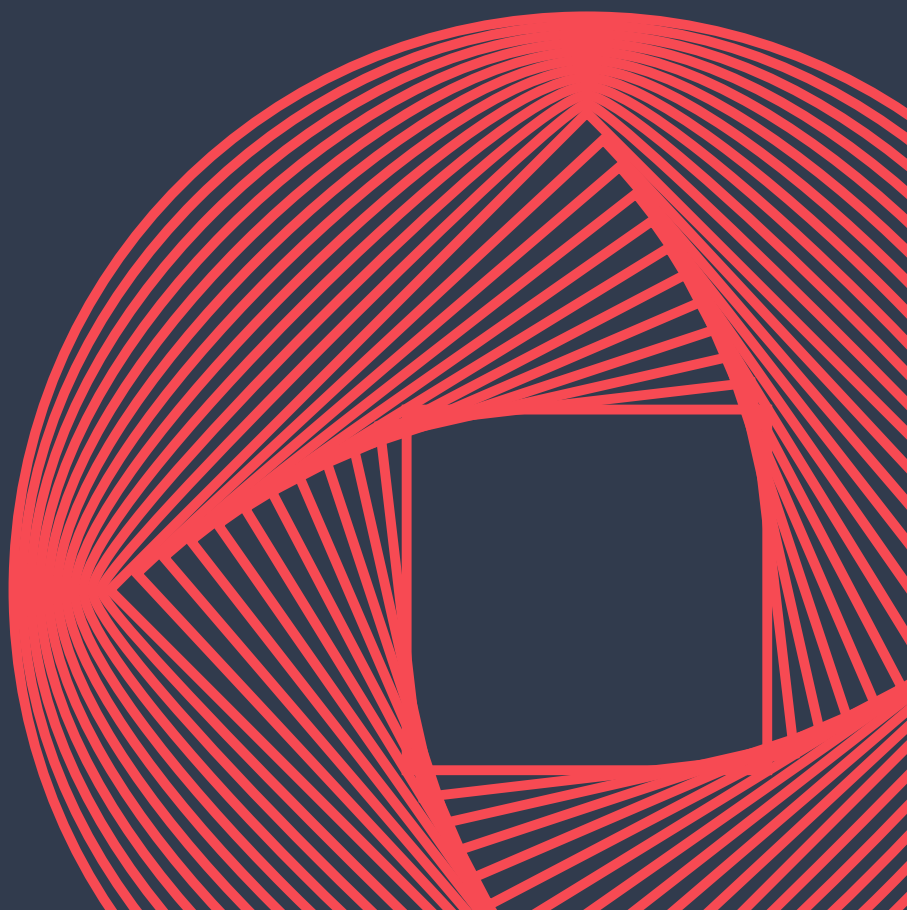
Security for AI Buyer's Guide

How to Evaluate Solutions to Secure your AI Transformation



pillar

Executive summary

With almost all organizations now leveraging AI technologies, security has emerged as a top concern among business leaders. The effectiveness of AI security hinges on the ability to identify, understand, and mitigate risks throughout the entire AI lifecycle. Staying ahead of potential threats, modern security teams must equip themselves with security platforms that anticipate and address AI-specific risks, enabling governance while not impeding rapid adoption of AI across the organization. Conventional security approaches often fall short in providing the necessary visibility, interpretability, and control; organizations require to manage AI-related risks, particularly in the face of emerging threats and evolving regulatory requirements.

To navigate the complex and rapidly changing landscape of AI security, organizations must adopt a comprehensive and adaptive approach to safeguarding their AI systems and assets.

This buyer's guide distills the essential criteria for selecting the optimal AI security solution and partner, drawing from our extensive experience working alongside AI and security leaders across various industries.

By reading this guide, you will gain insights into:

- The unique challenges of security for AI
- The new attack surface AI creates
- The foundational pillars of a robust AI security platform
- Request for Proposal template to help you evaluate your AI security vendor

Table of contents

Introduction

Generative AI (GenAI) is revolutionizing the workplace, driving innovation, competitiveness, and productivity. However, organizations face the Multiple challenges while adopting AI's potential and managing regulation, security and privacy risks.

The regulatory landscape is rapidly evolving, as exemplified by President Biden's Executive Order (E.O. 14110), the EU's AI Act, and NIST's AI Risk Management Framework. These initiatives underscore the critical need for responsible AI governance and oversight. Data, IT, and security leaders must navigate this complex terrain, balancing AI's benefits against perceived risks and uncertain accountability.

