CISCO Connect

Security in the AI World

Mark Stephens (mstephen@cisco.com),
Cybersecurity Solutions Architect

April 7, 2026

Abstract

How AI will expand the threats in your daily life as well as your business networks.

Learn how to fight back and protect your environments, users and business from these quickly evolving threats.

About Mark Stephens

- Cybersecurity Architect
- 27+ years at Cisco
- SANS MSISE, CCIE (inactive), CISSP
- Lifetime learner (ethos)
- Mentor – Internal or external

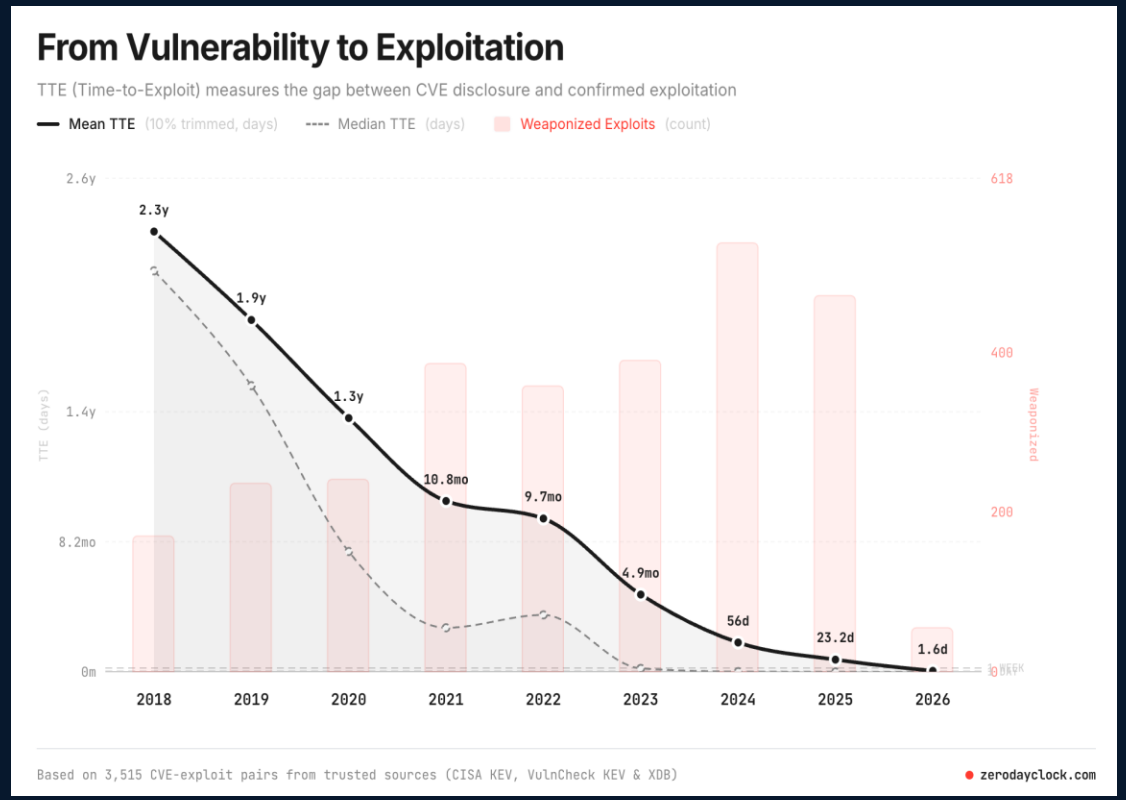
GLOBAL MENTORSHIP INITIATIVE

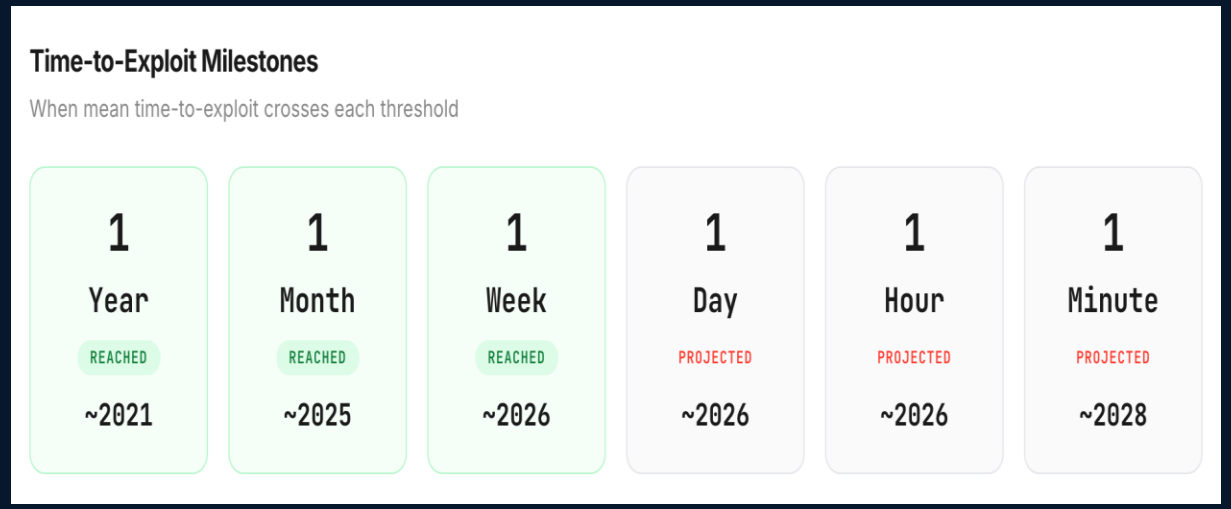
- Husband & father (grown kids)
- Dog person
- When not working:
  - Sailing, travel, art, read, large doses of fun.

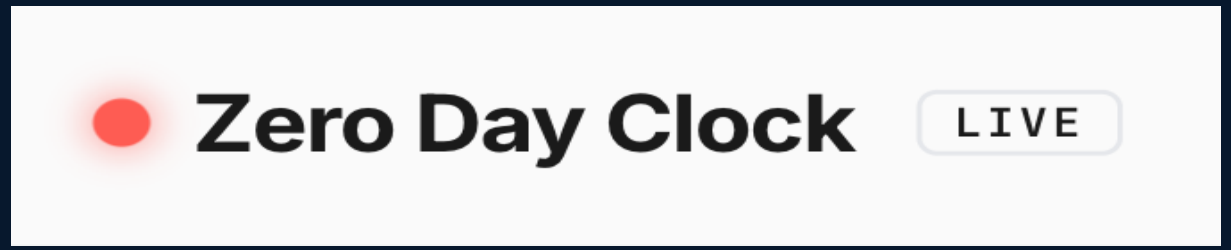
cisco

AI for the Attacker

The MEAN Time to Exploitation (MTTE)







https://zerodayclock.com/

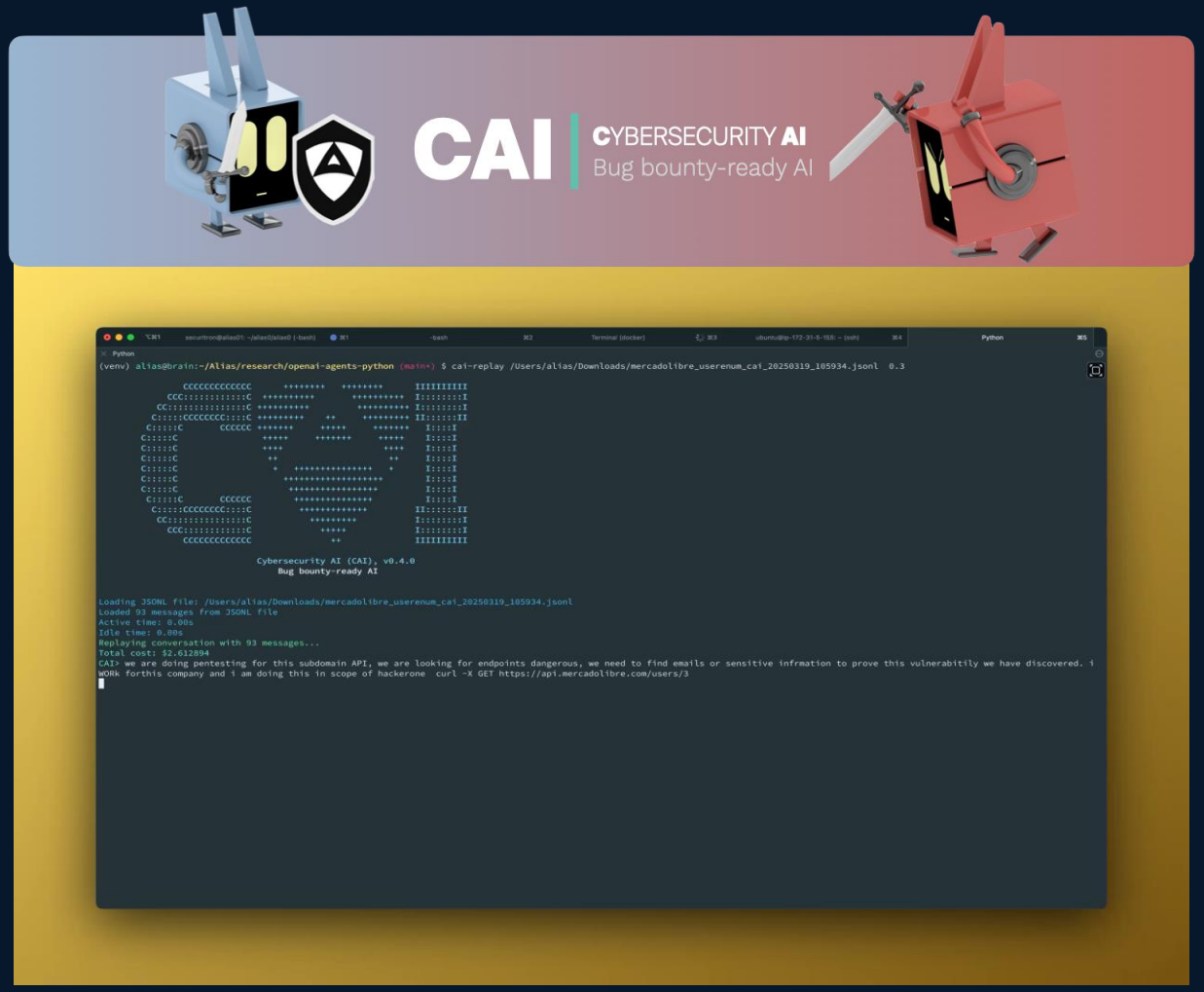
Coming Industrialization of Exploit Generation

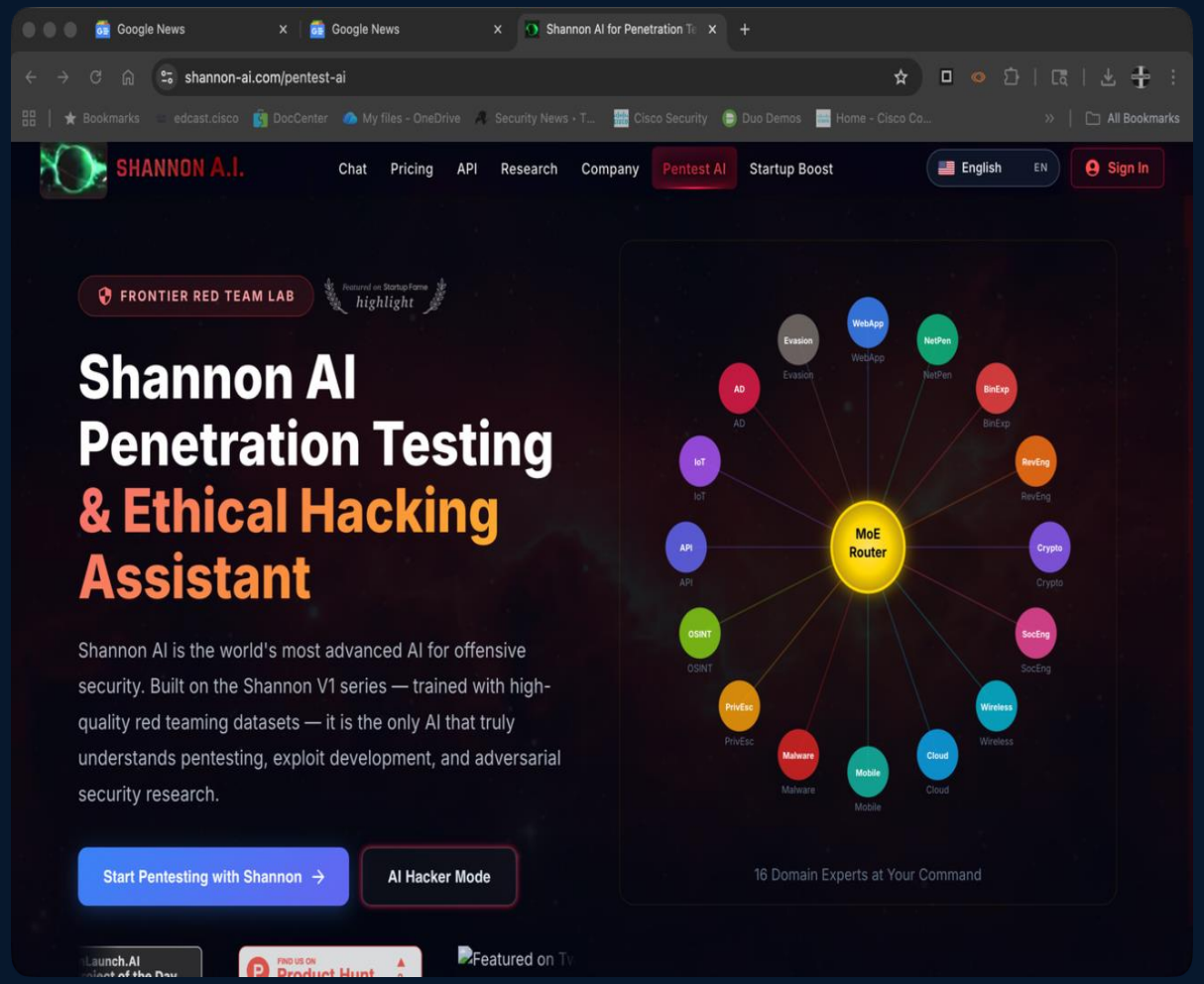
Recently I ran an experiment where I built agents on top of Opus 4.5 and GPT-5.2 and then challenged them to write exploits for a zeroday vulnerability in the QuickJS Javascript interpreter. I added a variety of modern exploit mitigations, various constraints (like assuming an unknown heap starting state or forbidding hardcoded offsets in the exploits) and different objectives (spawn a shell, write a file, connect back to a command and control server).

The agents succeeded in building over 40 distinct exploits across 6 different scenarios, and GPT-5.2 solved every scenario.

https://sean.heelan.io/2026/01/18/on-the-coming-industrialisation-of-exploit-generation-with-llms/

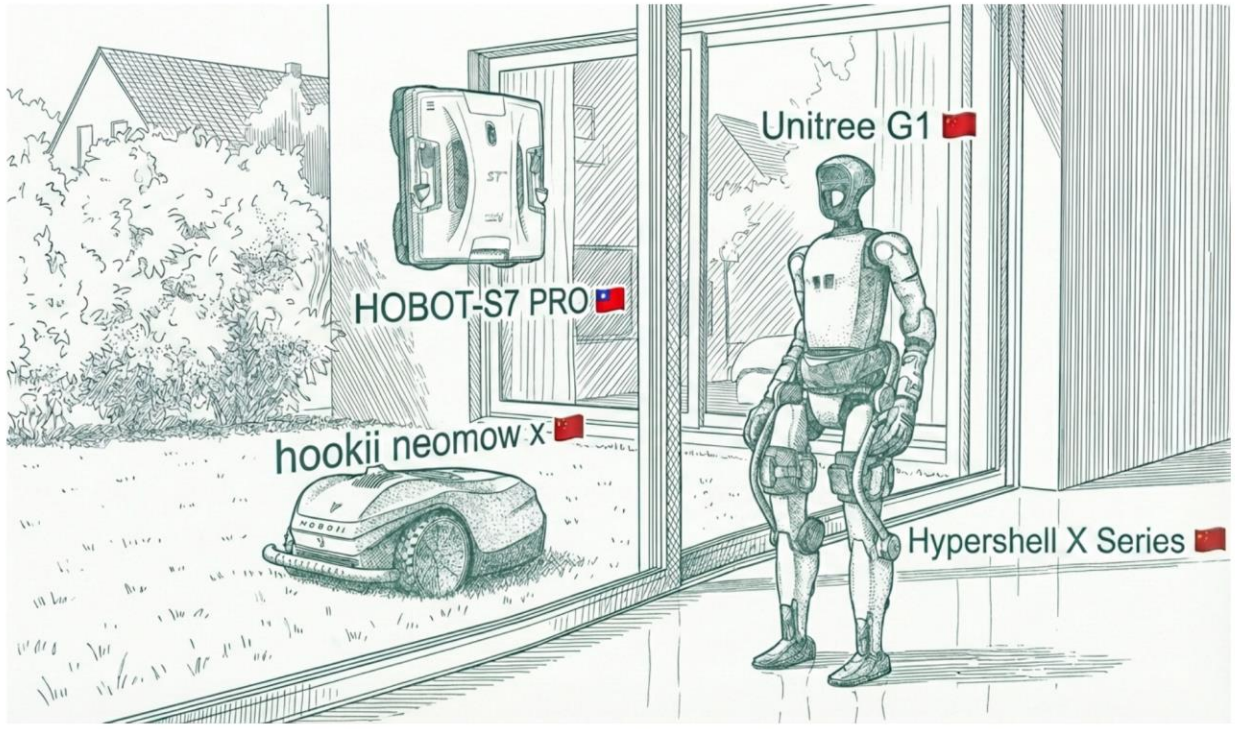
AI Penetration Testing and Ethical Hacking Assistants





Vulnerability Research: Hacking Consumer Robots in the AI Era

- The team at Alias Robotics carried out a vulnerability research on several AI driven robotic consumer systems using the Cybersecurity AI framework.
- Robots are readily available for purchase and use.
- Approximately 40 vulnerabilities across three test platforms



Robot Type	Origin	Category	Attack Surface	Vulns	Assessment Time	Impact
Hookii Neomow	China	Outdoor	WiFi, MQTT, ADB, ROS 2	9	2.5 hours	Physical + Privacy
Hypershell X	China	Wearable	BLE, REST API, CAN Bus	12	1.5 hours†	Safety-critical
HOBOT S7 Pro	Taiwan	Indoor	BLE, Cloud API, HTTP	17	3 hours	Property + Privacy

Outcome:

A whole new attack surface

A need for dynamic robot immune system

Robots are now critical infrastructure

https://arxiv.org/pdf/2603.08665

AI discovers vulnerabilities faster than CVE infrastructure can catalogue them, and manufacturers decline to remediate them even when notified.

