CISCO Connect

Cisco Secure AI Factory with NVIDIA

Johan Lagrand - NVIDIA

François Mauclin, Patrice Nivaggioli - Cisco

May 21st, 2026

Agenda
AI-Ready Data Centers

1. Cisco Secure AI Factory with NVIDIA 12h50
2. Cisco Unified Edge 13h40
3. Scaling AI 14h30
4. AI-Ready Network 15h20
5. RetEx Crédit Agricole (Isovalent) 16h10

A year is a long time in AI

Reasoning commoditized

DeepSeek

Infrastructure a national priority

EU AI Act

Adversarial Symmetry

Glasswing

Agents are the real deal

OpenClaw

DeepSeek | Tokenomics | Tokenmaxxing | Probablistic Waste | Vibe Coding | Content Explosion | Stargate | EU AI Act | FraudGTP | Glasswing | OpenClaw | Humans-aaS

cisco

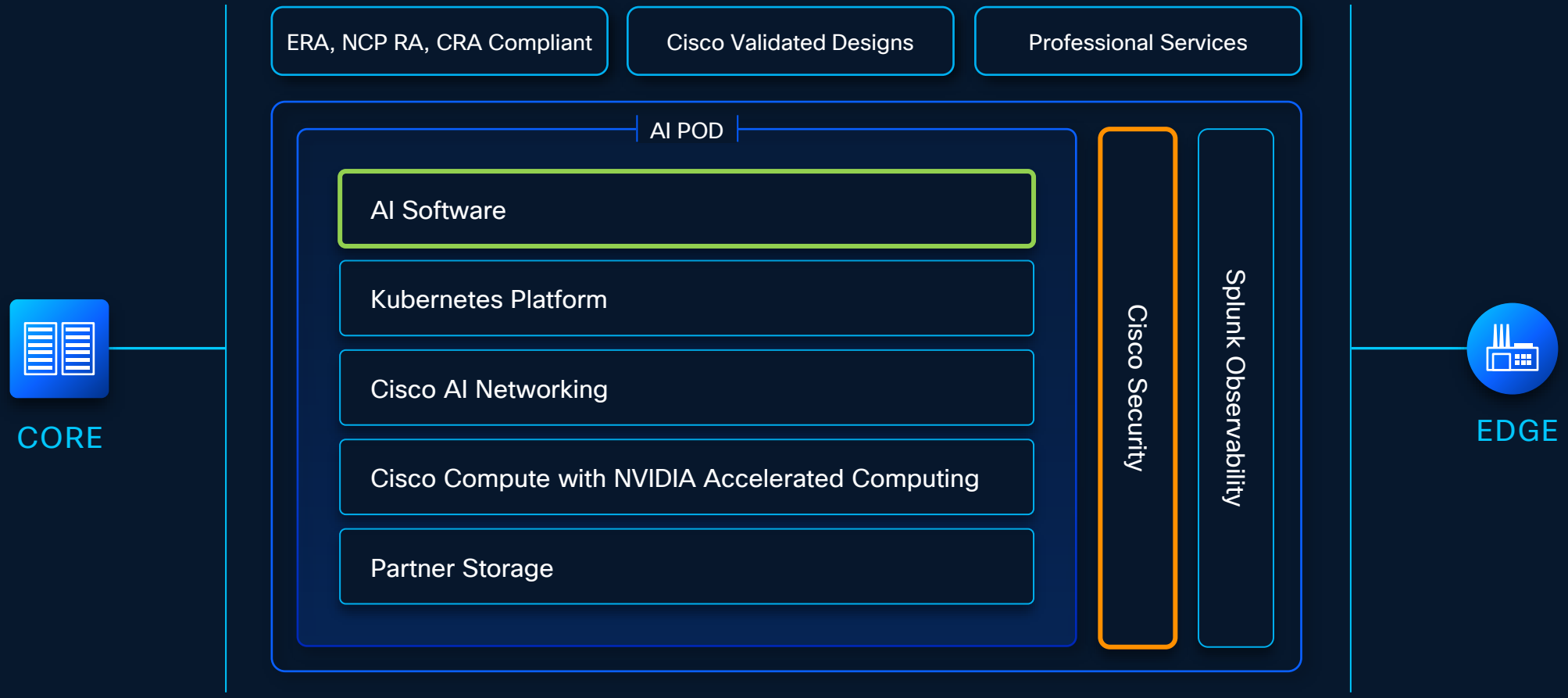
Protect the agents from the world

Protect the world from the agents

Deploy optimised infrastructure to manage agents at machine speed & scale

Cisco Secure AI Factory with NVIDIA