The urgency of addressing these challenges is clear. Gartner predicts that by 2026, over 80% of enterprises will have deployed generative AI in production environments, up from less than 5% in early 2023. This rapid adoption necessitates robust security measures tailored to AI's unique challenges.

This buyer's guide aims to navigate you through the complex landscape of AI security platforms. It will help you understand the key challenges, identify essential features, and make informed decisions in selecting a security solution that aligns with your organization's AI strategy while safeguarding against potential risks.

The unique challenges and risks of security for AI

AI as “Software 2.0”

AI represents a fundamental shift from traditional, logic-based software to a new paradigm often referred to as "Software 2.0." Unlike conventional software that relies on explicit programming rules, AI systems operate on goal-based models driven by data and prompts rather than code. This shift introduces unique challenges for security, as the behavior of AI models is shaped by their training data, given instructions and the objectives they are optimized to achieve, rather than by deterministic logic. Organizations must adapt to securing not just the code but also the data and meta-instructions that define AI app behavior, requiring a new approach to software assurance and risk management.

New threats and risks

AI systems introduce novel vulnerabilities that traditional security measures may not address. Adversarial attacks can manipulate input data to deceive AI models, potentially leading to incorrect outputs or decisions as well as leakage of sensitive data. These attacks can be subtle and difficult to detect, exploiting the very learning mechanisms that make AI powerful. As AI becomes more integrated into critical systems, the potential impact of such attacks grows, necessitating innovative security approaches tailored to AI's unique characteristics.

AI as a black box

The complexity of many AI models often renders their decision-making processes opaque. This lack of interpretability poses challenges for security teams trying to identify vulnerabilities or explain AI-driven outcomes. Without clear insight into how an AI arrives at its conclusions, it becomes difficult to ensure the system is behaving as intended or to detect when it has been compromised. This opacity also complicates efforts to build trust in AI systems among users and stakeholders.

New Compliance controls

The rapid evolution of AI technology has prompted governments worldwide to develop new regulatory frameworks specifically for AI systems. Recent initiatives like the EU AI Act and the Biden-Harris Administration's executive order address issues such as data privacy and ethical implications. Organizations must now implement robust monitoring and governance for their AI systems, conducting regular audits and ensuring transparency in their AI operations.

Lack of visibility

Many organizations struggle to maintain a comprehensive view of their AI deployments and usage. AI models may be developed and implemented across various departments, often without centralized oversight. This distributed nature makes it challenging to track AI assets, monitor their activities, and ensure consistent security practices. Without clear visibility, organizations risk overlooking potential vulnerabilities or failing to apply necessary security controls uniformly across all AI implementations.

New skills

The specialized nature of AI security presents a significant challenge for many security professionals. The field requires a unique blend of skills spanning traditional cybersecurity, data science, and machine learning. Many security teams lack the specific expertise needed to effectively secure AI systems, identify AI-specific threats, or implement appropriate safeguards. This knowledge gap can leave organizations vulnerable as they struggle to keep pace with rapidly evolving AI technologies and the associated security implications.

Recognizing that organizations will continue to accelerate AI implementation, security leaders must proactively guide this process, ensuring robust security and governance protocols are integrated from the start.

AI Creates a New attack Surface

The integration of AI, especially LLMs, into business processes creates novel attack surfaces that malicious actors can exploit. These new vulnerabilities extend beyond traditional cybersecurity concerns and require specialized knowledge and tools to address effectively. New threats for example:

- Prompt injection attacks
- Model and data poisoning
- Data extraction
- Adversarial examples

The Databricks AI Security Framework (DASF) provides an excellent visual representation of the new AI security landscape and its associated risks, revealing that out of these 55 risks, 35 are novel and specific to AI systems.

This framework adopts a holistic approach to mitigating cyber risks in AI systems and aligns with and complements other leading AI security frameworks, including MITRE ATLAS, OWASP LLM Top 10, and NIST guidelines.