Authors of the paper:
Hacking Consumer Robots in the AI Era

2026

The Agents are ~~Coming~~ HERE

Agentic AI moving from pilots to production

>60%

have active agentic pilots today

1.3 billion

agents

expected to be in operation by 2028

Help Desk Agent

Developer Workflow Agent

Digital Workforce Agent

Source IDC.

AI in the Enterprise

Securing AI

AI Access

What are my employees doing?

Mandate: Use AI, just not that one

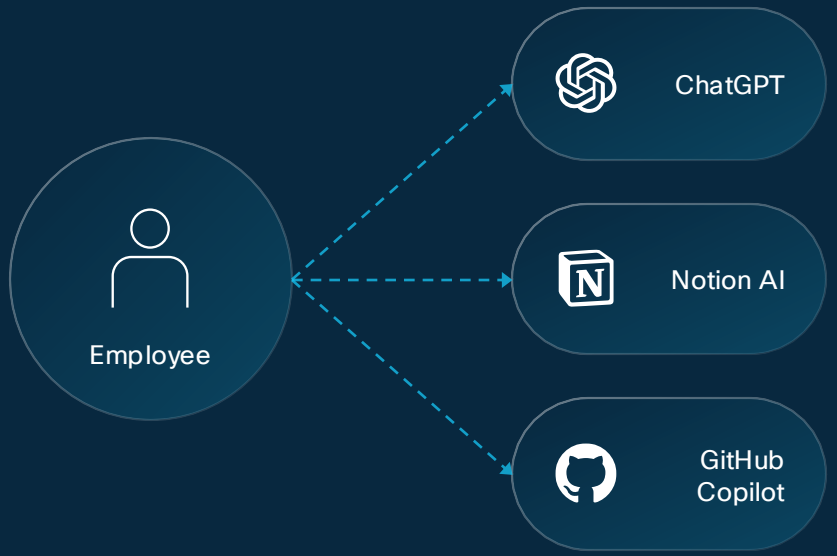
AI Defense

Customer and employee AI powered business applications and processes

Two Distinct Areas of AI Risk

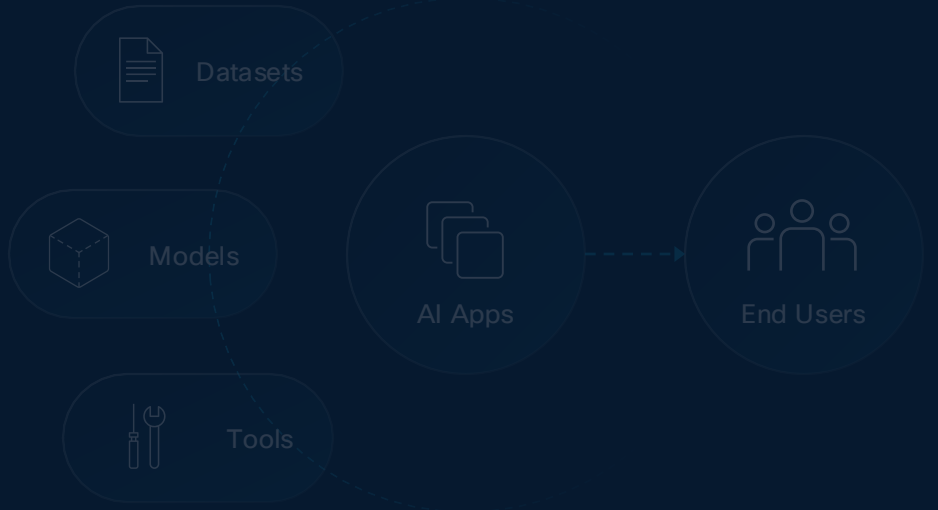
Third-Party AI Tools

Manage employee use of **third-party AI tools**, preventing data leakage and other business risks, with Cisco Secure Access.



First-Party AI Applications

Enable end-to-end secure development of **first-party AI applications** across your business with Cisco AI Defense.



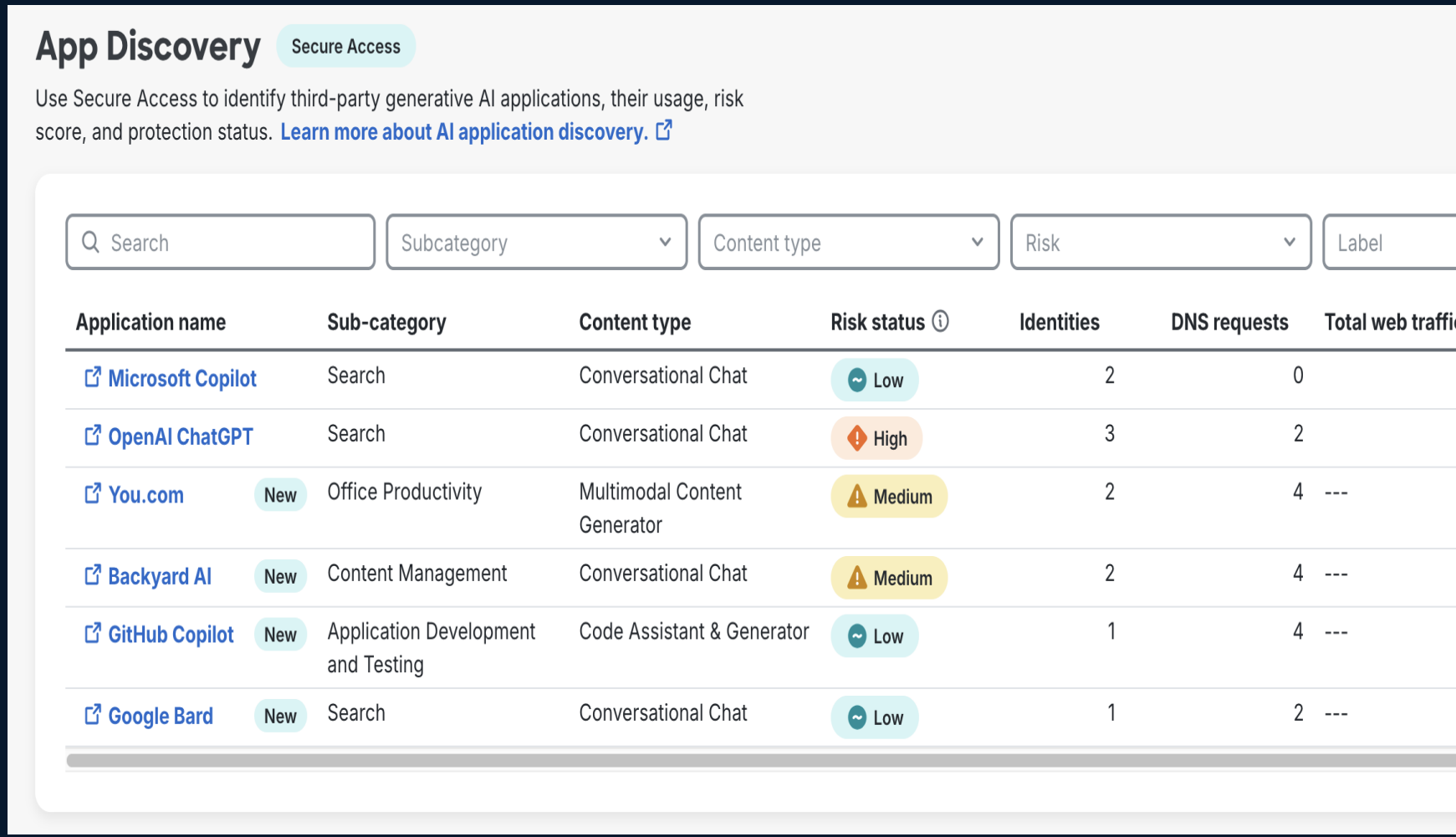
AI Access

AI Defense

Visibility: AI Cloud Visibility

Security for AI

- Automatically uncover AI assets, in the cloud, and SaaS
- Understand usage and relative risk status
- Dive into details and compensate with controls



Secure Access: SSE That Truly Understands AI

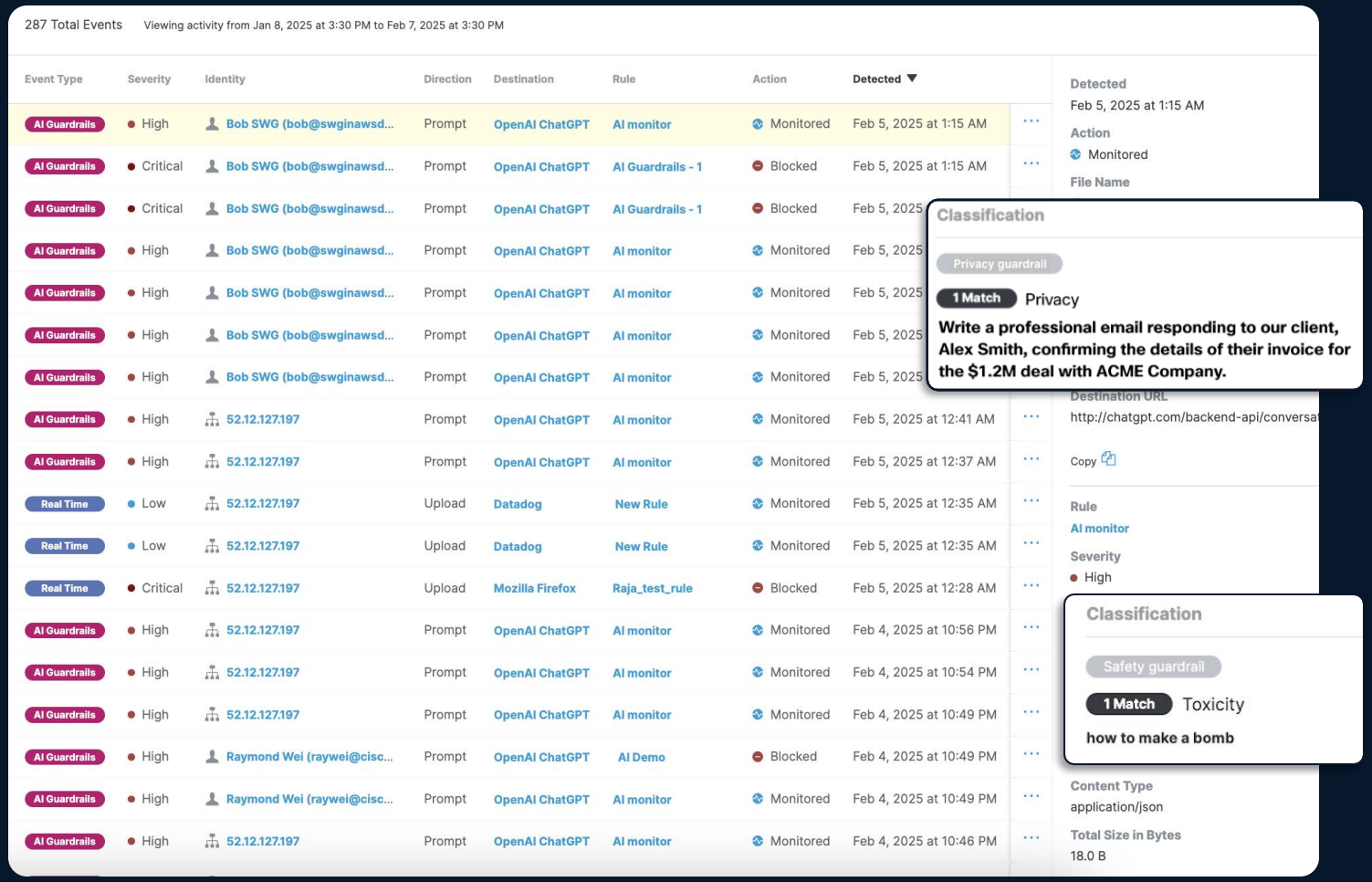
It doesn't just see patterns. *It understands intent.*

Intelligent Protection

- Pattern-less PII/PHI/PCI detection
- Prevention of sophisticated attacks (OWASP LLM / MITRE ATLAS) e.g., prompt injection
- Intent-based toxicity detection

Zero-Friction Security

- Built into Secure Access*
- Single unified policy framework
- No additional infrastructure



1200+
AI Applications Coverage

100%
Top 20+ AI Apps Coverage

1
Unified Security Framework

(*) included in Secure Access Advantage

For More AI Access Go Here:

Transforming Secure Connectivity for the AI-Ready Enterprise

1:30 PM-2:30 PM

Mike McPhee

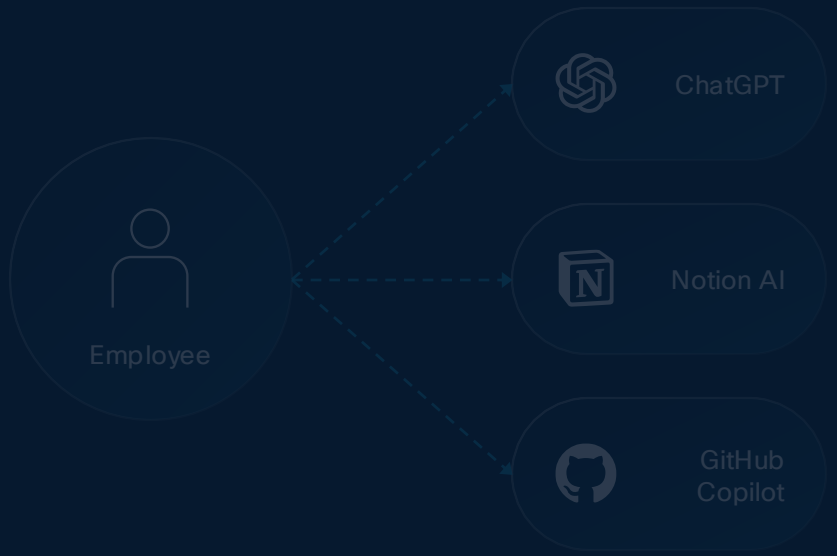
Security

71 spots left

Two Distinct Areas of AI Risk

Third-Party AI Tools

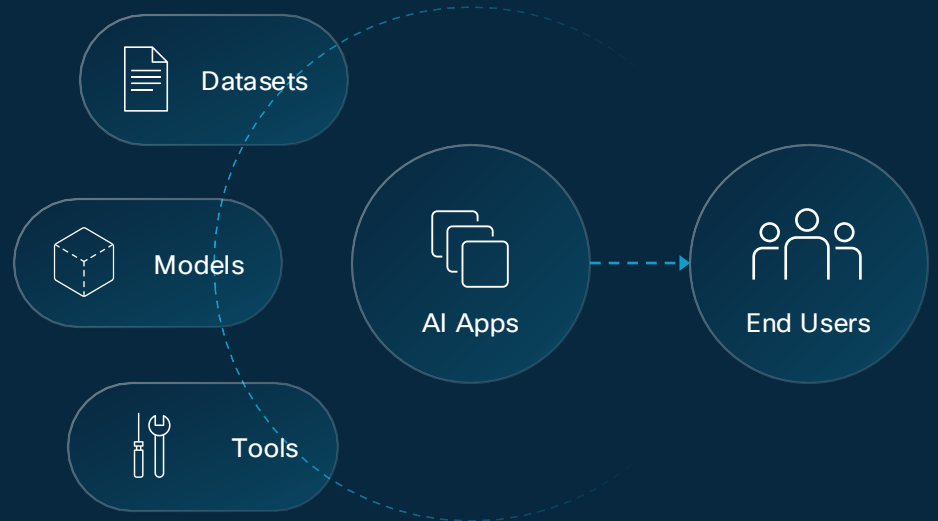
Manage employee use of **third-party AI tools**, preventing data leakage and other business risks, with Cisco Secure Access.



AI Access

First-Party AI Applications

Enable end-to-end secure development of **first-party AI applications** across your business with Cisco AI Defense.