Eliminating friction with a modular reference architecture



NVIDIA AI Enterprise Software

Cloud Native Software Platform for Production AI

AI solutions
Fastest path to production

NVIDIA Blueprints

Standard APIs and performance optimized models

Reference workflows for AI development

CUDA-X

NVIDIA

SDKs and Frameworks with full-stack acceleration

Enterprise-grade
Built for business

Security

Development flexibility

Validated deployment guidance

SLAs with NVIDIA support

Run everywhere
Cloud Native and Certified

Cloud | Data center | Workstation | Edge

NVIDIA Certified

kubernetes

NVIDIA NPN Cloud Partners

aws

Microsoft Azure

Google Cloud

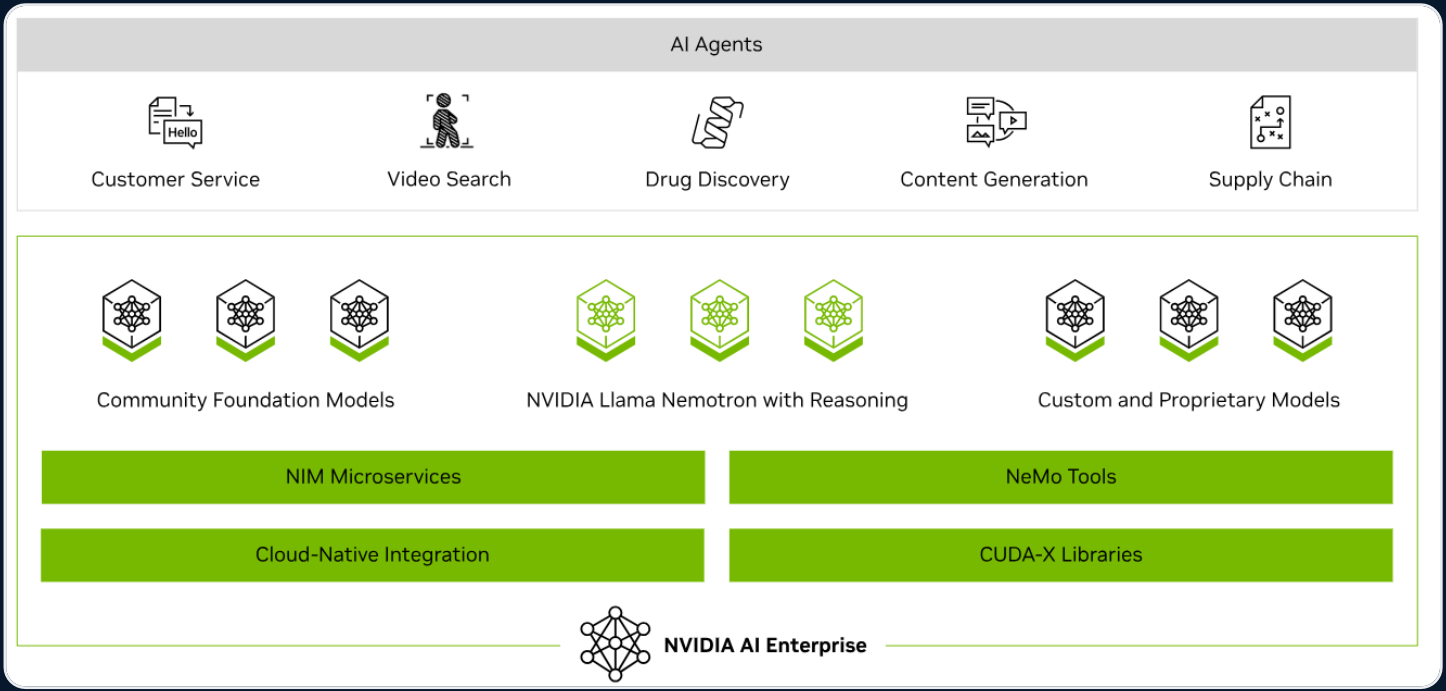
ORACLE CLOUD Infrastructure

cisco

NVIDIA Enterprise Software

The NVIDIA Enterprise tools in the Cisco Secure AI Factory with NVIDIA provide support for each step in the training, optimization, and deployment of AI agents.

