White paper Cisco public

Comparing AI Application Security to Traditional Cybersecurity

Contents

What is AI security?

The rapid evolution and proliferation of artificial intelligence is already reshaping many aspects of our daily lives, including how we work. AI systems leverage technologies like machine learning and natural language processing to perform tasks that would typically require human intelligence, unlocking a myriad of new personal and enterprise AI applications. However, this disruptive technology has also introduced a slew of new safety and security challenges for organizations to contend with.

The new threat landscape of AI upends many longstanding principles of cybersecurity. AI brings fundamentally new threats and vulnerabilities to software that existing tools and processes don't address. As AI applications serve more critical functions and handle greater volumes of sensitive data, bad actors and nation states are increasingly motivated to target them. Effective AI security requires a paradigm shift that considers unique risks while leveraging solutions purpose-built to mitigate them.

AI security refers to the set of measures that protect artificial intelligence systems from threats and vulnerabilities that might compromise their functionality or the data they handle.

Despite widespread AI adoption, only 45% of organizations report having the resources and expertise to conduct comprehensive AI security assessments, while 86% have already experienced AI-related security incidents

Source: cisco.com

AI applications vs. Traditional applications

There are several fundamental differences between AI applications and traditional applications. These differences make AI security unique and particularly complex when compared to the traditional, well-established cybersecurity practices used in the past.

At a high level, there are six primary distinctions between AI applications and traditional applications:

1. **AI applications are largely non-deterministic.** Traditional applications are designed to support a specific set of operations, and their outputs are consistently determined by preceding user inputs. On the other hand, Large Language Models (LLMs) rely on natural language and are trained on vast, complex datasets, so one singular input can lead to inconsistent results. This also means bad actors can reuse malicious prompts with varying outcomes.

2. **Attacks on AI applications exploit new and unique vulnerabilities.** Threat actors may target various components of AI architecture, such as models, agents, Model Context Protocol (MCP) servers, datasets, and pipelines, with techniques that are less understood or entirely unprecedented. This contrasts with attacks on traditional applications, which largely exploit known vulnerabilities or processes.

3. **AI applications blur the boundaries between data and code.** In traditional applications, users can take specific, predefined actions to interact with specific portions of the data–clicking a button to read and modify a financial report, for example. AI effectively erases those boundaries as training data is codified into the models, and developers have less control over what end users are privy to.

4. **Threat actor interest in AI is still being realized.** Traditional applications are standardized,and adversary interest is defined by the application's purpose and the data it contains. Interest in AI applications is less clear and varies based on deployment and capabilities. Motivations can include misuse, data theft, repurposing, and exploitation of AI systems. Even in situations where an AI application is deployed for internal use, insiders can use them to streamline information gathering and data exfiltration.

5. **AI applications require purpose-built security measures.** Traditional cybersecurity practices for systems such as networks, endpoints, and cloud applications are more well-understood, and a myriad of solutions already exist to address these needs. As an automated, agentic technology, AI introduces unique risks that require purpose-built security solutions to address.

6. **AI applications can act autonomously (agentic behavior).** Unlike traditional applications and even earlier AI models, modern agentic systems can initiate tasks, call APIs, retrieve documents, or take actions without direct user input. This autonomy introduces new security concerns, such as chaining actions that lead to unanticipated harm, triggering external effects, or escalating access.

Recent Cisco® studies have shown that 86% of companies deploying AI have already experienced an AI security incident such as data compromise or an attack on AI infrastructure in the past 12 months.

Source: Cisco Cybersecurity Readiness Index

AI application security vs. Traditional application security

Despite the many ways in which AI differs so radically from existing technologies, it is still fundamentally software. As such, it still relies heavily on familiar cybersecurity measures to protect underlying infrastructure, control access, and more.

In fact, AI security borrows many familiar concepts from traditional cybersecurity but implements them in new and unique ways.

Below are some of the most prominent categories of application security. In each section, we'll examine how these concepts have been historically applied to traditional applications, then adapt them to the new paradigm of AI.