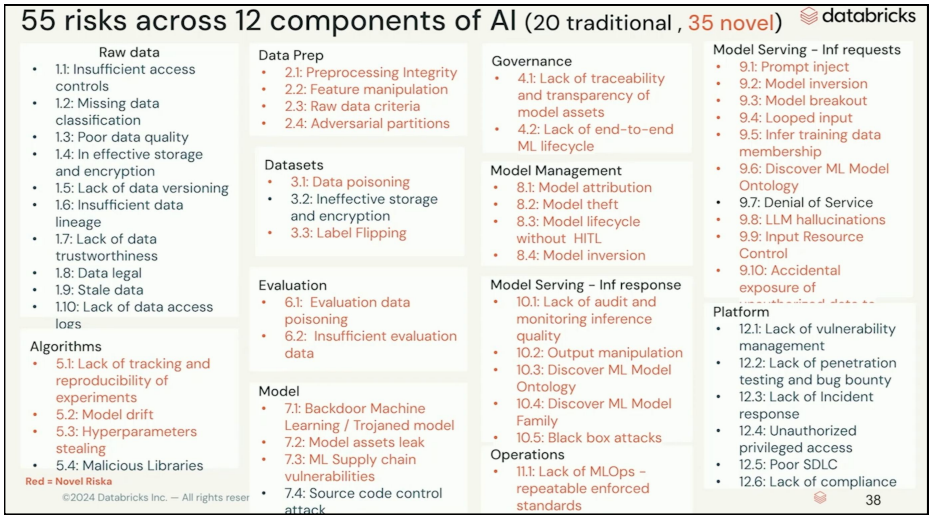


Image source: Introducing the Databricks AI Security Framework (DASF) to Manage AI Security Risks

Understanding this new attack surface is crucial for several reasons:

- Traditional security tools and practices may not be equipped to detect or prevent these AI-specific threats.
- The potential impact of successful attacks on AI systems can be far-reaching, affecting decision-making processes, data integrity, and brand reputation.
- New regulatory compliance requires organizations to demonstrate adequate protection against these new types of risks.
- As AI becomes more prevalent in critical systems, the potential for cascading failures due to AI security breaches increases.

Generative AI in Your Organization

The implementation of Generative AI (GenAI) in organizations can be categorized into three main areas: Development, Applications, and Tool usage. Each of these areas represents a different approach to incorporating AI into business processes and operations:

Development

- **Building GenAI-powered apps:** Organizations are developing their own applications using powerful language models such as GPT, Claude, and Mistral. This allows for customized AI solutions tailored to specific business needs.
- **Fine-tuning/RAG and evaluating LLMs:** Companies are adapting pre-existing Large Language Models (LLMs) to their specific use cases through fine-tuning techniques or implementing Retrieval-Augmented Generation (RAG). This process must involve rigorous evaluation of the models to ensure they meet performance and safety standards.

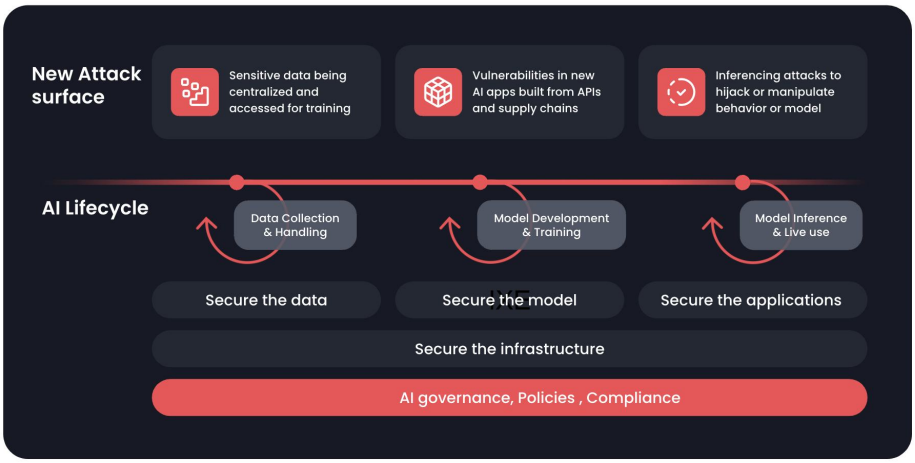
Production

- **Deploying internal or external-facing GenAI apps:** Organizations are implementing AI-powered applications for both internal use and customer-facing interactions. Examples include chatbots for customer service, data classifiers for information management, and AI agents for task automation.
- **Integrating GenAI capabilities into processes and workflows:** AI is being incorporated into existing business processes to enhance efficiency and decision-making. This could involve using AI for data analysis, document processing, or workflow optimization.

Tool Usage

- **Using public GenAI apps:** Employees are leveraging publicly available AI tools like ChatGPT or Gemini for various tasks such as content creation, research, or problem-solving.
- **Using GenAI Features in Apps or SaaS:** Organizations are adopting AI-enhanced features within existing software applications or Software-as-a-Service (SaaS) platforms. Examples include Glean, GitHub Copilot for code assistance or Microsoft 365 Copilot for productivity enhancement in office applications.

The diagram below illustrates the AI lifecycle and its associated new attack surfaces, emphasizing the critical need for comprehensive security measures throughout each stage of AI development and deployment. In the following section, we outline the primary risks for each phase of the lifecycle, accompanied by real-world examples.