Production-ready software for agentic AI



Deploy the latest state-of-the-art AI models

Explore the NVIDIA NIMs catalog of enterprise-ready, performance-optimized models for efficient inference and reasoning.

Build and manage data flywheels with NeMo

Discover powerful, ready-to-use model training, evaluation, and guard railing tools and RAG building blocks for optimizing agentic AI.

Customizable blueprints for your use case

Reference workflows for building fast, high-performance, and secure agentic systems using the latest machine learning best practices.

Software for AI

NVIDIA

NVIDIA Enterprise	NeMo	NIM	Blueprints
NVIDIA Run:ai	AI Workload & GPU Orchestration		

cisco

NVIDIA Enterprise Software

The NVIDIA Enterprise tools in the Cisco Secure AI Factory with NVIDIA provide support for each step in the training, optimization, and deployment of AI agents.

Accelerate Time To Value with NVIDIA Software

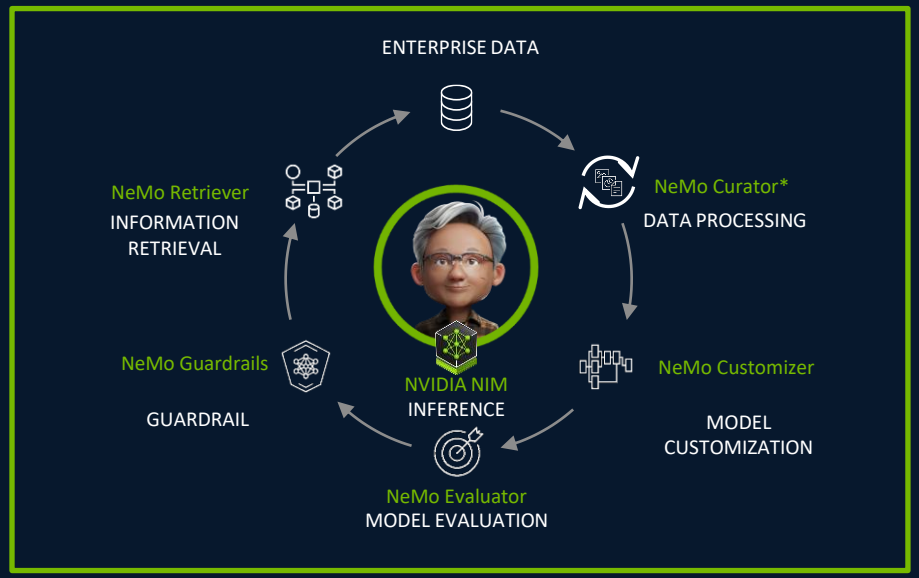
Portable: Run and scale cloud-native containers anywhere, maintaining control of data and apps

Easy: Spin up production-ready endpoints in minutes with built-in observability, autoscaling and continuous updates

Enterprise Ready: Gain confidence with NVIDIA-secured and verified software including monitoring and patching

Performant: Leverage the best performing inference technology, preoptimized and continuously updated

