cybersecurity

About Us

Software Analyst Cybersecurity Research (SACR) delivers in-depth analysis of the ever-evolving cybersecurity industry. Specializing in SOC, Identity, Network, Cloud, AppSec, and AI Security, our mission is to empower CISO's, security leaders, investors, and cybersecurity professionals with the knowledge they need to navigate this complex field.