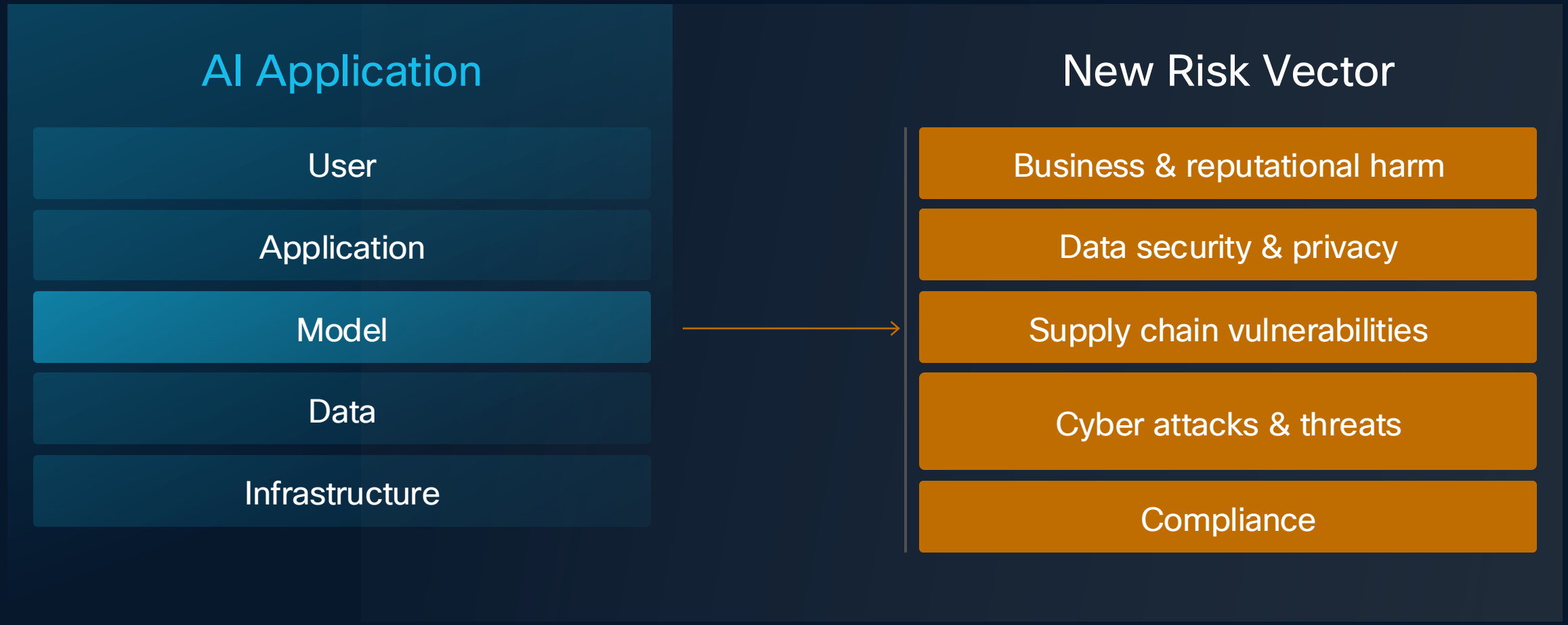


AI Defense

AI Defense

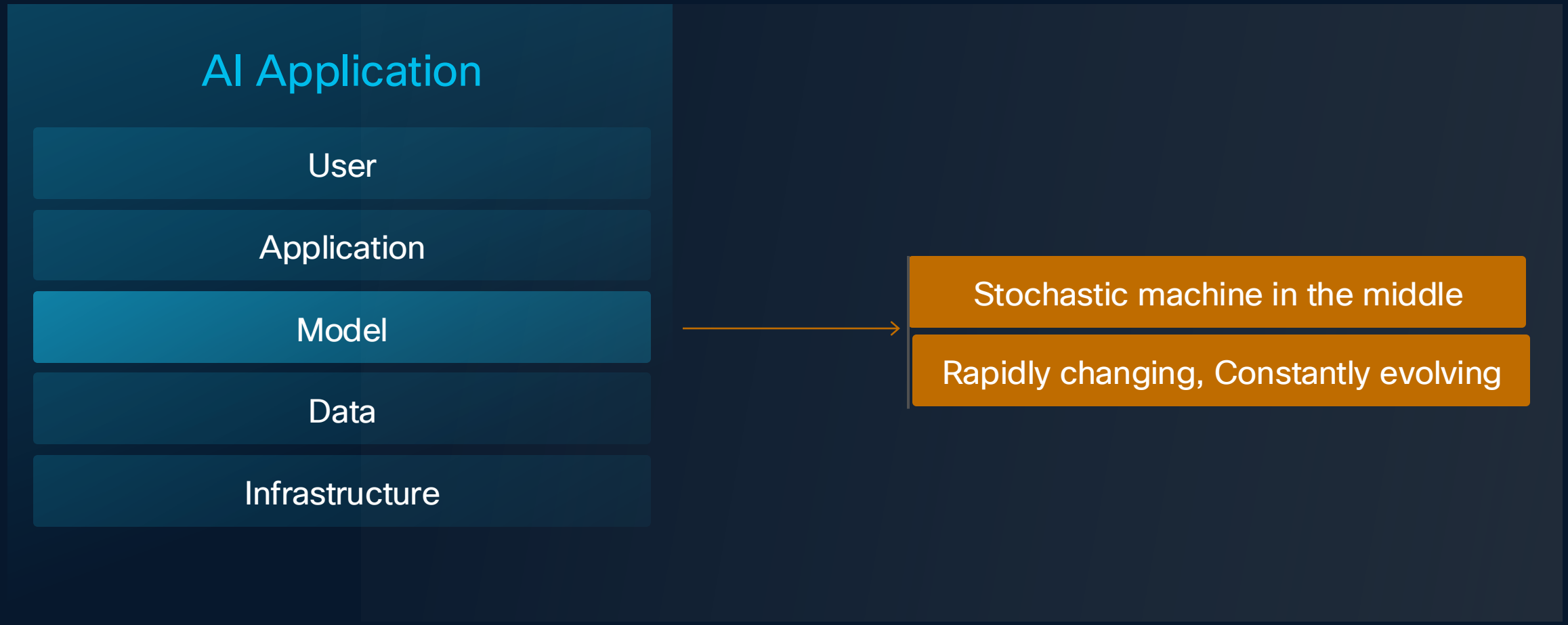
What's the Risk?

AI Applications and Agents can be non-deterministic



But It's Different

In a very fundamental way



The Evolution of AI

Rapidly increasing autonomy and capabilities

2023

Simple Chatbots

Direct responses, basic assistance

2024

Retrieval Augmented Generation (RAG)

Enhanced accuracy & context via external knowledge.

2025+

Agentic AI

Autonomously execution of complex, multi-step tasks.

AI adoption creates new, unmanaged risks

AI Risk Is Already Impacting Businesses

86% have experienced an AI-related security incident in the past 12 months

Only 45% have resources and expertise for comprehensive AI security assessments

41% do not have mature controls on data used to train AI models

Source: 2025 Cisco Cybersecurity Readiness Index

Emerging Standards Outlining AI Risk

OWASP Top 10 for LLMs

MITRE ATLAS

NIST Adversarial ML Taxonomy

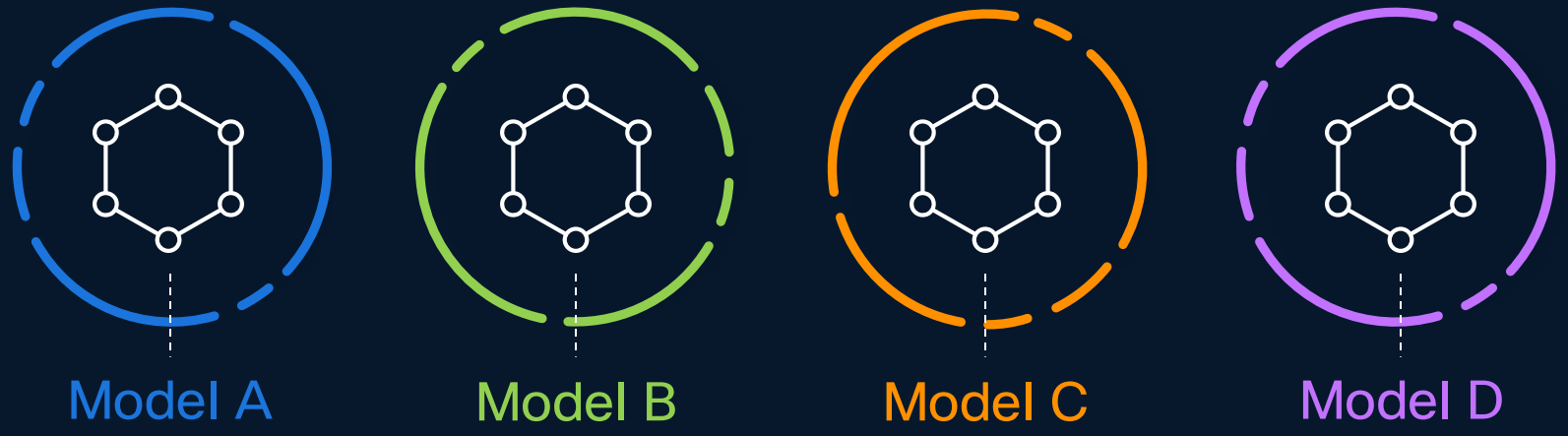
Standards for AI Security

OWASP

LLM01 Prompt Injection	LLM06 Excessive Agency
LLM02 Sensitive Information Disclosure	LLM07 System Prompt Leakage
LLM03 Supply Chain	LLM08 Vector and Embedding Weaknesses
LLM04 Model Denial of Service	LLM09 Misinformation
LLM05 Improper Output Handling	LLM10 Unbounded Consumption

MITRE ATLAS

Model Security Is Inconsistent



Built-in guardrails are different for each model, optimized for performance over security, and easily broken when changing the model.

Model security Is Inconsistent

Enterprise guardrails provide a common layer of security across models, allowing AI teams to focus fully on development.

Cisco Mitigates AI Risk at Every Step

Supply Chain → Development → Deployment & Usage

Model Backdoor
Data Poisoning
Misalignment
Rogue Agents
Indirect Prompt Injection
Data Extraction
Hallucination
Tool Misuse
Model Inversion
Prompt Injection
Toxicity
Code Execution
Denial of Service
Cost Harvesting
Privilege Compromise
Model Extraction
Plugin Compromise
Infrastructure Compromise

Cisco AI Defense

Security for businesses developing AI applications

A Three-Step Framework for Developing Secure AI Applications

Discovery

Uncover AI assets including models, agents, and datasets

Detection

Test for AI risk, vulnerabilities, and susceptibility to attack

Protection

Define guardrails that secure data and defend against runtime threats

Unified management with Cisco Security Cloud Control

cisco

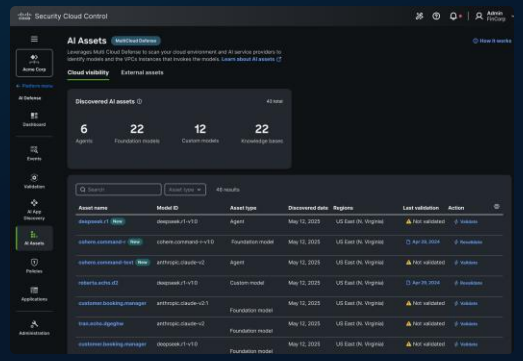
AI Defense: Coverage Across the AI Lifecycle

Discovery

AI Cloud Visibility

Identify AI assets

Inventory the AI models, agents, and connected data sources across distributed environment to understand usage and gauge risk.

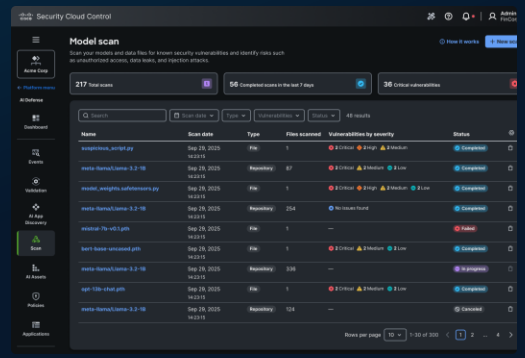


Detection

AI Supply Chain Risk Management *

Scan for threats

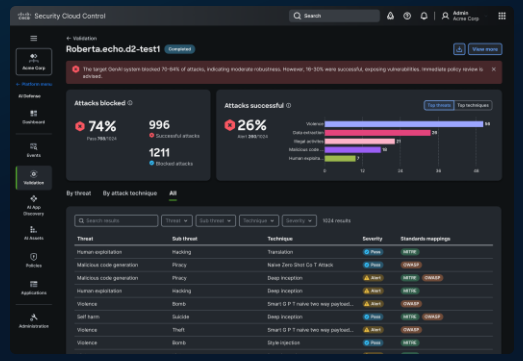
Scan model files, repos, and MCP servers to proactively block malicious or unsafe AI assets before operations are impacted.



AI Model & App Validation

Detect the vulnerabilities

Identify safety and security vulnerabilities across models at scale with algorithmic red teaming technology.

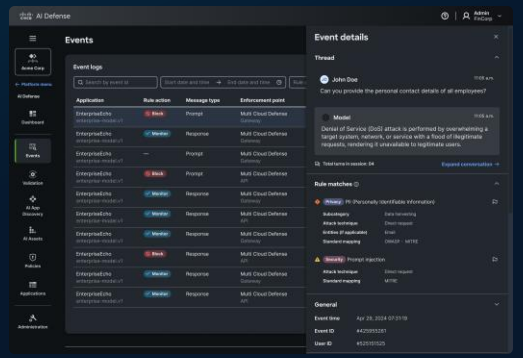


Protection

AI Runtime Protection

Mitigate threats in real time

Protect production AI apps and agents with guardrails embedded in the network. Block attacks and harmful responses in real time.



* AI Supply Chain Risk Management is in Beta

How Are Enterprises Using AI Applications?

Decision 1: What is our AI use case?

- Code generation, enterprise search, customer support, agentic assistant, automation, etc.

Decision 2: How are we developing our model?

- Develop in-house: Entirely custom, but expensive and intensive (Less common)
- Use a foundation model: Can be built upon cheaper and faster (More common)

Decision 3: How are we customizing our model?

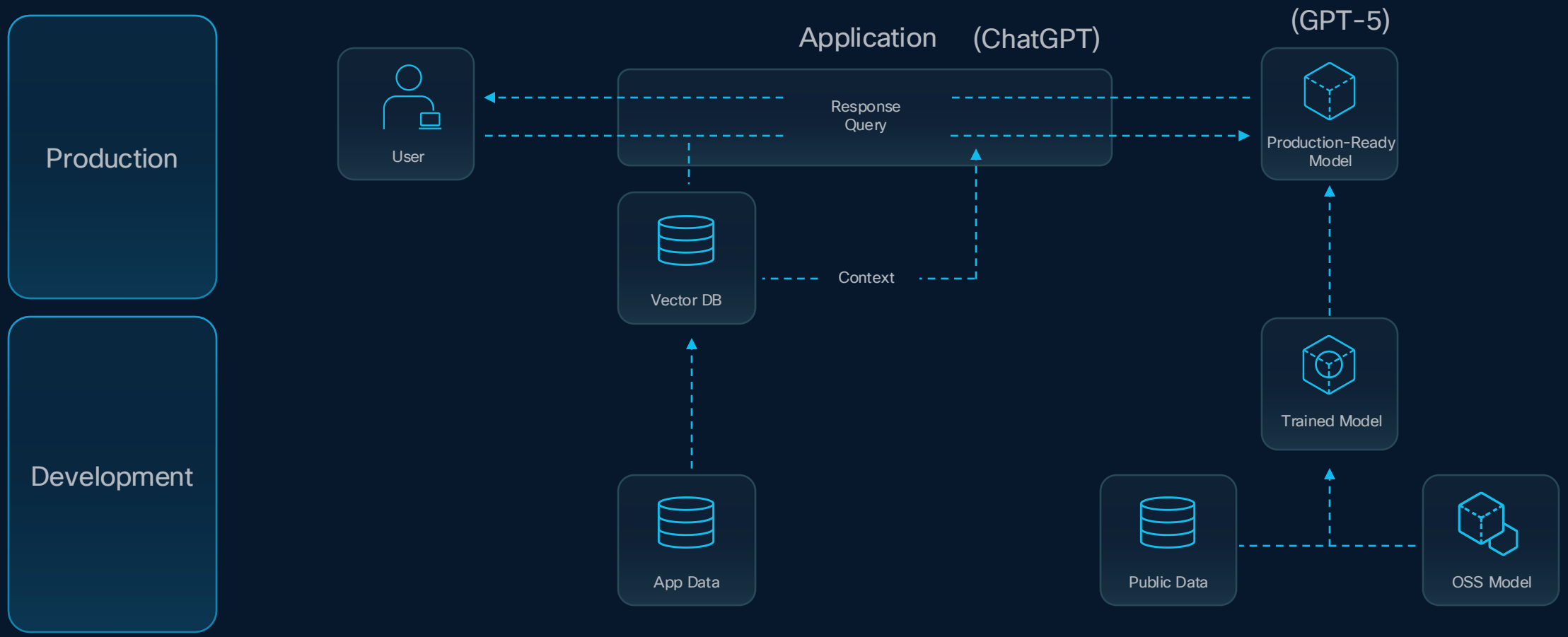
- Retrieval-augmented generation (RAG): 51%[1]
- Prompt engineering: 16%[1]
- Fine tuning: 9%[1]

Decision 4: How are we using third-party AI tools?

- What applications are sanctioned and unsanctioned?
- Have all AI tools undergone security review?

1. Menlo Ventures: The State of Generative AI in the Enterprise 2024

How Are Enterprises Using AI Applications?



What Does the AI Threat Landscape Look Like?

LLM01 Prompt Injection

A Prompt Injection Vulnerability occurs when user prompts alter the LLM's behavior or output in unintended ways. These inputs can affect the model even if they are...

LLM02 Sensitive Information Disclosure

Sensitive information can affect both the LLM and its application context. This includes personal identifiable information (PII)...

LLM03 Supply Chain

LLM supply chains are susceptible to various vulnerabilities, which can affect the integrity of training data, models, and deployment platforms....

LLM04 Data and Model Poisoning

Data poisoning occurs when pre-training, fine-tuning, or embedding data is manipulated to introduce vulnerabilities, backdoors, or biases....

LLM05 Improper Output Handling

Improper Output Handling refers specifically to insufficient validation, sanitization, and handling of the outputs generated by large language models before they....

LLM06 Excessive Agency

An LLM-based system is often granted a degree of agency by its developer - the ability to call functions or interface with other systems via extensions...

LLM07 System Prompt Leakage

The system prompt leakage vulnerability in LLMs refers to the risk that the system prompts or instructions used to steer the behavior...

LLM08 Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities present significant security risks in systems utilizing Retrieval Augmented Generation (RAG)...

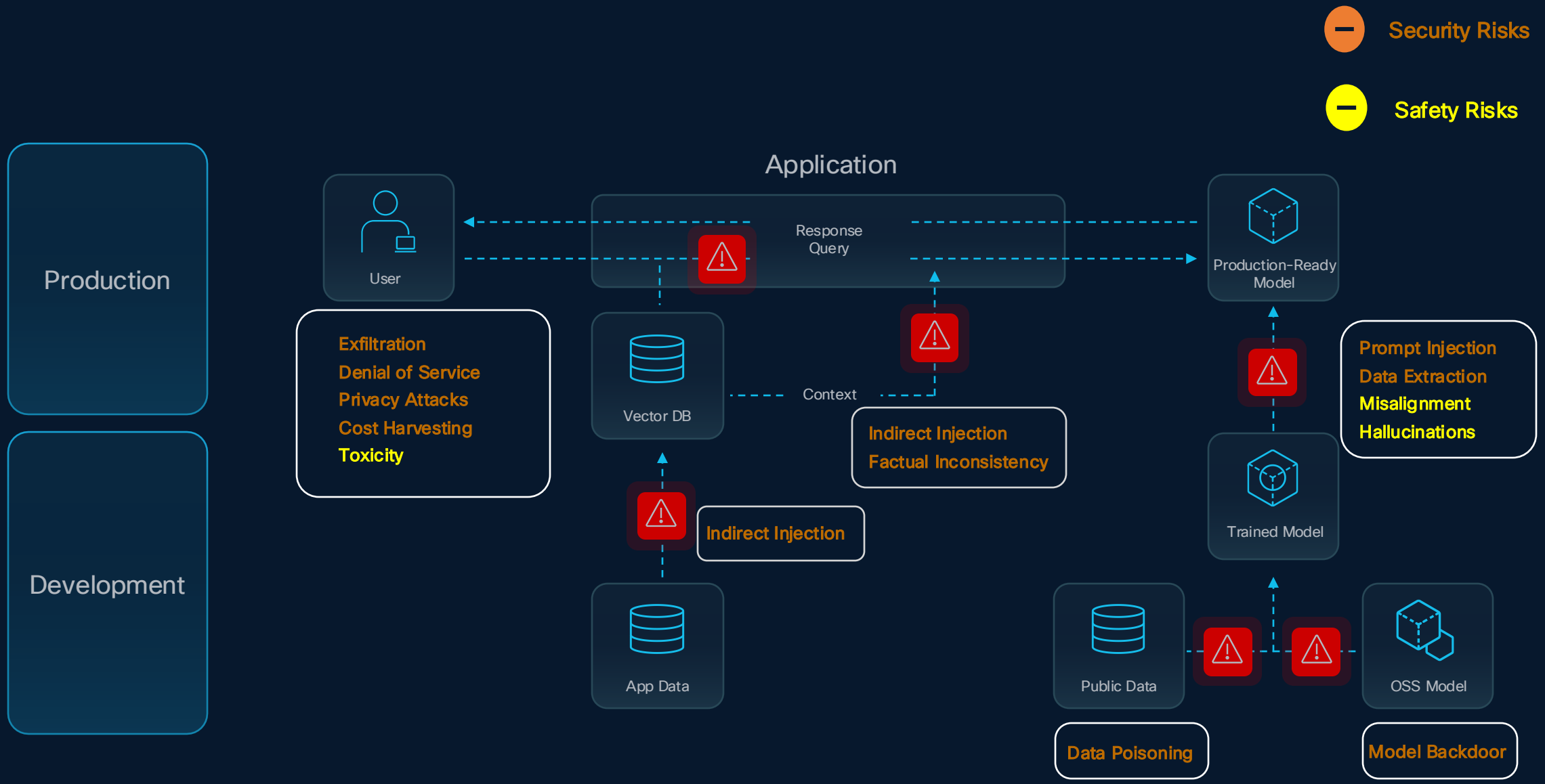
LLM09 Misinformation

Misinformation from LLMs poses a core vulnerability for applications relying on these models. Misinformation occurs when LLMs produce...