NeMo & NIM Microservices

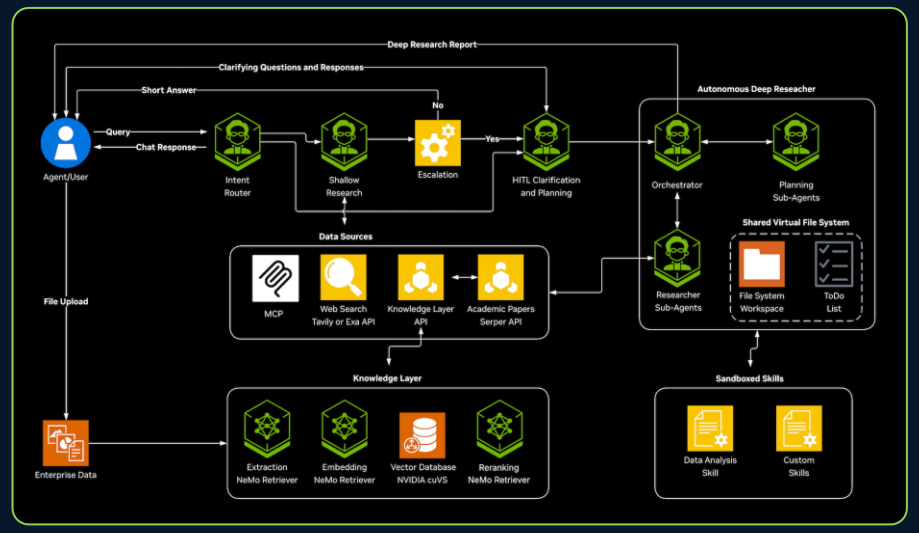


Software for AI				
NVIDIA	NVIDIA Enterprise	NeMo	NIM	Blueprints
	NVIDIA Run:ai	AI Workload & GPU Orchestration		

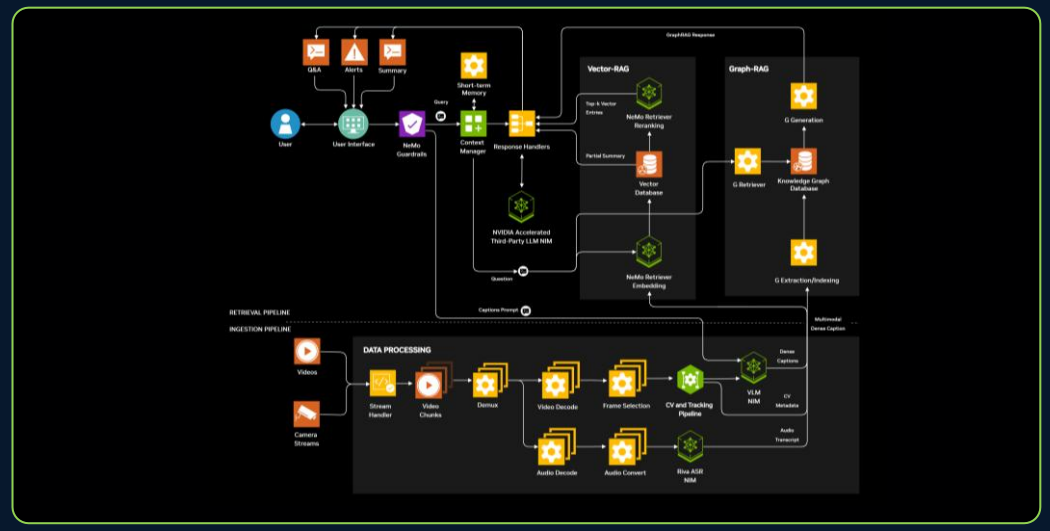
NVIDIA Enterprise – Blueprints for use cases

https://build.nvidia.com | https://catalog.ngc.nvidia.com

AI-Q Deep Research



Video Search & Summarization



Blueprints offer sample workload designs for common AI use cases. These blueprints leverage technology available in the NVIDIA Enterprise software suite. These blueprints are but a few of infinite use cases that can be developed with AI software.

Software for AI

NVIDIA

NVIDIA Enterprise — NeMo | NIM | Blueprints

NVIDIA Run:ai — AI Workload & GPU Orchestration

cisco

NVIDIA Run:ai

Software for AI

NVIDIA

NVIDIA Enterprise

NeMo

NIM

Blueprints

NVIDIA Run:ai

AI Workload & GPU Orchestration

Resource Management

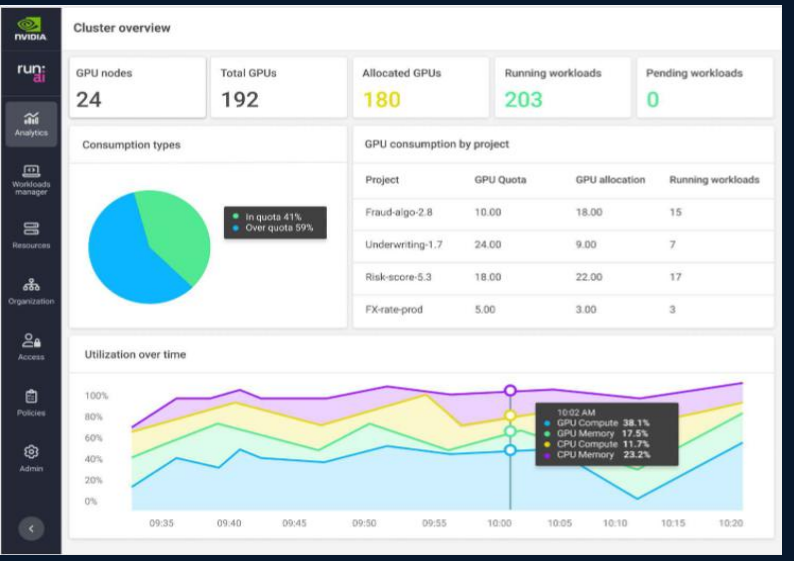
- Infrastructure Pooling
- Policy Engine