Open-source scanning

Traditional applications	AI applications
Software Composition Analysis (SCA) plays an important role in secure application development. SCA tools identify open-source dependencies in an application, detailing them in a Software Bill of Materials (SBOM). These dependencies are then analyzed to find any potential risks or known vulnerabilities. With modern software so reliant on third-party components, this is an integral application security practice.	AI application development relies heavily on components such as open-source models, public datasets, and third-party libraries. These dependencies can include vulnerabilities or malicious insertions that compromise the entire system. File scanning and model validation tools can proactively identify security vulnerabilities in open-source components of the AI supply chain, like models or MCP servers provisioned from Hugging Face. This enables developers to build AI applications with greater confidence. In addition, organizations are beginning to adopt an AI Bill of Materials (AI BOM) to provide visibility into these components, similar to how a SBOM is used for traditional software.

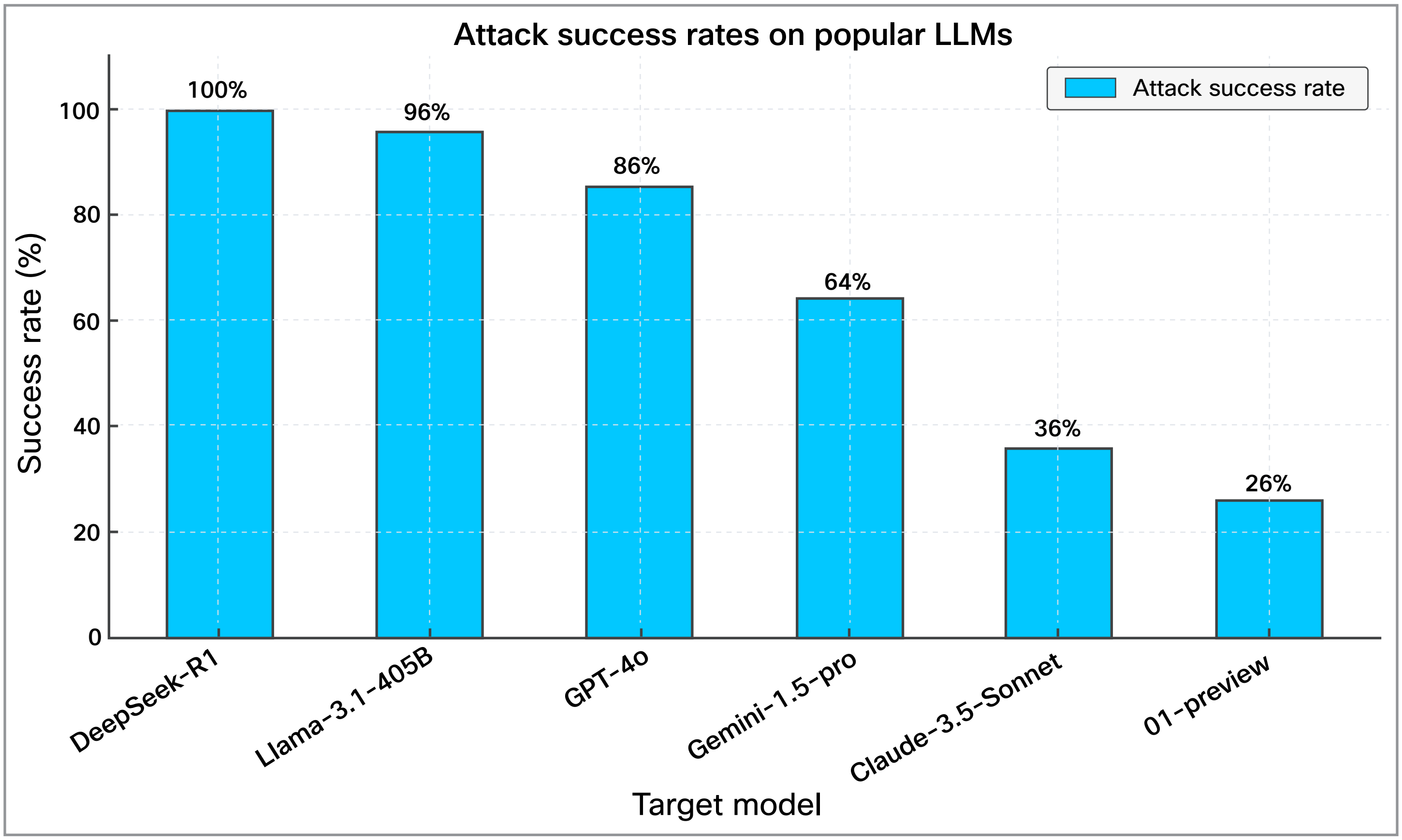
Case Study

DeepSeek Vulnerability

To evaluate the risks posed by open-source models, Cisco AI Defense conducted algorithmic red teaming against DeepSeek, a widely used large language model, along with several other popular frontier models. The test involved over 100 adversarial prompts spanning prompt injection, data leakage, impersonation, and other high-risk categories.