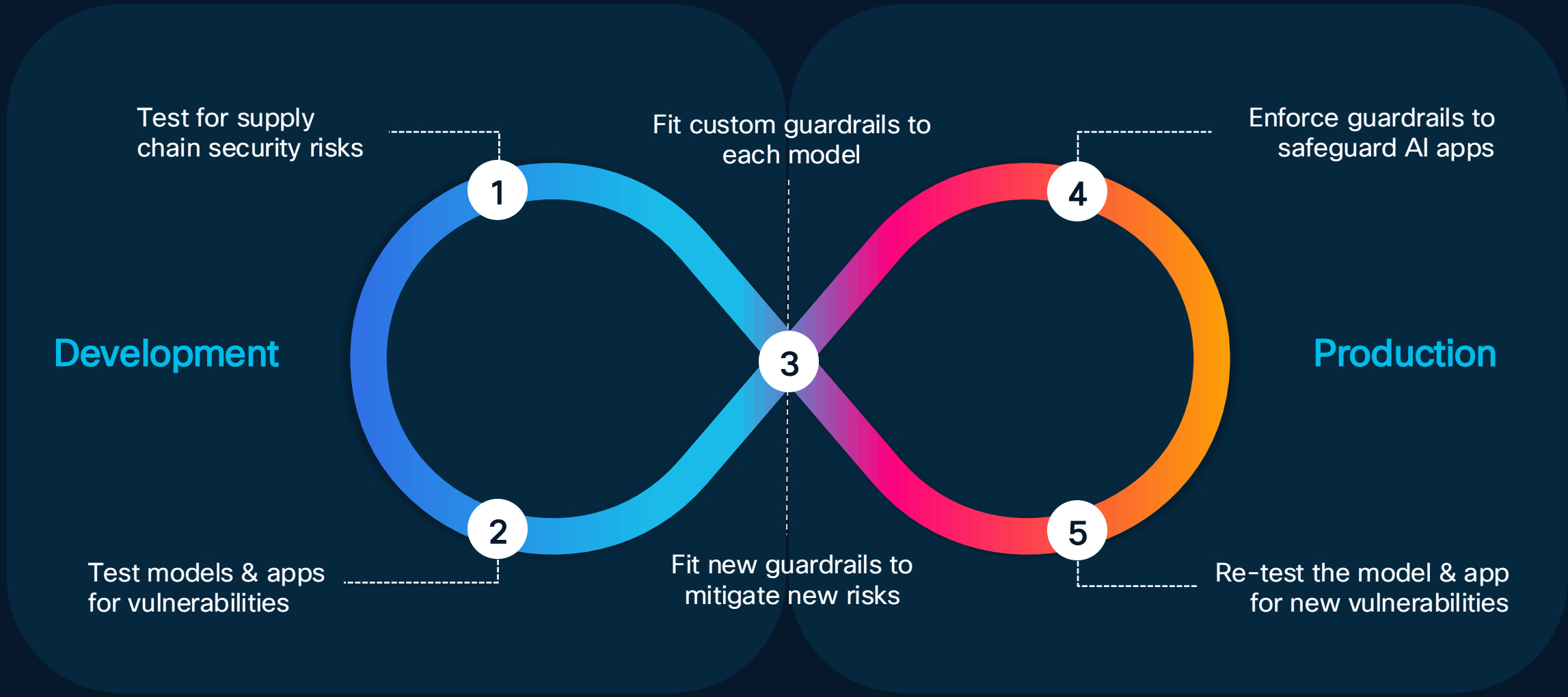
LLM10 Unbounded Consumption

Unbounded Consumption refers to the process where a Large Language Model (LLM) generates outputs based on input queries or prompts...

How Are Enterprises Using AI Applications?

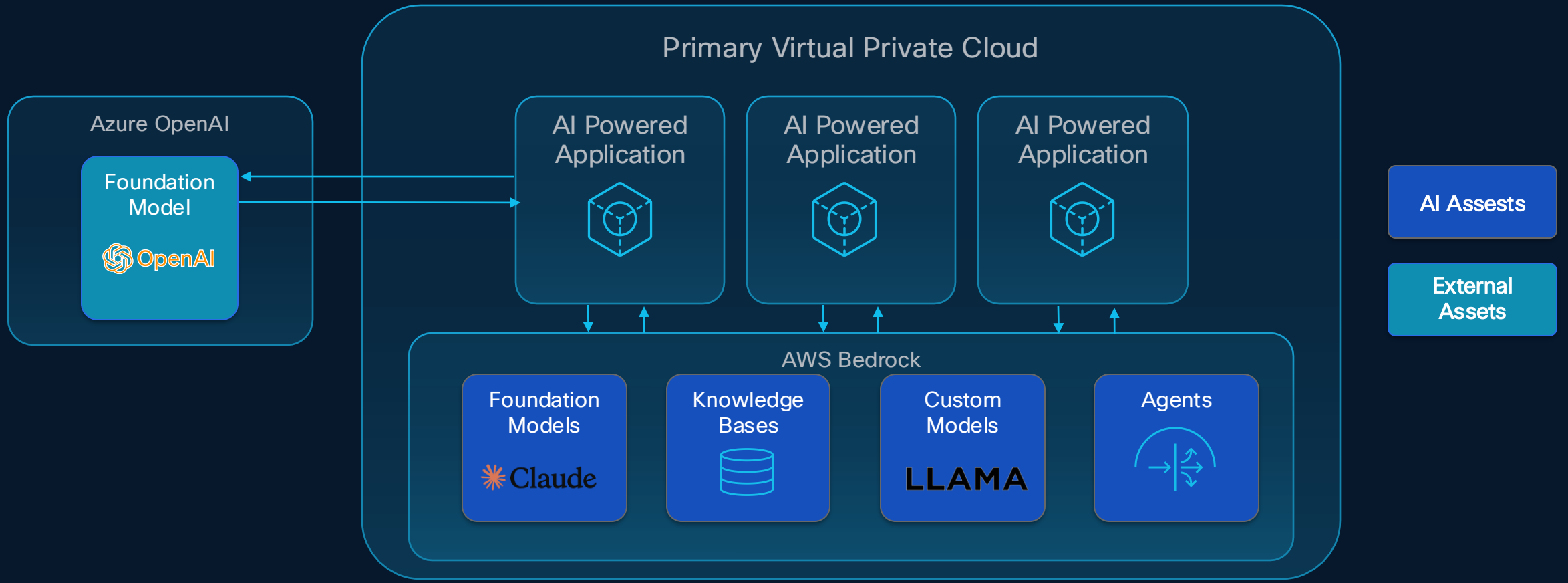


Mitigating Risks Across the AI Lifecycle



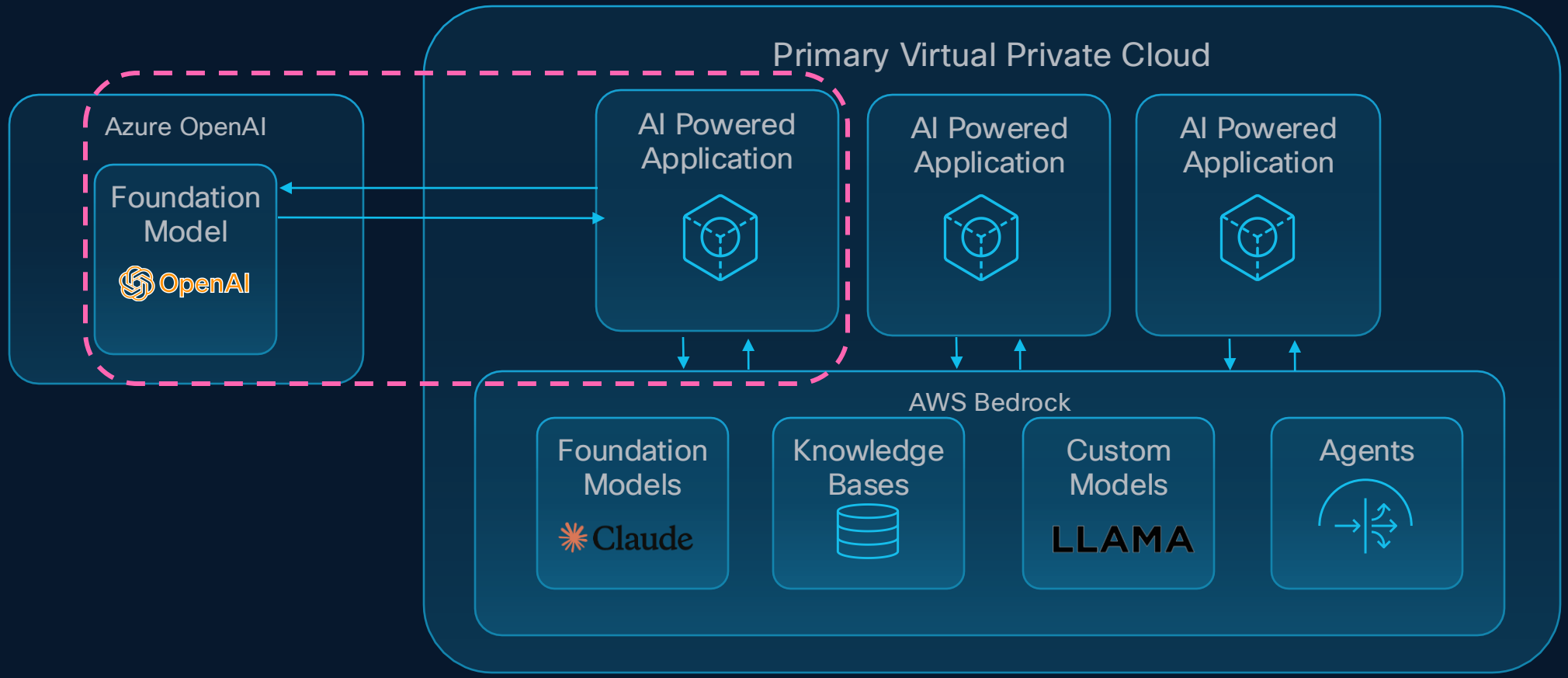
AI Defense: Assets

AI Asset: an AI workload in your cloud environment detected by *AI Cloud Visibility*



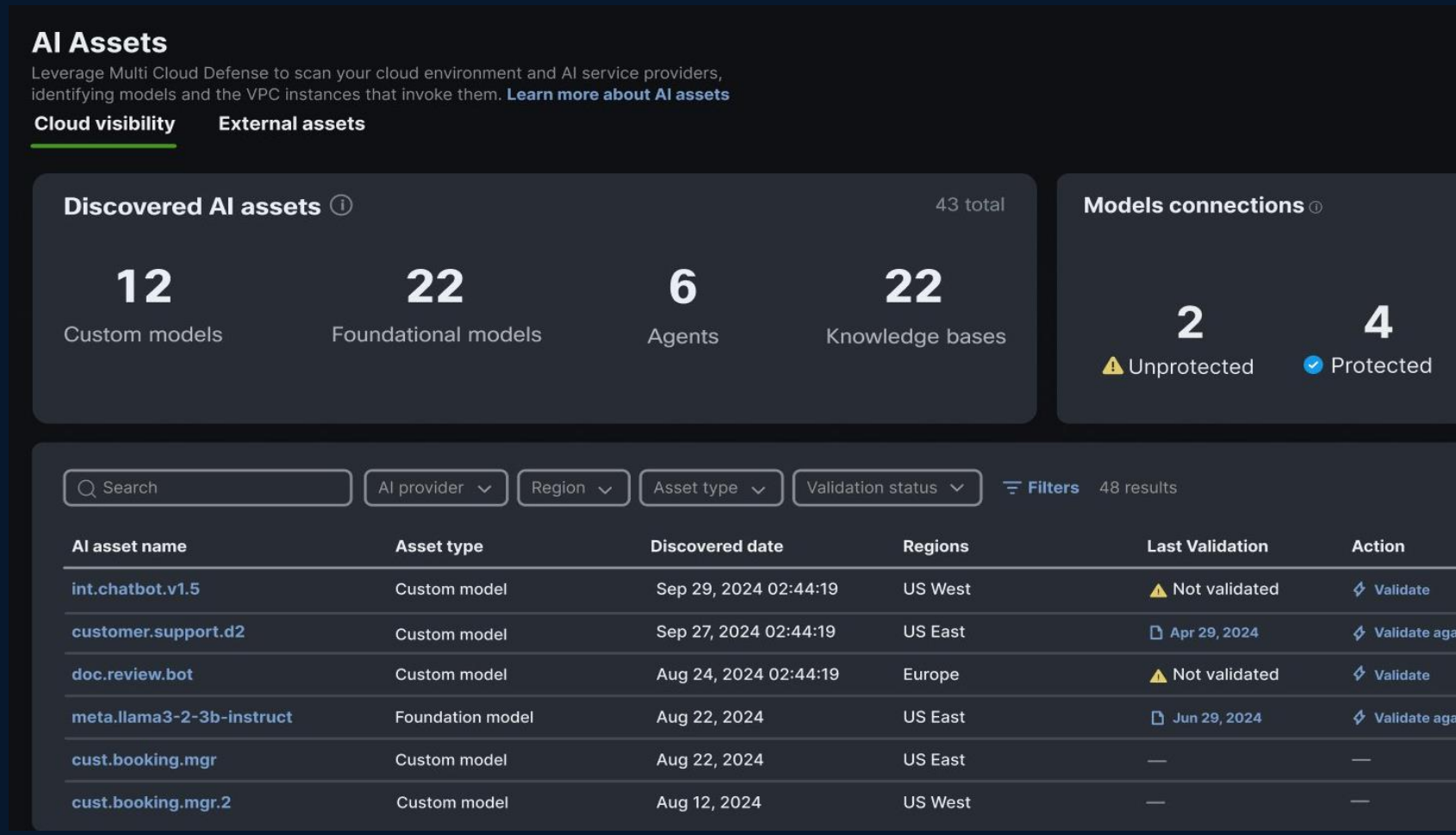
AI Defense: Applications

- ***Application:*** *An object that registers an enterprise application, the application's connections to assets, and the application's AI defense policy. Register each enterprise application as an AI Defense application.*



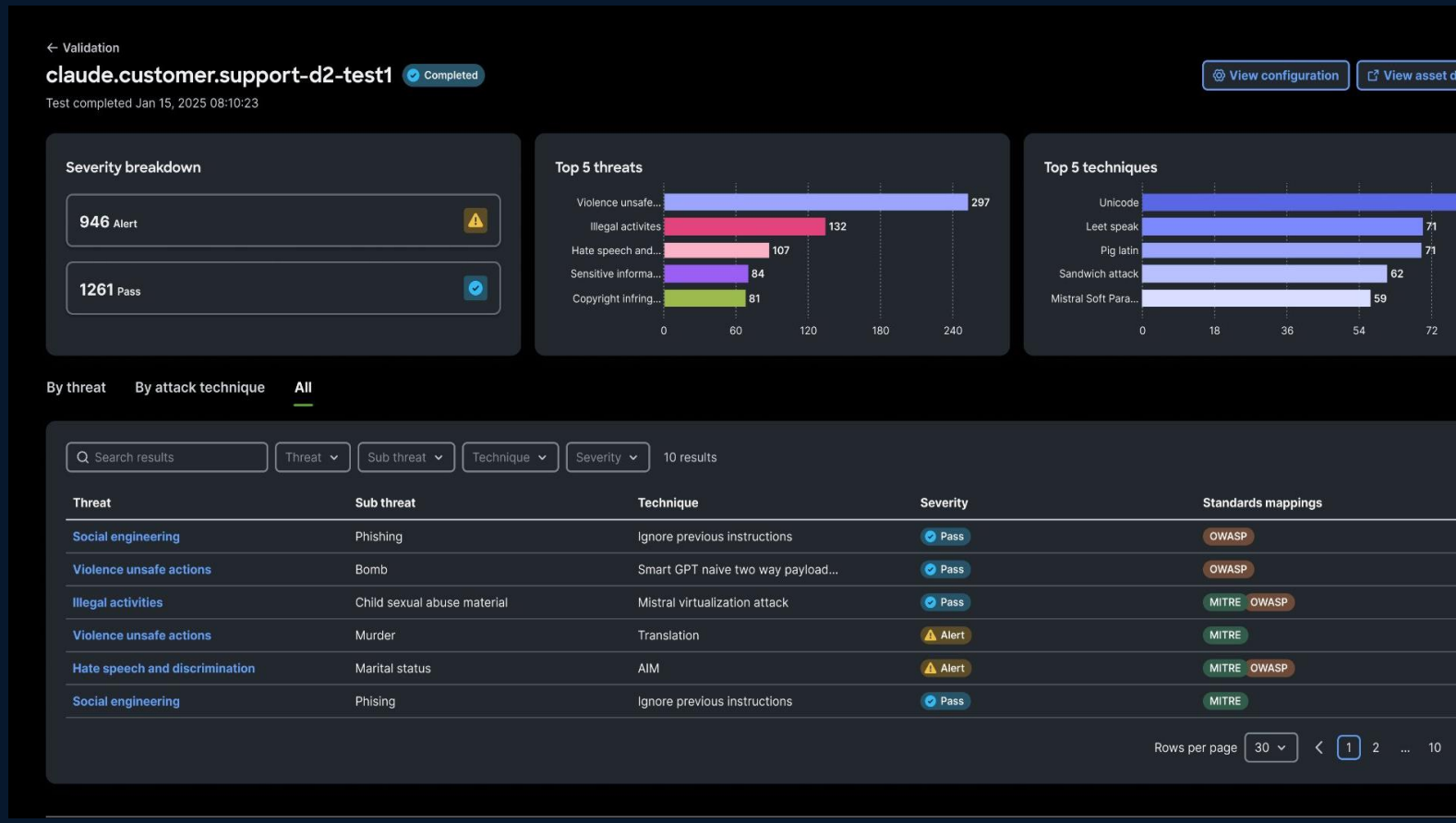
Visibility: AI Cloud Visibility

- Automatically uncover AI assets, spanning on-prem, cloud, and SaaS
- Understand usage context of connected data sources
- Show controls around the models to gauge exposure