AI Lifecycle Integration

- Scheduling
- GPU Orchestration

Workload Orchestration

- Scheduling
- GPU Orchestration



AI-Native Workload Orchestration

Purpose-built for AI workloads, NVIDIA Run:ai delivers intelligent orchestration that maximizes compute efficiency and dynamically scales AI training and inference.

Flexible AI Deployment

NVIDIA Run:ai supports AI workloads wherever they need to run, whether on prem, in the cloud, or across hybrid environments, providing seamless integration with AI ecosystems.

Unified AI Infrastructure Management

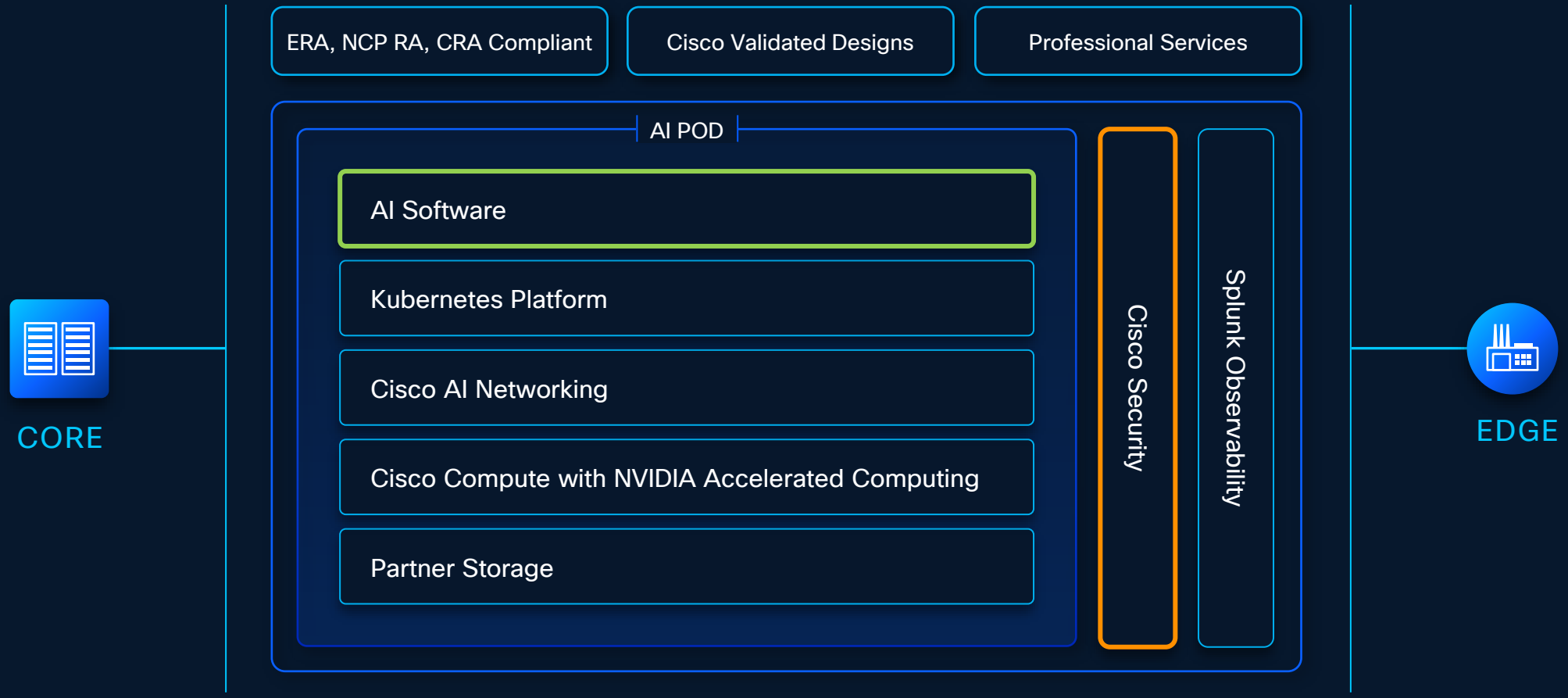
NVIDIA Run:ai provides a centralized approach to managing AI infrastructure, ensuring optimal workload distribution across hybrid, multi-cloud, and on-premises environments.