The result: DeepSeek failed every test with a 100% attack success rate. Compared against other commercial and open-sourced models, DeepSeek showed the highest rate of vulnerability to these threats.

This test highlighted the importance of pre-deployment model validation and the limitations of relying on native model guardrails alone. By catching weaknesses early, organizations can apply targeted protections before models reach production, closing security gaps that traditional scanning tools would miss.

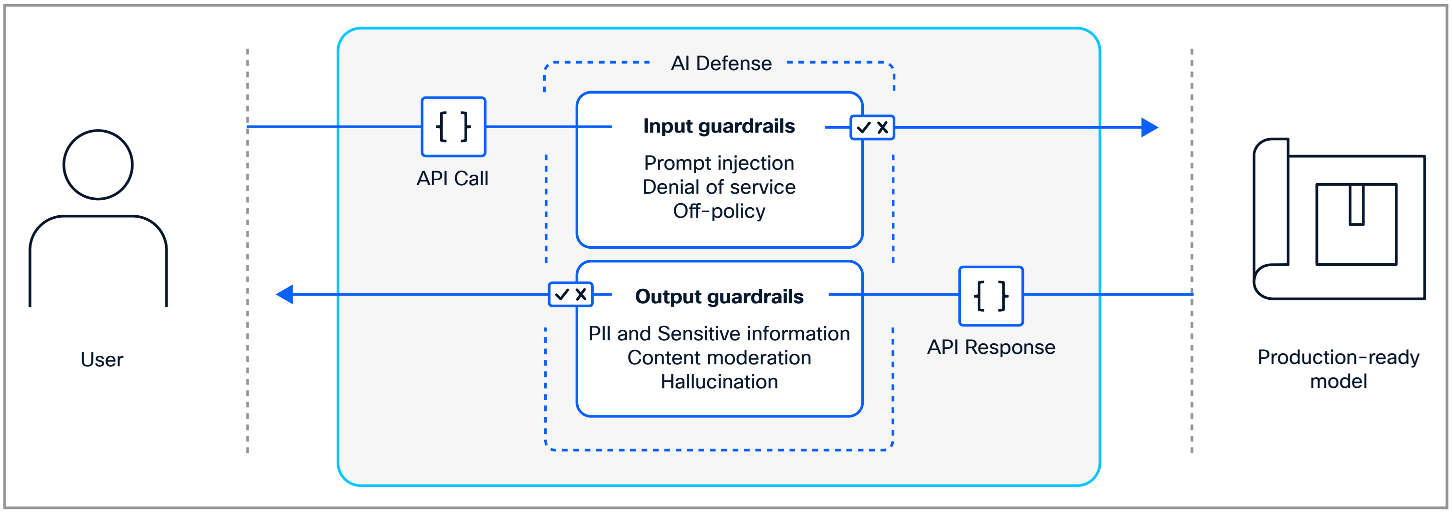


Vulnerability testing

Traditional applications	AI applications
Static and Dynamic Application Security Testing (SAST and DAST) are two complementary methods for software vulnerability testing. Static testing requires source code access and enables developers to identify and remediate vulnerabilities early. Dynamic testing is a black box methodology that evaluates software while it is running to discover vulnerabilities the same way an external adversary might.	Static testing for AI applications involves validating the components of an AI application, such as binaries, datasets, and models, for example, to identify vulnerabilities like backdoors or poisoned data. Dynamic testing for AI applications evaluates how a model responds across various scenarios in production. Algorithmic red-teaming can simulate a diverse and extensive set of adversarial techniques without requiring manual testing.