Common risks for each phase in the lifecycle

For each of the stages in the AI lifecycle, there are key risks that organizations must address to ensure secure and responsible implementation of generative AI technologies. These risks vary depending on the stage of AI development, deployment, and use:

Development phase main risks:

1. **Supply chain:** This risk involves vulnerabilities in the AI development pipeline, including potential compromises in model sources, training data, or third-party components used in AI system development.

   For example: In 2022, a vulnerability was discovered in PyTorch, a popular machine learning library. The flaw allowed attackers to execute arbitrary code by exploiting a weakness in the library's handling of certain inputs.

2. **Compliance:** Ensuring that AI development adheres to relevant regulations, industry and corporate standards, and ethical guidelines. This includes data protection laws, AI-specific regulations, and responsible AI principles.

   For Example, The EU AI Act took effect on August 1, 2024, revolutionizing AI regulation worldwide. This comprehensive legislation, sets new standards for AI development and use. It poses significant challenges for tech companies who must now align their practices with the EU's stringent rules.

Production phase main risks:

1. **Jailbreaking:** The risk of malicious actors bypassing AI system safeguards or restrictions, potentially leading to unauthorized access or misuse of the AI application.

   For example: On May 25, 2023, WIRED reported on a significant security flaw in generative AI systems. The research revealed that Chatbots like OpenAI's ChatGPT and Google's Bard are vulnerable to indirect prompt injection attacks. Security researchers say the holes can be exploited to bypass content filters and manipulate AI outputs in potentially harmful ways. This vulnerability demonstrates how AI models can be exploited through crafted inputs, posing significant security challenges.

2. **Data exfiltration**: The danger of sensitive data being extracted from AI systems is a significant risk. This can occur through malicious queries, model vulnerabilities, or insider threats. Both external hackers and privileged insiders may attempt to manipulate AI models to reveal confidential information, compromising intellectual property, personal data, and business-critical information

   For example: In November 2023, researchers discovered that OpenAI's custom GPTs were vulnerable to data leakage. By using simple prompt injection techniques, they could extract the initial instructions and upload files used to create these chatbots with nearly 100% success rate.

Tool Usage phase main risks:

1. **Sensitive data leakage:** The inadvertent exposure of confidential or personal information through AI system outputs or interactions, potentially violating privacy regulations or compromising organizational security.

   For example: In July 2023, Samsung faced issues with its AI-powered chatbot on its Korean website. Users reported that the chatbot was revealing sensitive information about Samsung's internal operations and upcoming products. This incident, while not a traditional data breach, demonstrated how AI systems can inadvertently expose confidential data, highlighting the risks associated with deploying AI in customer-facing applications without proper safeguards.

2. **Shadow AI:** The unauthorized or uncontrolled use of AI tools within an organization, potentially introducing security vulnerabilities, compliance issues, or inconsistencies in AI application across the enterprise.

   For example: On May, 2024, SC Media reported that 'Shadow AI' is on the rise, with sensitive data input by workers increasing by 156%. The article suggests that AI use in the workplace is growing exponentially, often without proper oversight or authorization. This trend of employees using unsanctioned AI tools poses significant security and compliance risks for organizations, as these applications may not meet company security standards or regulatory requirements.

The foundational pillars of a robust AI security platform

To effectively secure and manage AI systems in your organization, your chosen security platform must provide robust capabilities across four essential pillars. These pillars form the foundation of a comprehensive AI security strategy, addressing the unique challenges and risks associated with AI technologie across the entire lifecycle:

Visibility and Inventory Management

Providing clear insights into AI asset usage and management across the enterprise. Mapping of models, datasets and tools in use. In-depth logging of all app interactions for auditing and incident response.

Adversarial Resistance and Model Robustness

Strengthening AI models against potential attacks and enhancing their ability to handle unexpected inputs and threats.

User, Data, and Application Protection

Safeguarding the integrity of AI systems, user interactions, and sensitive data.

Governance, Compliance, and Transparency

Ensuring AI systems adhere to regulatory requirements and organizational policies while maintaining transparency in decision-making processes.

Translating these pillars into a practical checklist, the following section identifies the key use cases and capabilities you should evaluate when choosing a Security for AI platform.

Request for Proposal template

Selecting an AI security vendor is a critical decision that can impact your organization's risk posture and AI initiatives. We've developed a comprehensive Request for Proposal (RFP) template. This template is designed to help you thoroughly evaluate potential AI security vendors, ensuring the chosen solution aligns with your specific AI adoption use cases.