Open Architecture

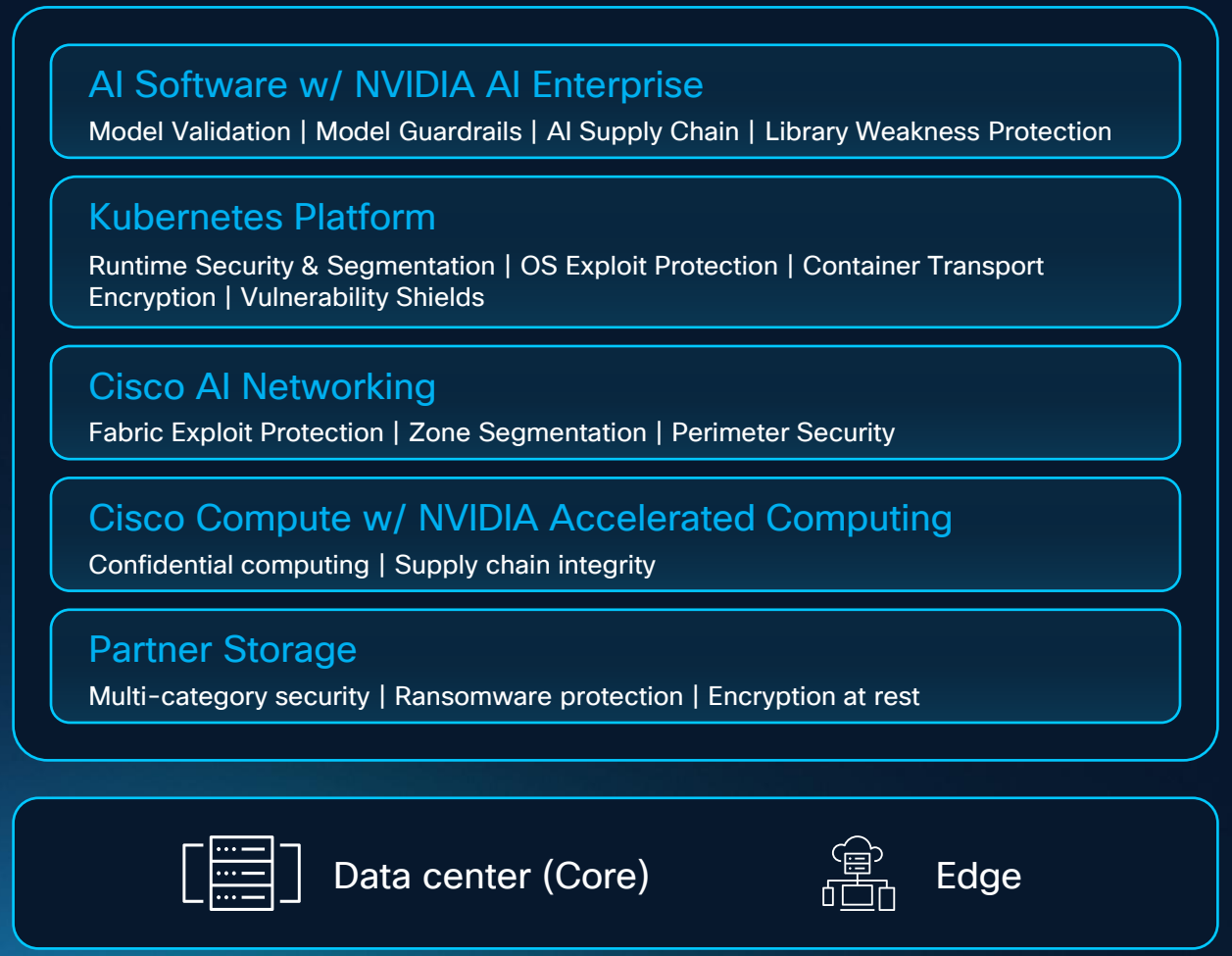
Built with an API-first approach, NVIDIA Run:ai ensures seamless integration with all major AI frameworks, machine learning tools, and third-party solutions.