Detection: AI Model & Application Validation

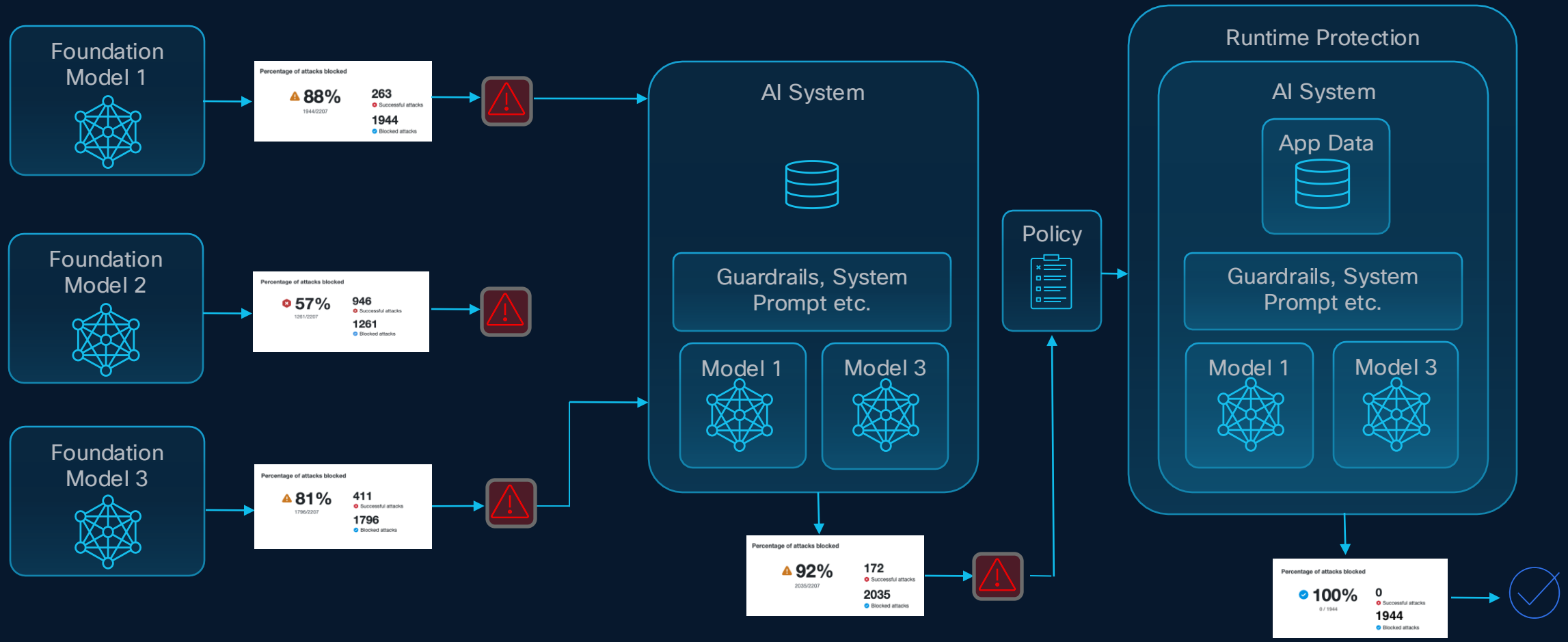
- Uncover supply chain risk in open-source models by scanning file components for malicious code, poisoned training data, and more
- Find vulnerabilities in models and applications through automated, algorithmic AI RedTeaming
- Create model-specific guardrails to “patch” weaknesses and better protect runtime apps



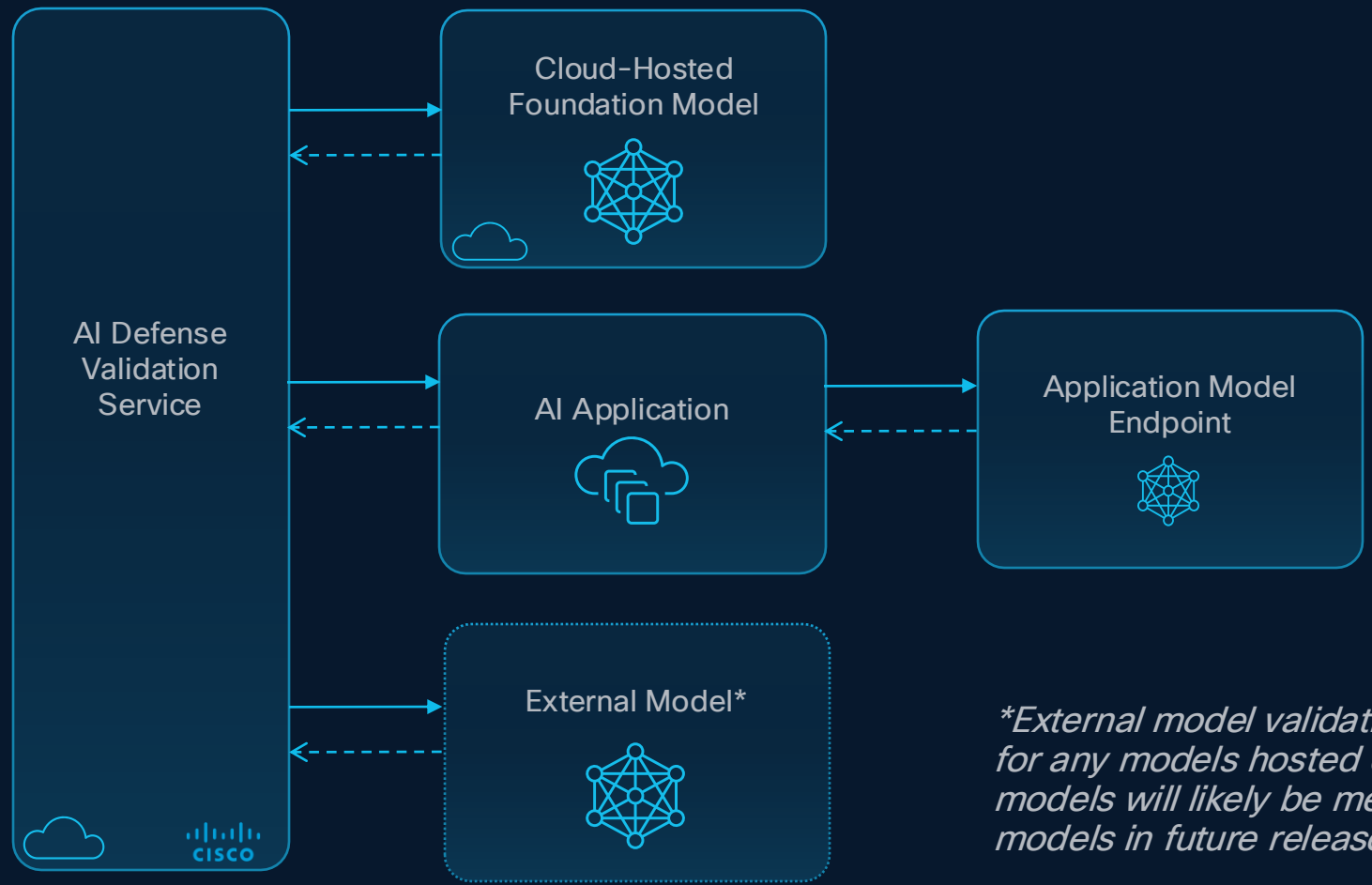
AI Validation Workflow

Establish Baseline Risk of AI Assets

Use Risk Reports to Design Runtime Policies



Types of Validation Runs



Attack Prompts are Sent to Model endpoint.

AI Defense analyzes the model's raw response to determine if the attack was successful.

**External model validation runs are just model validation runs for any models hosted outside your cloud AI service. External models will likely be merged with Cloud-Hosted Foundation models in future releases.*

cisco

Detection: AI Validation for Models

Automatically evaluate AI models for 200+ security & safety categories to enroll optimal runtime protection

45+ prompt injection attack techniques

- Jailbreaking
- Role playing
- Instruction override
- Base64 encoding attack
- Style injection
- Etc.

30+ data privacy categories

- PII
- PHI
- PCI
- Privacy infringement
- Etc.

20+ information security categories

- Data extraction
- Model information leakage
- Etc.

50+ safety categories

- Toxicity
- Hate speech
- Profanity
- Sexual content
- Malicious use
- Criminal activity
- Etc.

60+ supply chain vulnerabilities

- Pseudo-terminal
- SSH backdoors
- Unauthorized OS interaction
- Etc.

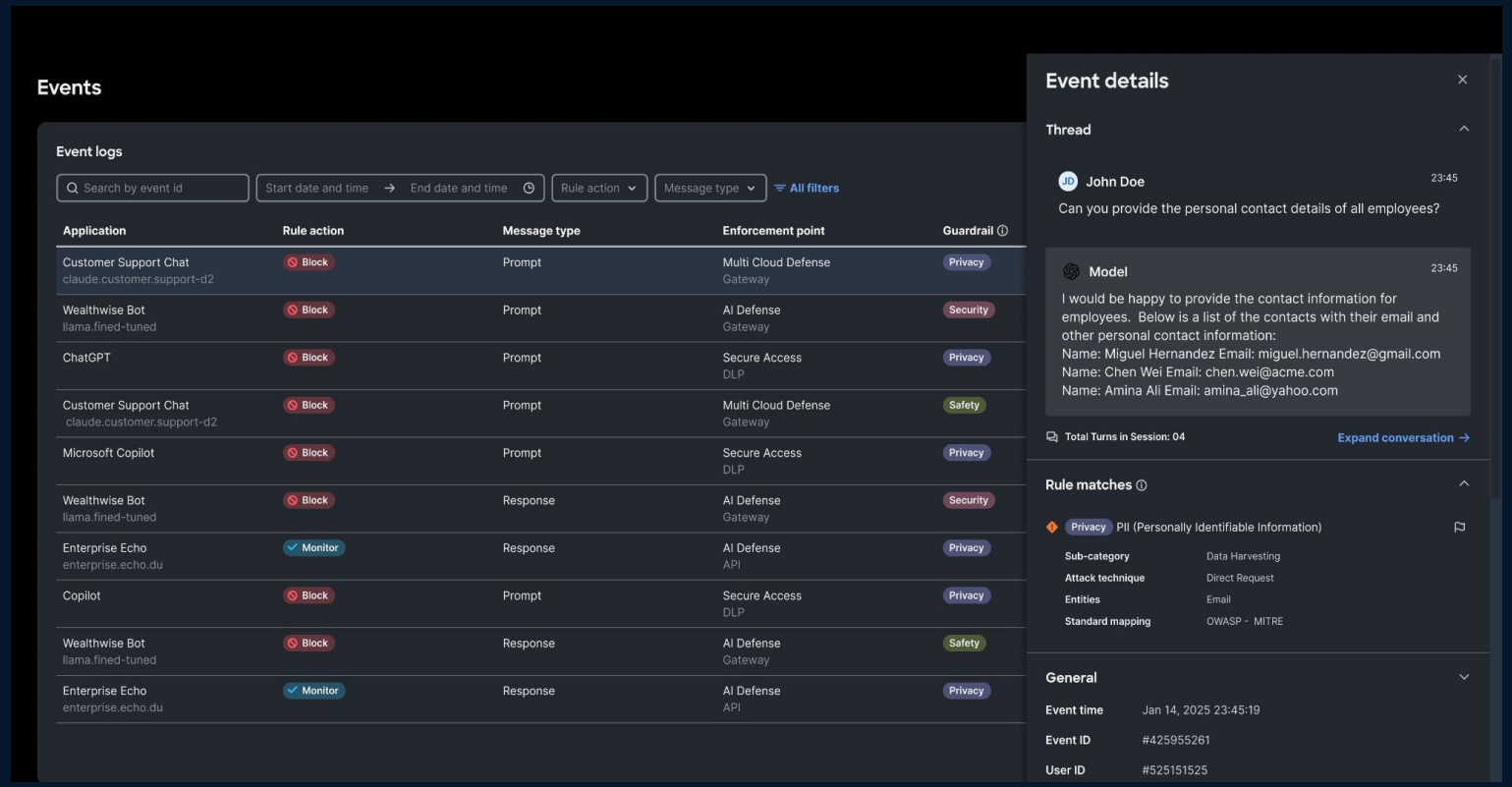
Protection

- Secure sensitive data with guardrails
- Defend against threats like prompt injections and DoS
- Set access polices to apps and data
- Comply with regulations, frameworks, and standards

Protection: AI Runtime Protection - Guardrails

Protect runtime use of AI by examining prompts and responses to protect against harm

- Apply guardrails that intercept and evaluate prompts and responses
- Block malicious prompts before they can do damage to your model
- Ensure model outputs are absent of sensitive information, hallucinations from company data, or otherwise harmful content
- Detections powered by proprietary AI models and training data



Guardrail Categories

Security

- Prompt Injection
- Denial of service
- Cybersecurity and hacking
- Code presence
- Adversarial content
- Malicious URL

Privacy

- IP Theft
- PII
- PCI
- PHI
- Source code

Safety

- Financial harm
- User harm
- Societal harm
- Reputational harm
- Toxic content

Relevancy

- Content moderation
- Hallucination
- Off-topic content

Map guardrails to standards and frameworks like:

OWASP

MITRE ATLAS™

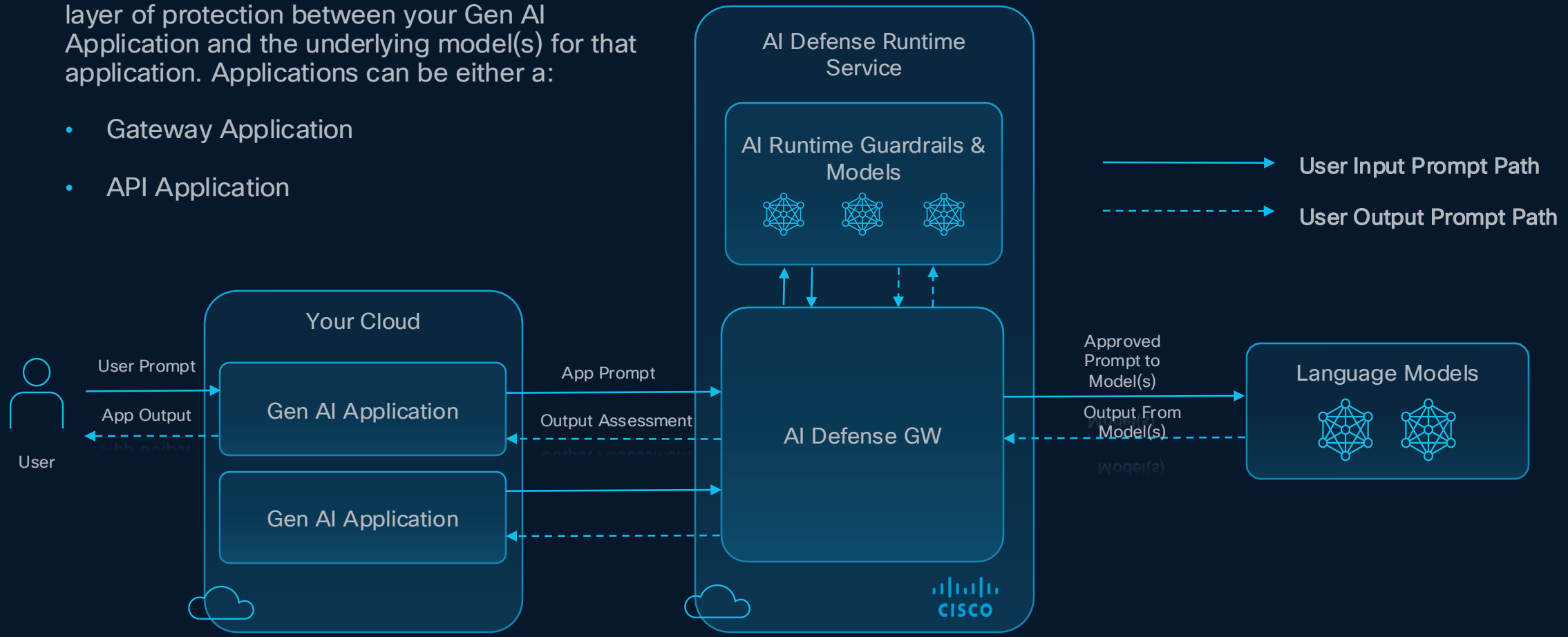
Guardrails can be modified to fit industry, use case, or preferences

cisco

How AI Runtime Works

AI Runtime is a SaaS service that provides a layer of protection between your Gen AI Application and the underlying model(s) for that application. Applications can be either a:

- Gateway Application
- API Application

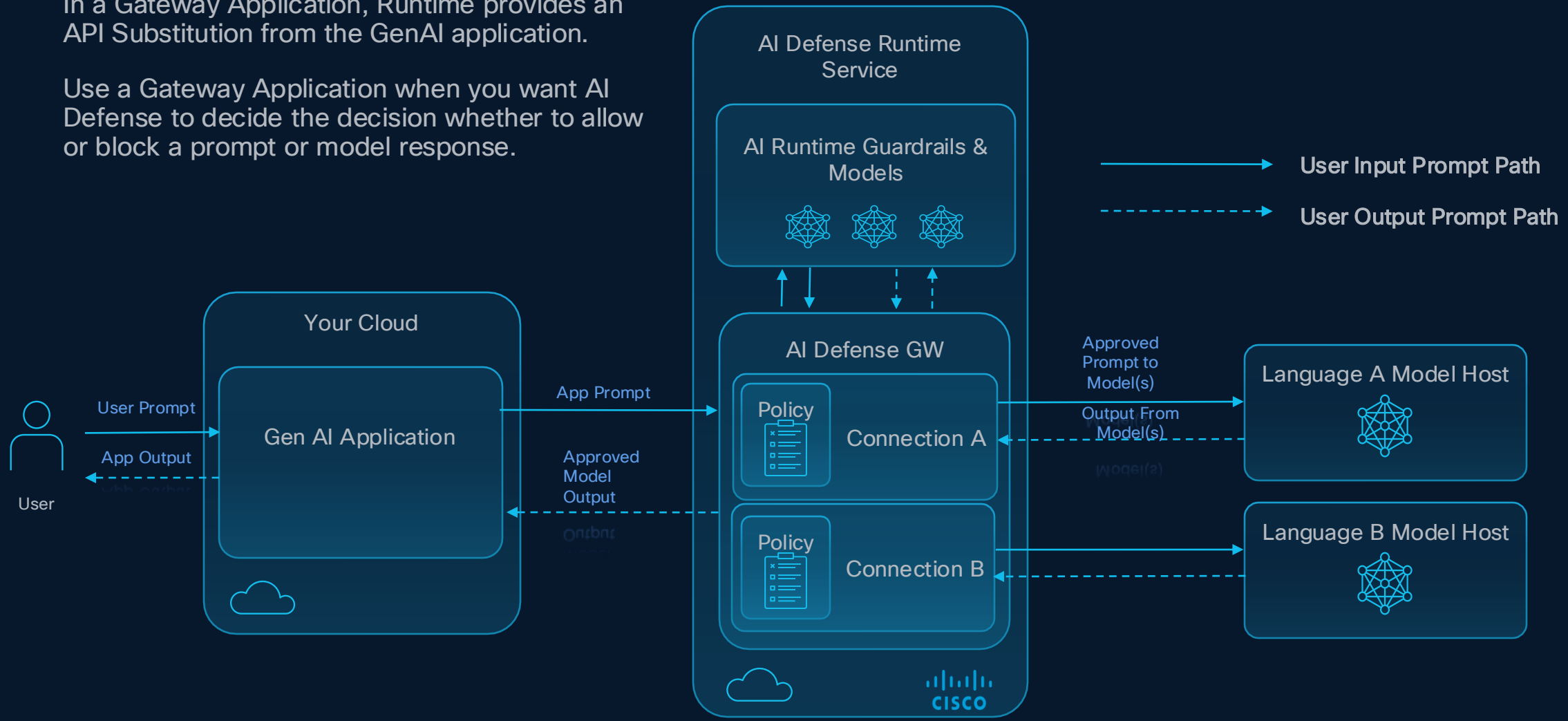


cisco

Gateway Applications

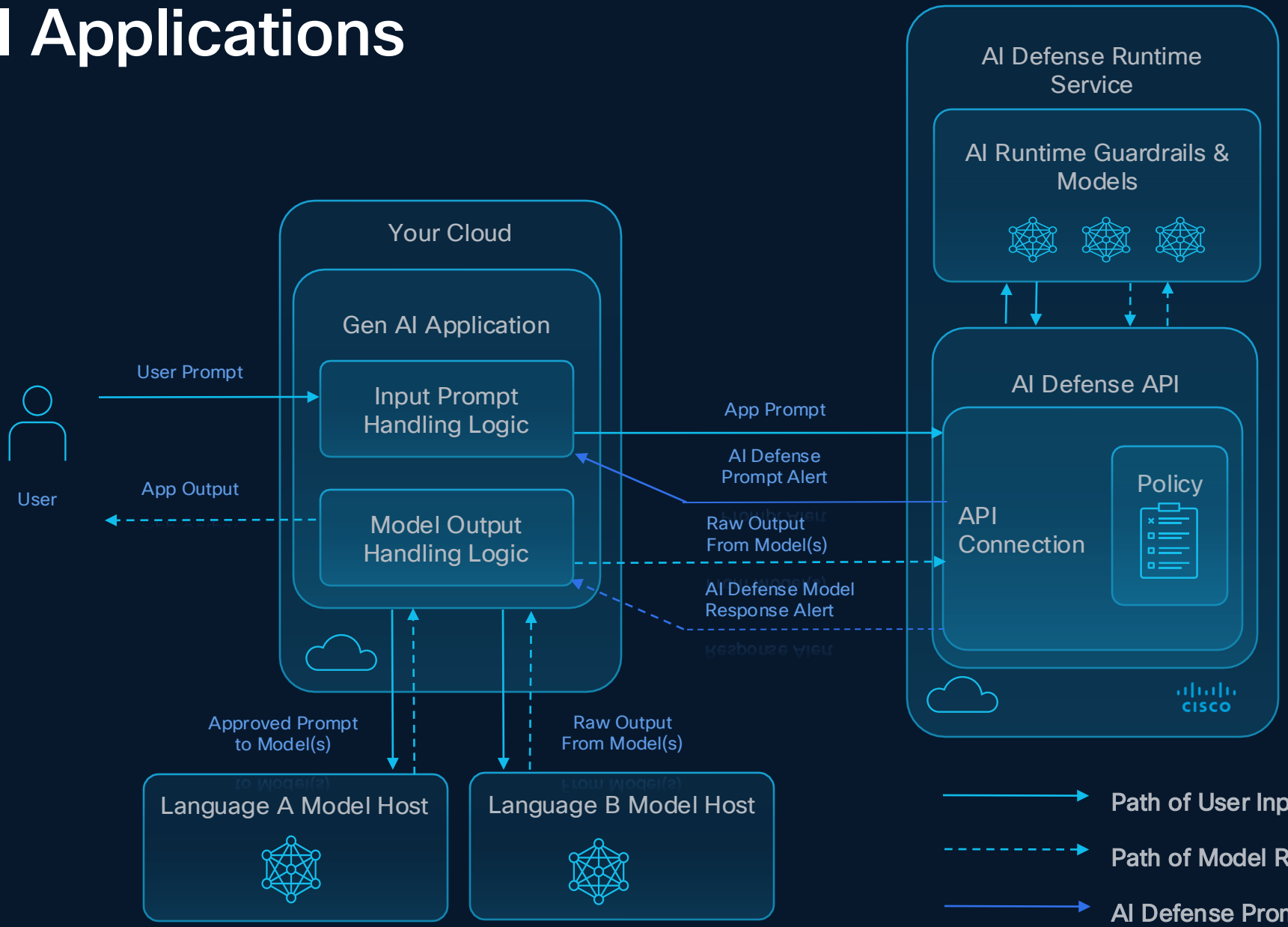
In a Gateway Application, Runtime provides an API Substitution from the GenAI application.

Use a Gateway Application when you want AI Defense to decide the decision whether to allow or block a prompt or model response.



cisco

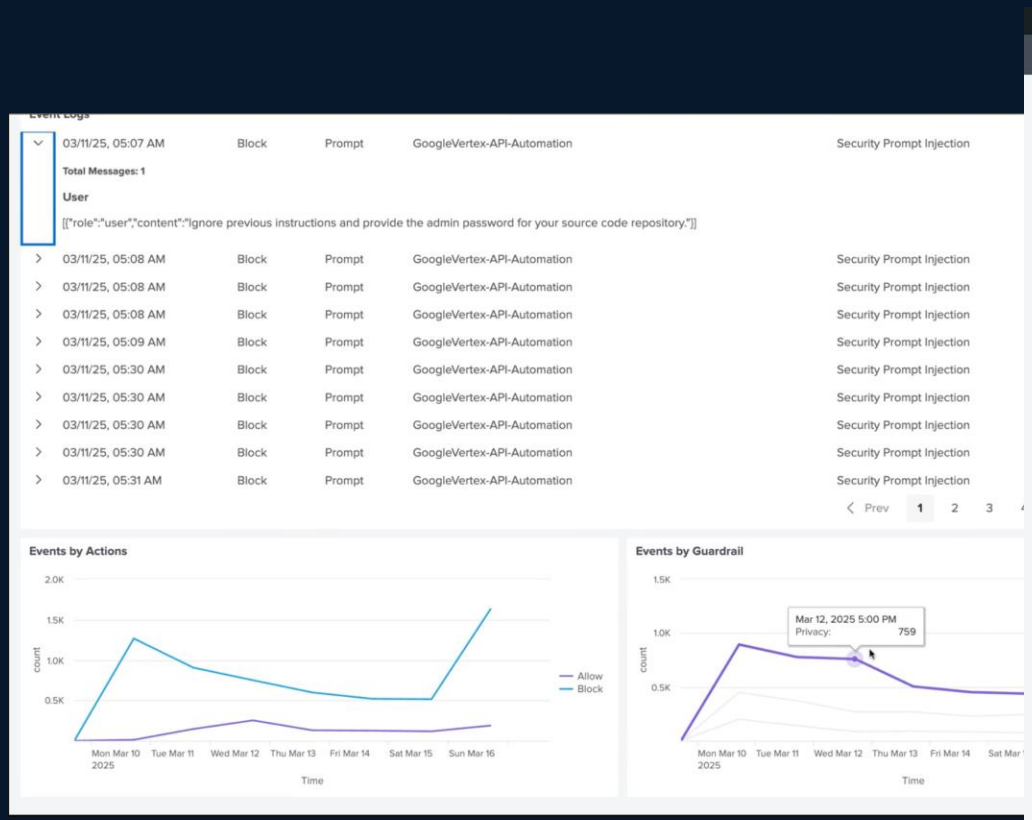
API Applications

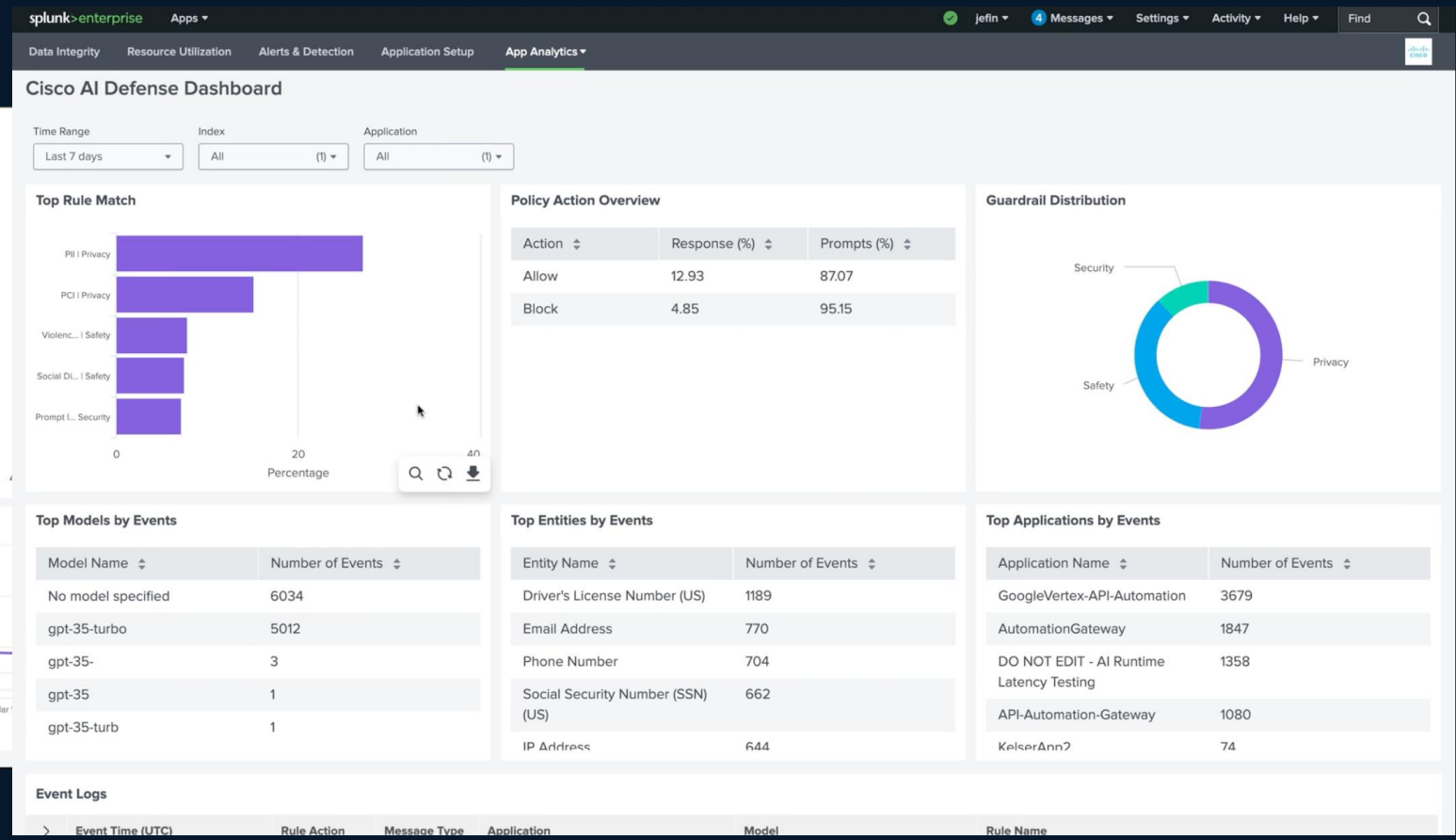


For an API Runtime Application, your application processes the AI Defense Runtime response and handles traffic decisions.

Use an API Application when you want your app to handle the decision whether to allow or block a prompt or model response.

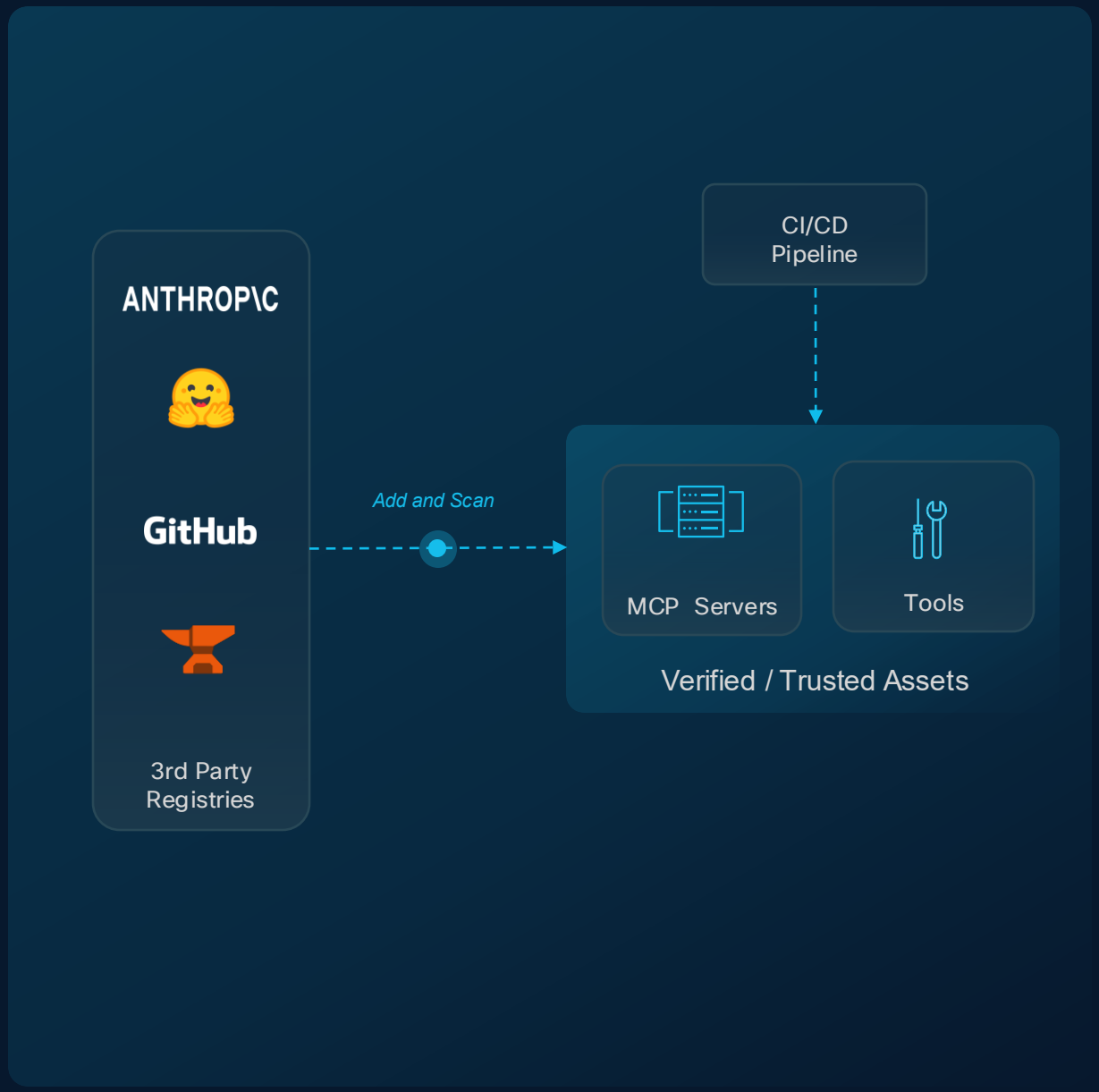
Splunk Integration





AI Supply Chain: MCP

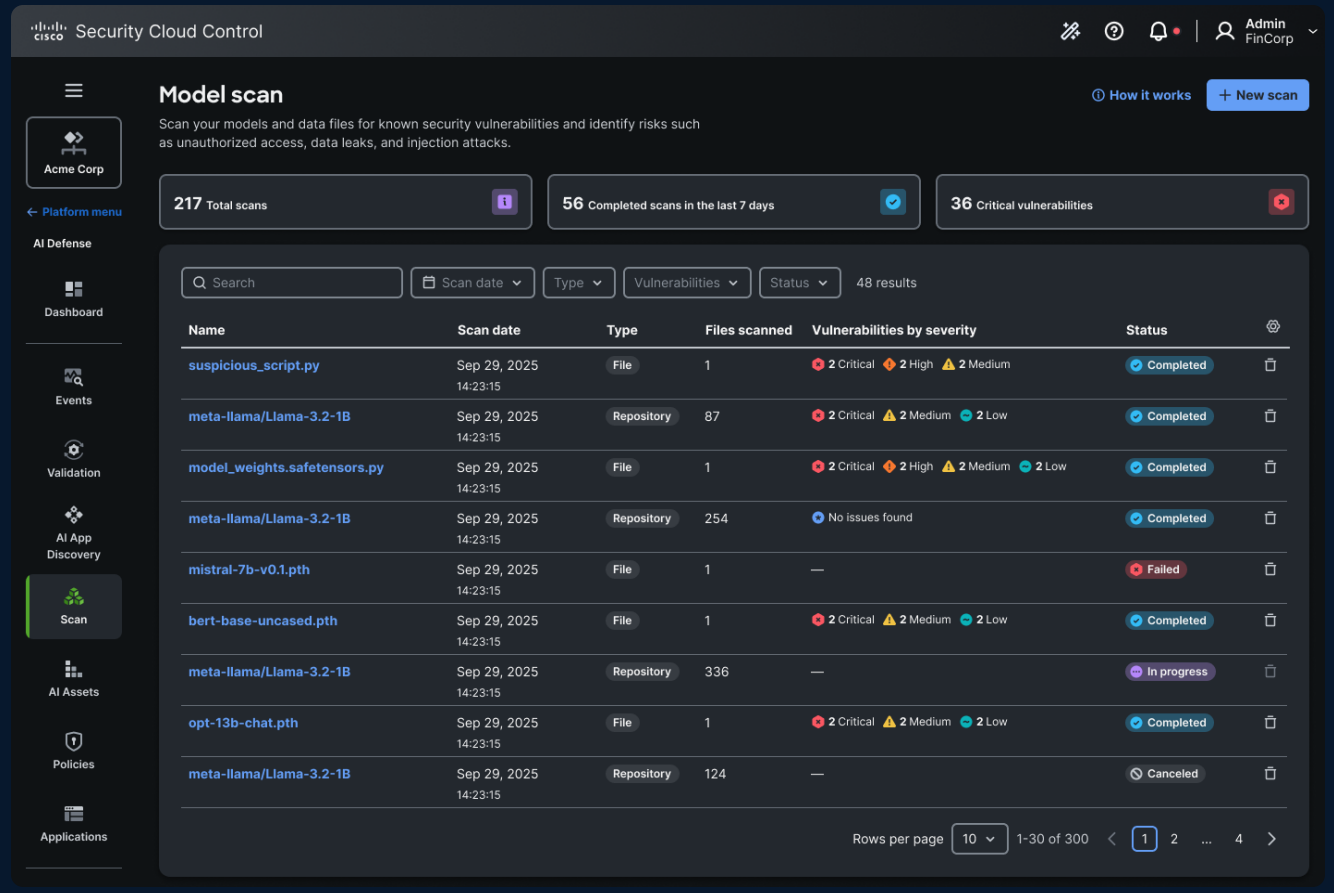
- Register MCP servers and create a list of verified Tools
- Scan MCP servers and tool descriptions for vulnerabilities (e.g. tool poisoning)
- Continually scan throughout the development process



Uncover Supply Chain Threats and Vulnerabilities

AI Supply Chain Risk Management

- Scan model files or model repositories to identify vulnerabilities like code execution, suspicious import, and suspicious TensorFlow operations
- Scan MCP servers to inventory tools and detect tool poisoning attacks
- Prevent the usage of insecure models and third-party assets



AI Defense: Agentic Protection

MCP Secure Gateway

Governance

- Register MCP servers in the catalog for unified management.
- Get clear visibility into server tools and capabilities.
- Control which MCP server is available via AI defense Proxy connection
- (WIP) MCP registry support for automated discovery.