Requirement	Vendor Coverage
AI Asset Discovery and Inventory	
Automatically detect and catalog AI/ML models across your environment	
Classify models based on type, framework, and potential risk level	
Identify unauthorized or shadow AI deployments within the organization	
Provide a centralized inventory of all AI assets for governance and compliance	
Adversarial Resistance and App Robustness	
Harden AI systems against adversarial examples and LLM-focused attacks	
Evaluate model robustness using pre-deployment testing	
Implement techniques to improve model resilience to malicious inputs	
Provide ongoing assessment of model vulnerability to emerging attack vectors and as new models are leveraged	
Secure LLM Integration	
Protect integrations between LLMs, enterprise applications, and data repositories	
Implement secure API gateways for LLM interactions	
Provide fine-grained access controls for LLM usage within the organization	
Monitor and log all LLM interactions for security and compliance purposes	

Request for Proposal template

Requirement	Vendor Coverage
Content Anomaly Detection and Filtering	
Detect and block unacceptable inputs that could compromise enterprise decision-making and confidentiality	
Filter out inaccurate, hallucinatory, copyrighted, illegal, or otherwise unwanted outputs	
Support dynamic, prompt-aware, and contextual policy enforcement based on user roles and departments	
Integrate AI-based detection with traditional techniques like keyword matching and rule engines	
Data Protection Mechanisms	
Prevent data leakage and compromise of confidential data used in LLMs, both with the LLM provider and internally	
Provide data obfuscation capabilities on-demand to mask sensitive information	
Ensure protection of training data against poisoning attacks	
Implement strong access controls and encryption for AI-related data	
Runtime Application Security	
Defend against prompt injection attacks that attempt to bypass filters and policies	
Secure APIs and integrations between LLMs, vector databases, and enterprise apps	
Prevent unauthorized access to model states, parameters, and architecture	
Provide runtime monitoring for detecting anomalous model behavior and drift	
Governance, Risk, and Compliance Alignment	
Support evolving AI governance structures and policies	
Provide model validation, testing, and audit capabilities	
Integrate with IT GRC and security operations workflows	
Align AI security measures with leading frameworks, including the OWASP Top 10 for LLMs and MITRE ATLAS	

Request for Proposal template

Requirement	Vendor Coverage
Audit and Enforcement	
Provide detailed audit logs of AI model usage, inputs, and outputs	
Enable real-time monitoring of AI system behavior and performance	
Implement automated policy checks to ensure compliance with AI governance frameworks	
Generate customizable reports for stakeholders on AI risk posture and policy adherence	
Support forensic analysis capabilities for investigating AI-related security incidents	
Adaptive Threat Response	
Offer regular updates to address the rapidly evolving AI threat landscape	
Provide threat intelligence specific to AI and ML systems	
Implement automated threat detection and response mechanisms	
Support integration with existing security information and event management (SIEM) systems	
Flexible Deployment Options	
Support various generative AI deployment models from provider-managed to self-managed implementations	
Offer hybrid deployment options to balance security and performance requirements	
Provide scalability to meet enterprise performance needs	
Ensure compatibility with major generative AI/LLM platforms	

pillar

Build, run, and use AI with confidence

About Pillar

Pillar offers a groundbreaking, proactive approach to securing AI systems and activities, providing comprehensive protection throughout the entire AI lifecycle. Our platform addresses the key capabilities outlined in this buyer's guide, delivering robust solutions for in-depth logging, risk assessment, advanced adversarial resistance, and adaptive protection.

Pillar's Unique Approach:

Cyber and AI experts

Pillar's founding team comprises experts with extensive backgrounds in adversarial cybersecurity, threat intelligence, and AI. Our management team brings over 50 years of combined cybersecurity experience. This unique combination of deep expertise in both AI and security is crucial for delivering robust protection against the complex, evolving threat landscape.

Powered by Real-World Threat Intelligence

Leveraging strategic integrations with leading LLM ecosystem solutions, our proprietary detection and evaluation engines are trained on large datasets of real-world attacks. This enables us to analyze and process vast amounts of app interactions from thousands AI applications and over 500,000 chat conversations. Our comprehensive approach delivers highly contextual alerts with minimal false positives.

End-to-End AI Lifecycle Security

Pillar provides a single, integrated platform to secure the entire AI lifecycle - from development through production to usage. This comprehensive approach offers unparalleled visibility and control, ensuring real-time protection and compliance at every stage of the AI journey.

Proactive & adaptive security

Our proprietary red teaming capabilities empower teams to identify and mitigate real AI app risks continuously and automatically. This proactive stance allows for ongoing monitoring of changes in apps or models, routinely re-evaluating exposure posture to stay ahead of emerging threats.

www.pillar.security

pillar