Application firewalls

Traditional applications	AI applications
Web Application Firewalls (WAF) act as barriers between traditional web applications and the internet, filtering and monitoring HTTP traffic to block malicious requests and attacks like SQL injection and Cross-Site Scripting (XSS). These reverse proxy solutions operate based on a set of defined policies that can be easily modified to cover new vulnerabilities or reflect unique security requirements.	The emergence of generative AI applications has given rise to a new class of AI guardrail solutions designed to protect against real-time threats to AI systems. These solutions, effectively serve as model-agnostic guardrails, examining AI application traffic bi-directionally to identify and prevent various failures and attacks. This enables teams to enforce policies and mitigate threats to AI applications such as Personal Identifiable Information (PII) leakage, prompt injection, and Denial of Service (DoS) attacks.



Data loss prevention

Traditional applications	AI applications
Data Loss Prevention (DLP) solutions prevent the exposure of sensitive data through negligence, misuse, or exfiltration. Different forms of DLP exist to cover networks, endpoints, and the cloud. DLP comprises various tools to help with data identification, classification, monitoring, and protection. The effectiveness of these solutions relies heavily on sufficient visibility, accurate classification, and robust policy implementation, among other things.	The rapid proliferation of AI and the dynamic nature of natural language content required a change to how enterprises approach DLP. DLP for AI applications examines inputs and outputs to combat sensitive data leakage through malicious actions and benign or inadvertent interactions. Input DLP includes policies that restrict file uploads, block copy-paste functionalities, or restrict access to unapproved AI tools altogether. Output DLP uses guardrail filters to ensure model responses do not contain PII, intellectual property, or other forms of sensitive data, helping protect against intentional exfiltration and accidental disclosure.

Cisco research shows that external threats are responsible for **58%** of AI-related security incidents, compa red to **42%** caused by internal vulnerabilities, indicating that external attack vectors remain the dominant risk in AI deployments.

Source: Cisco Cybersecurity Readiness Index

The need for AI-designed security solutions

Traditional cybersecurity solutions will continue to play a vital role in protecting AI infrastructure, managing user access, and addressing the variety of other security requirements inherent to all software.

Still, when it comes to the security of AI applications, traditional solutions fall short. New supply chains, new development processes, and a myriad of new safety and security risks require a solution that is purpose-built for AI.

The threat landscape of AI is rapidly evolving as adversaries and threat researchers continue to push boundaries and uncover new vulnerabilities. A dedicated AI security solution effectively decouples security from development to offer better, more flexible defenses that are continuously updated to stay ahead of emerging threats.

According to McKinsey, **78%** of organizations are already using or implementing AI in at least one business function, an increase from **72%** in 2024 and **55%** the year prior.

Source: McKinsey - The State of AI Report

How Cisco AI defense protects AI applications

Cisco AI Defense provides comprehensive, end-to-end protection for AI applications from initial sourcing to real-time deployment. The platform is built around complementary components:

AI Cloud Visibility
Gain complete visibility into AI assets across public cloud environments. Cisco AI Defense automatically inventories models, agents, datasets, and knowledge bases, including those hosted externally or provisioned by individual teams. This enables organizations to understand usage, track exposure, and apply security policies consistently, even in decentralized environments.

AI Supply Chain Risk Management
Modern AI development relies heavily on open-source models, datasets, MCP servers, and third-party APIs. Cisco AI Defense helps organizations scan these components for vulnerabilities, malicious insertions, or poisoned data. By integrating model validation into the development pipeline, organizations can build with greater confidence and mitigate upstream risk before it propagates.

AI Model and Application Validation
Cisco's automated red teaming tests models against hundreds of attack techniques from prompt injection and data leakage to impersonation and harmful content. These assessments are informed by the latest research from Cisco's AI security team and integrate seamlessly into existing CI/CD workflows.

AI Runtime Protection
Real-time, model-agnostic guardrails monitor and enforce policies at the point of execution. Inputs and outputs are inspected bi-directionally to prevent unsafe, non-compliant, or adversarial behavior. These protections are continuously updated and can be customized to align with organizational risk profiles. In addition to models, runtime protection is expanding to cover MCP guardrails, ensuring that security extends beyond the model itself.

Ready to bring end-to-end security to your AI applications?

Get started with a demo of Cisco AI Defense.