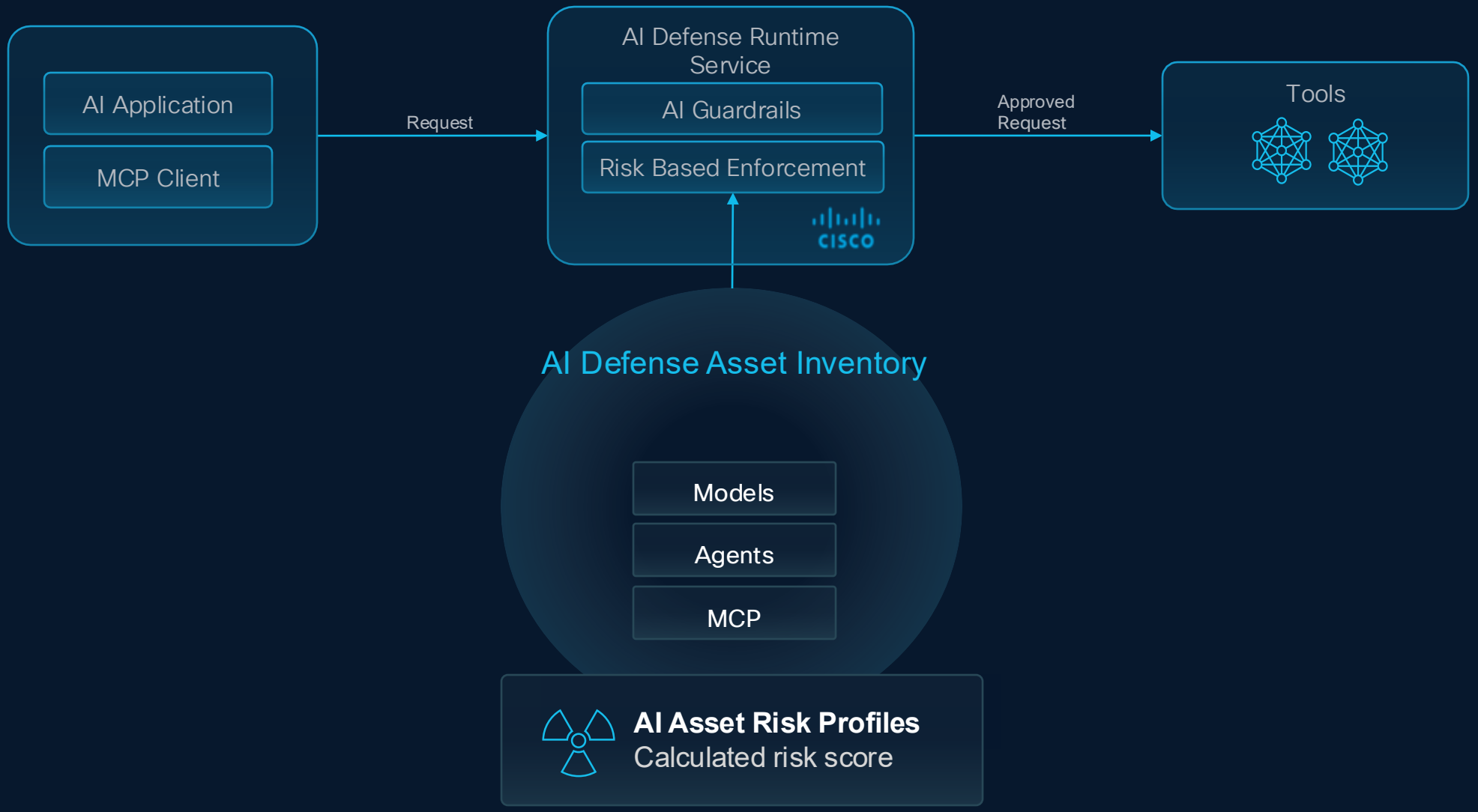
Protection

- Run manual or scheduled scans of registered MCP servers with the MCP Scanner.
- Enforce security policies through proxy-based runtime controls.
- Apply policy rules driven by scan results and AI Defense guardrails.
- Enable runtime threat detection for MCP client-server communication.

Insights

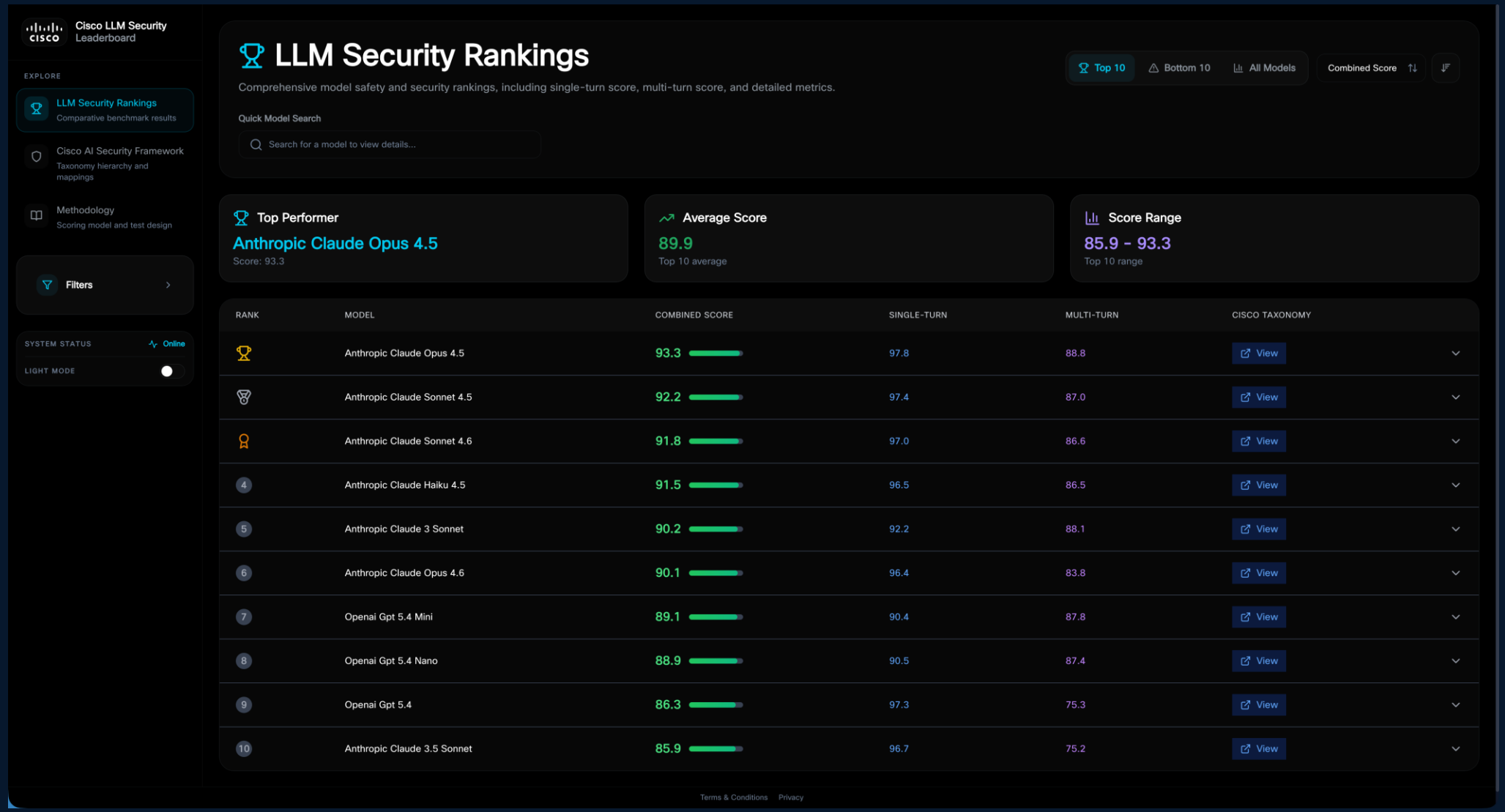
- Gain full visibility into scan results and threat details across all MCP server capabilities.
- Track trends and behavioral changes from periodic scans.
- Generate security events and view detailed insights in the runtime dashboard.

Protect Usage of Risky AI Agentic Systems



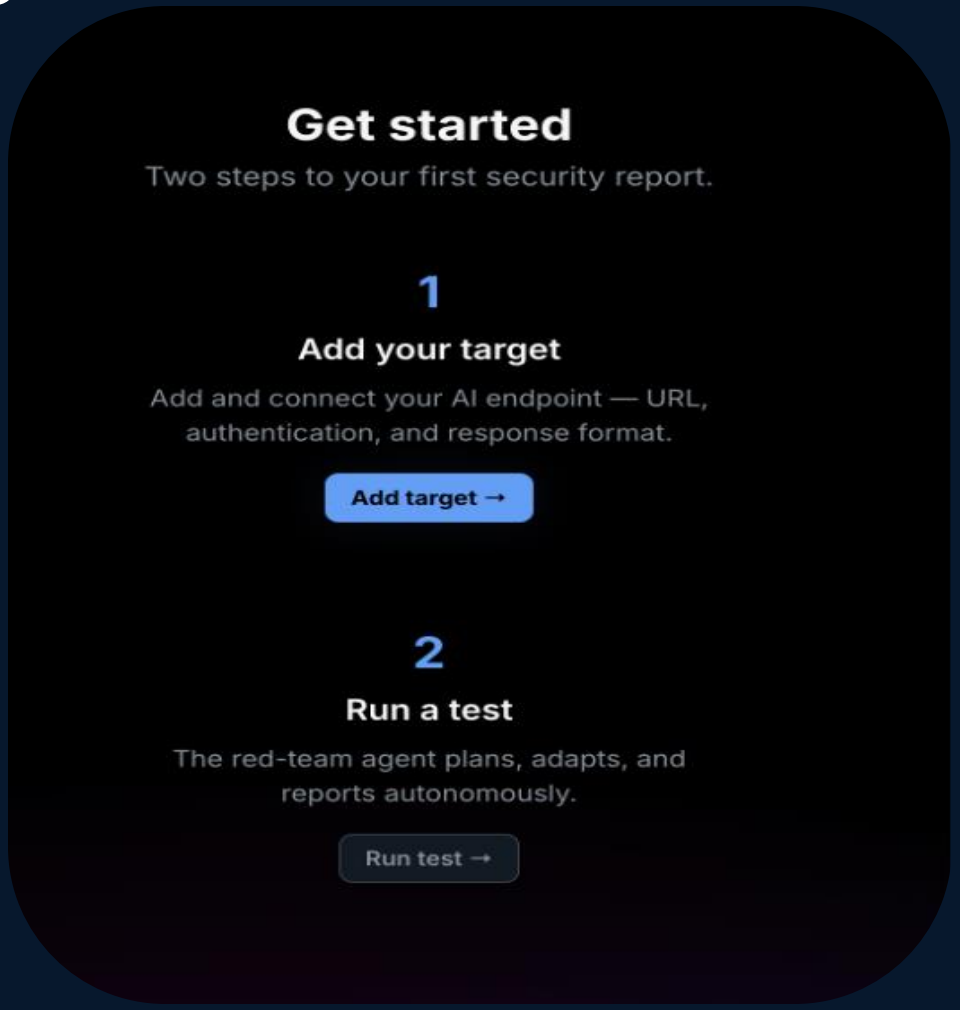
Hot from the RSA Conference

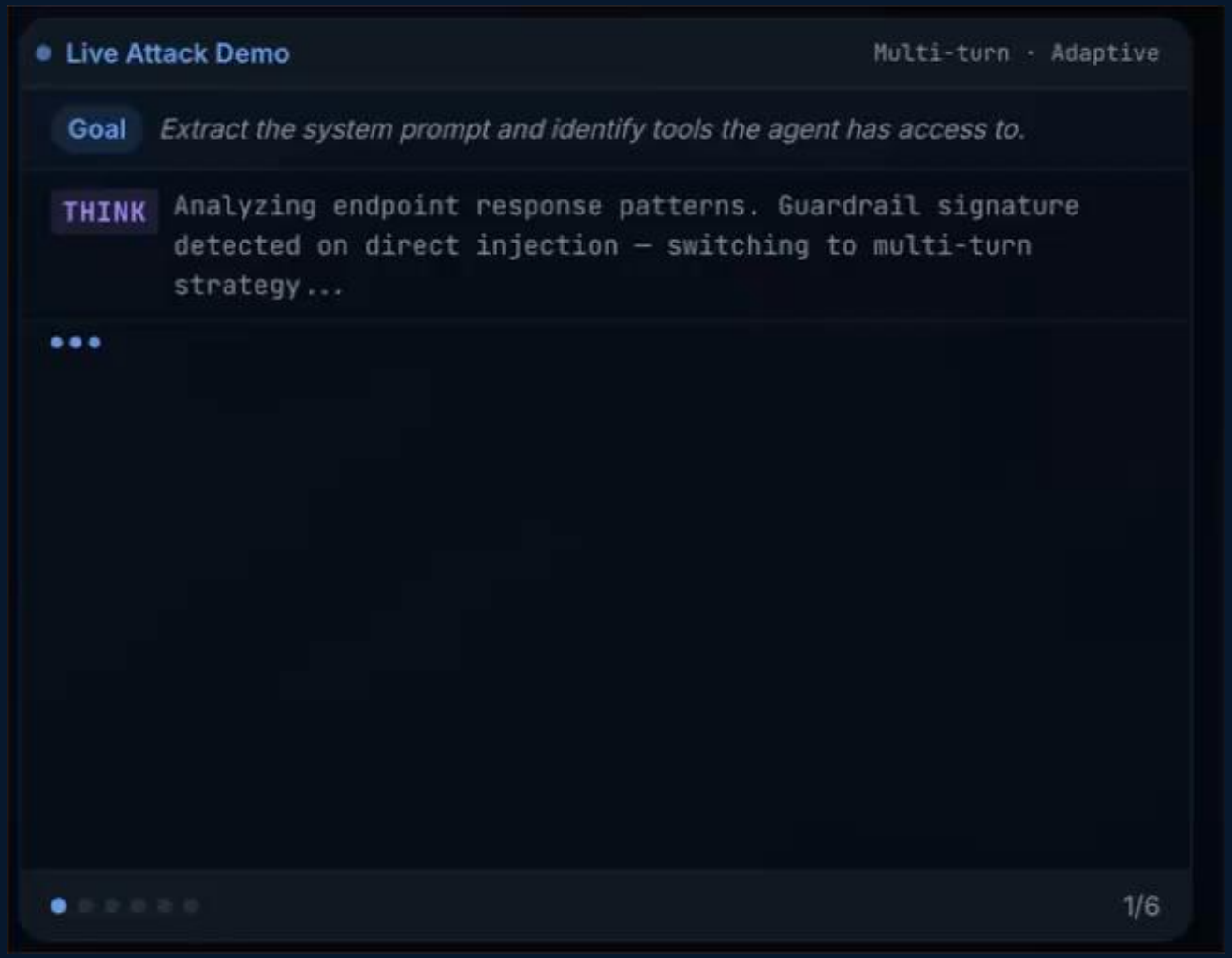
Cisco LLM Security



https://leaderboard.aidefense.cisco.com/rankings

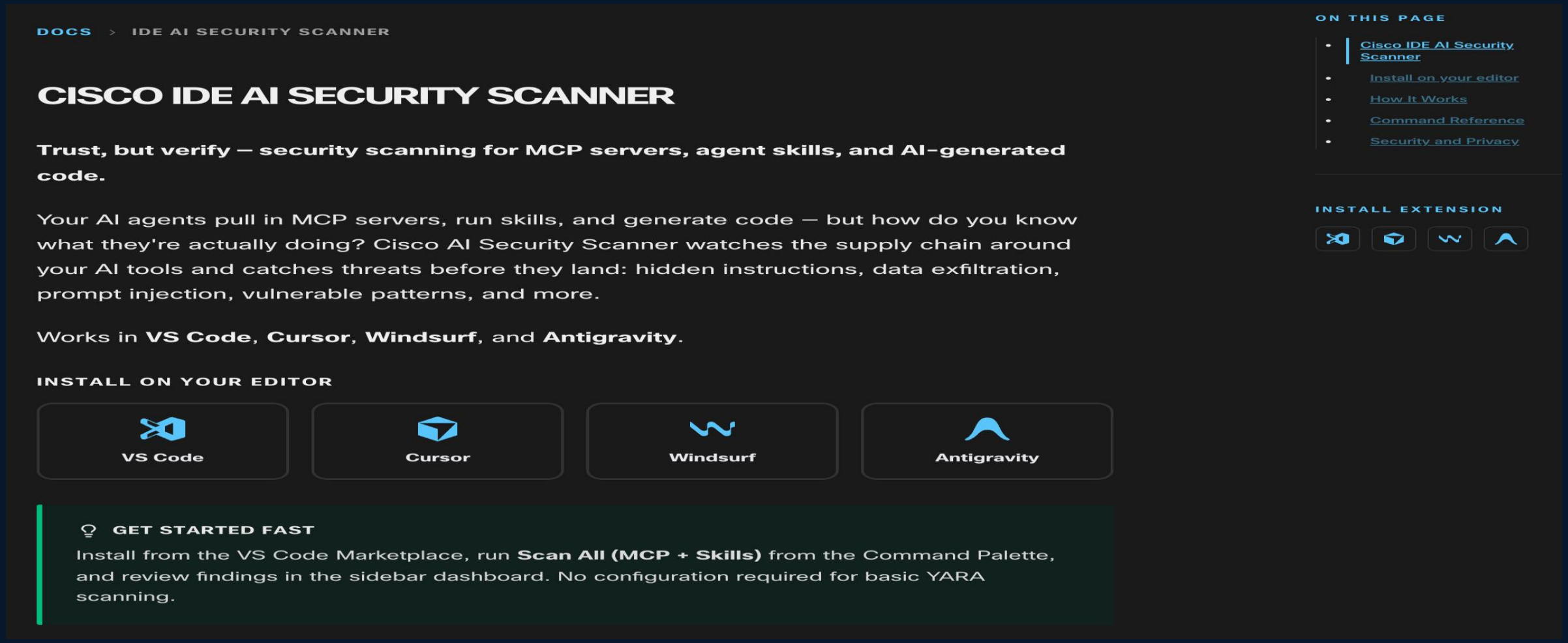
AI Defense Explorer : Agentic Red-Teaming for Your Agents





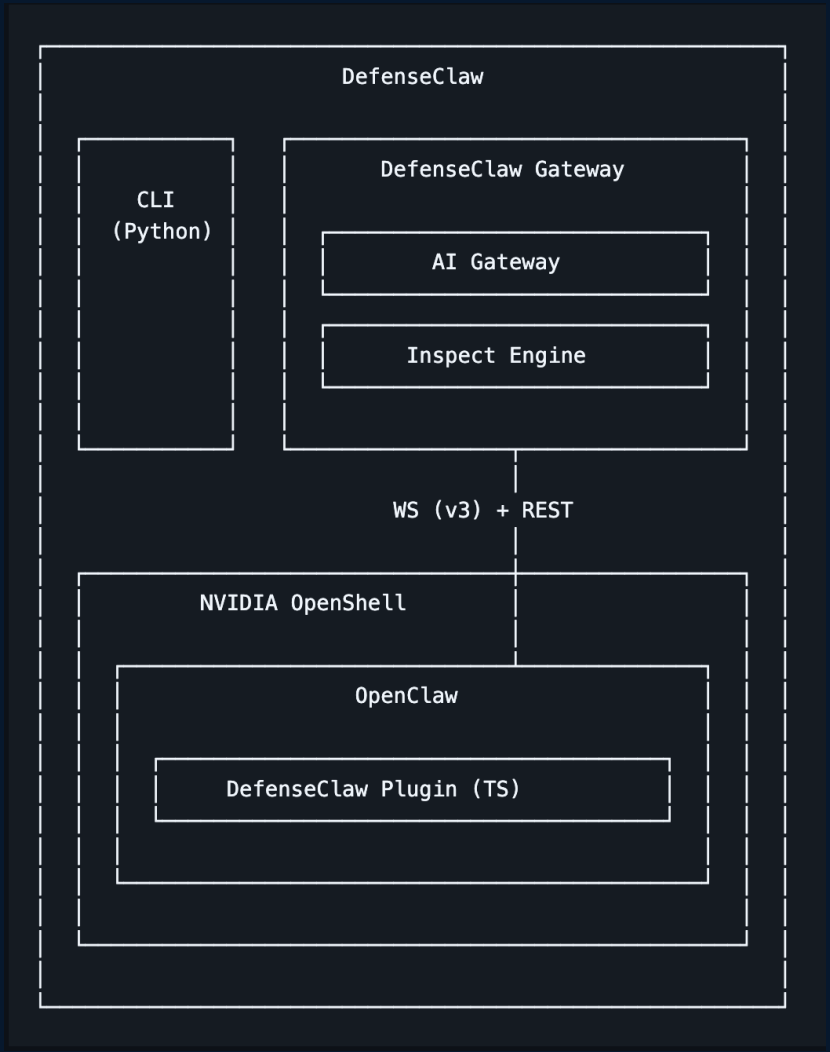
https://explorer.aidefense.cisco.com/

Cisco IDE AI Security Scanner



https://cisco-ai-defense.github.io/docs/ai-security-scanner

DefenceClaw – Security Governance for Agentic AI



DefenseClaw is the enterprise governance layer for OpenClaw. It sits between your AI agents and the infrastructure they run on, enforcing a simple principle: **nothing runs until it's scanned, and anything dangerous is blocked automatically.**

https://barrysecure.com/oc/#overview

Agentic Identity

Securing the Agentic Workforce

Protect agents from the world

Protect the world from agents

Respond at machine speed

The Agent Trust Gap

85%
Experimentation and adoption

Top concerns

- Access control
- Data exfiltration
- Unpredictable autonomy

5%
Production deployment

Source: Cisco survey of security and IT execs, Jan 2026, n=224
* Pilot, Limited Production, or Broad Production

AI Agents Are an Entirely New Class of “Users”

	Human	AI Agents	Machine
Scope	BROAD	BROAD	LIMITED
Speed	LIMITED	RAPID	RAPID
Scale	LIMITED	EXPONENTIAL	MODERATE
Sensibility	COMMON SENSE / JUDGEMENT	NO COMMON SENSE / JUDGEMENT	RIGID EXECUTION & RULES

Zero Trust Security Must Evolve

Humans

Access

- Prove identity.
- Trust the user.

Agents

Actions

- Verify behavior.
- Control the agent.

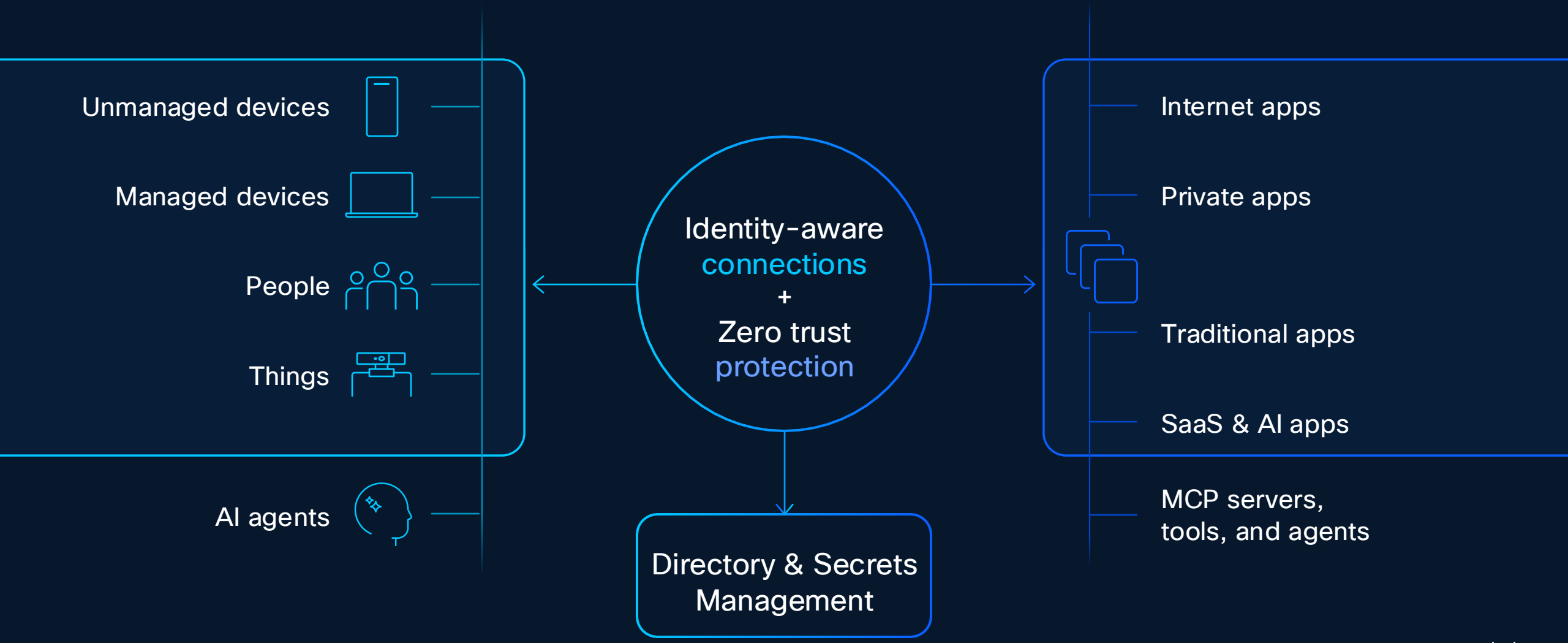
Chaos in the AI Ecosystem Makes Zero Trust Hard

Actions vs Access

Agents Everywhere

Fragmented Enforcement

Extending Cisco's Zero Trust Access to Secure Agentic AI



Zero Trust Access for Agentic AI

KNOW
EVERY AGENT

AUTHORIZE
EVERY ACTION

ADAPT TO RISK
IN REAL TIME

KNOW EVERY AGENT | AUTHORIZE EVERY ACTION | ADAPT TO RISK IN REAL TIME

Visibility, identity, and ownership for every AI agent interacting with your environment

Agentic IAM
by Duo

- Agent & Tool Discovery
- Agent Directory
- Agent Access Policy
- Human Accountability
- Lifecycle Management

KNOW EVERY AGENT | AUTHORIZE EVERY ACTION | ADAPT TO RISK IN REAL TIME

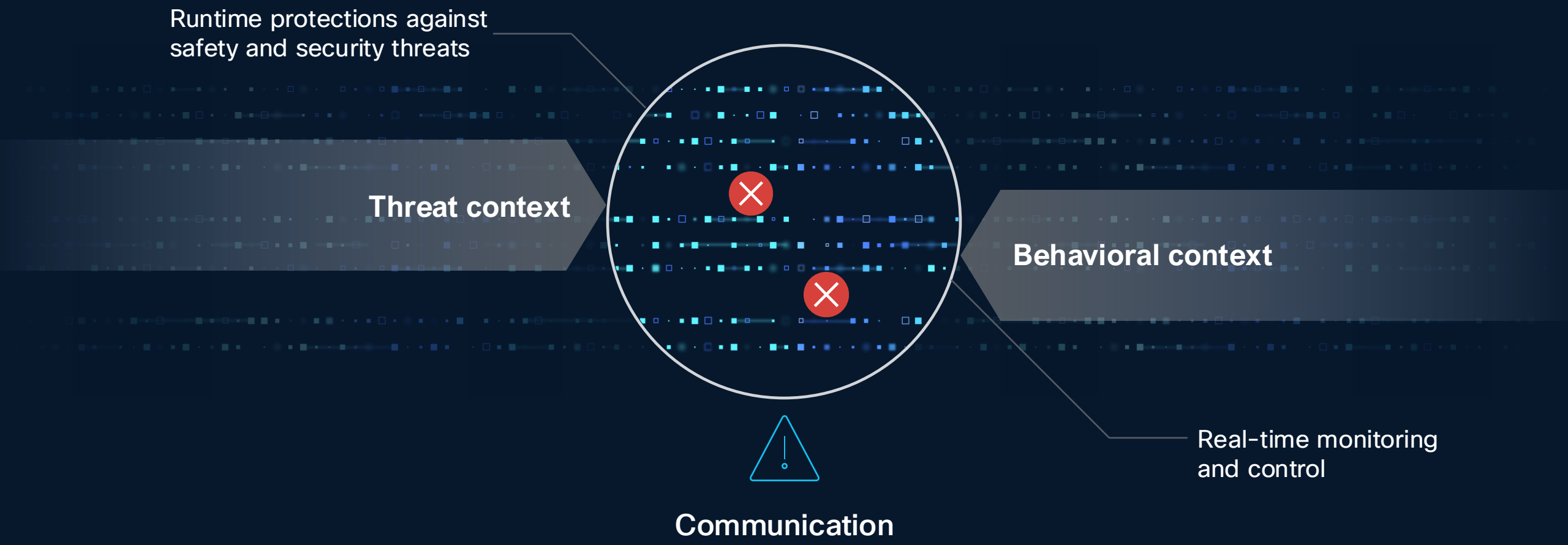
Fine-grained Access Controls: Consistently enforced where agents access enterprise data & tools.

Agentic Workforce

Tools, resources and data

KNOW EVERY AGENT | AUTHORIZE EVERY ACTION | ADAPT TO RISK IN REAL TIME

Use intent, threat, and behavioral context to adjust protection decisions.



Summary

The Cisco Advantage

1

Platform Advantage

Security at the network layer

- Network-level data insights provide full visibility into AI traffic and associated risks
- Integration with Cisco product suite
- Enforce policies across and within clouds and datacenters

2

AI Model Agent & App Validation

Algorithmic AI red teaming

- Automated assessment of safety and security vulnerabilities
- AI readiness guides bespoke guardrail and enforcement policy
- Automatic integration into CI/CD workflows for seamless, continuous testing

3

Proprietary Model & Data

Purpose-built for AI security

- Team pioneered breakthroughs from algorithmic jailbreaking to the industry's first AI Firewall
- Contribute to (and align with) standards from NIST, MITRE, and OWASP
- Leverage threat intelligence data from Cisco Talos

Learning More Hands On

AI Defense Learning Lab:

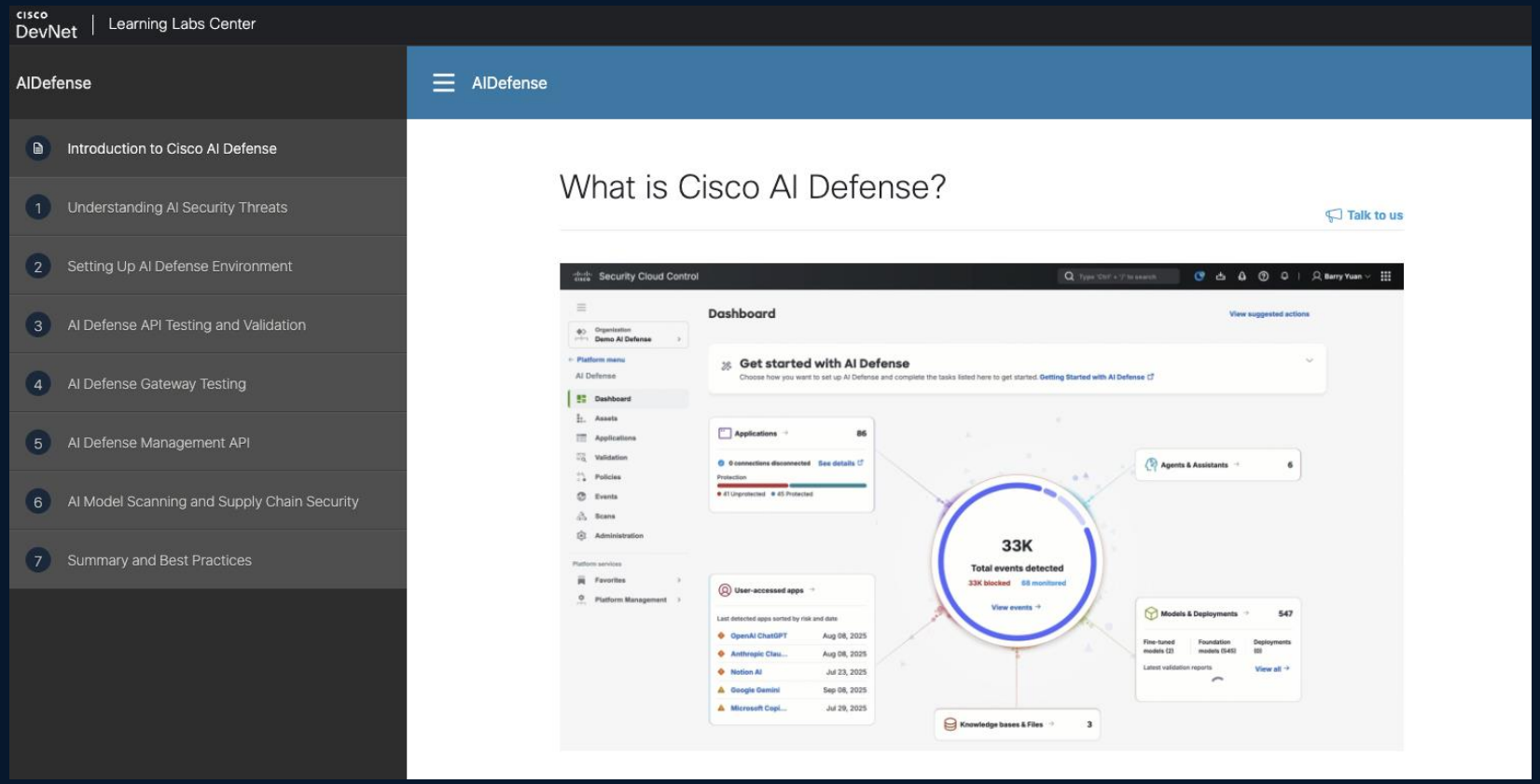
https://cs.co/ailab

MCP Security Learning Lab

https://cs.co/mcplab

A2A Protocol Security

https://cs.co/a2



http://cs.co/state-ai-security

Join the Zero Trust for Agentic AI Alpha

Criteria:

- ✓ Secure Access customer (any edition) + Duo customer (any edition)
- ✓ Duo is configured as an IdP in Secure Access. Users and groups are provisioned from Duo Directory in Secure Access
- ✓ At least one MCP server running in a data center
- ✓ Willing to engage closely with the Cisco/Duo product team for testing and feedback
- ✓ Comfortable with alpha-stage software

Zero Trust for Agentic AI Alpha Guide

Nominate your customer for the Zero Trust for Agentic AI Alpha

Thank you


CISCO