Cisco Secure AI Factory with NVIDIA

Eliminating friction with a modular reference architecture



Full stack protection, threat detection, investigation, response engineered in the solution

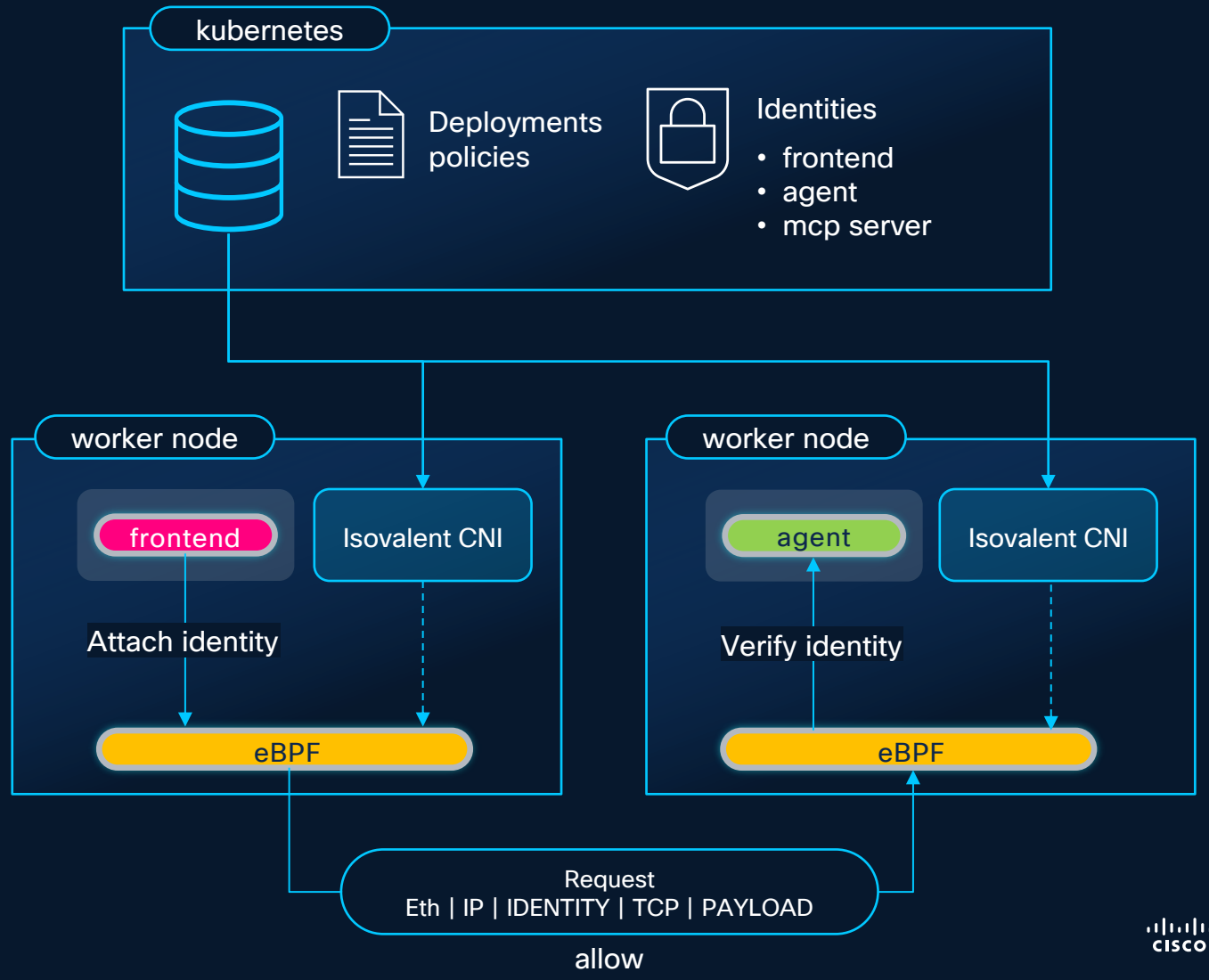


Identity-based Workload Security

Network filtering

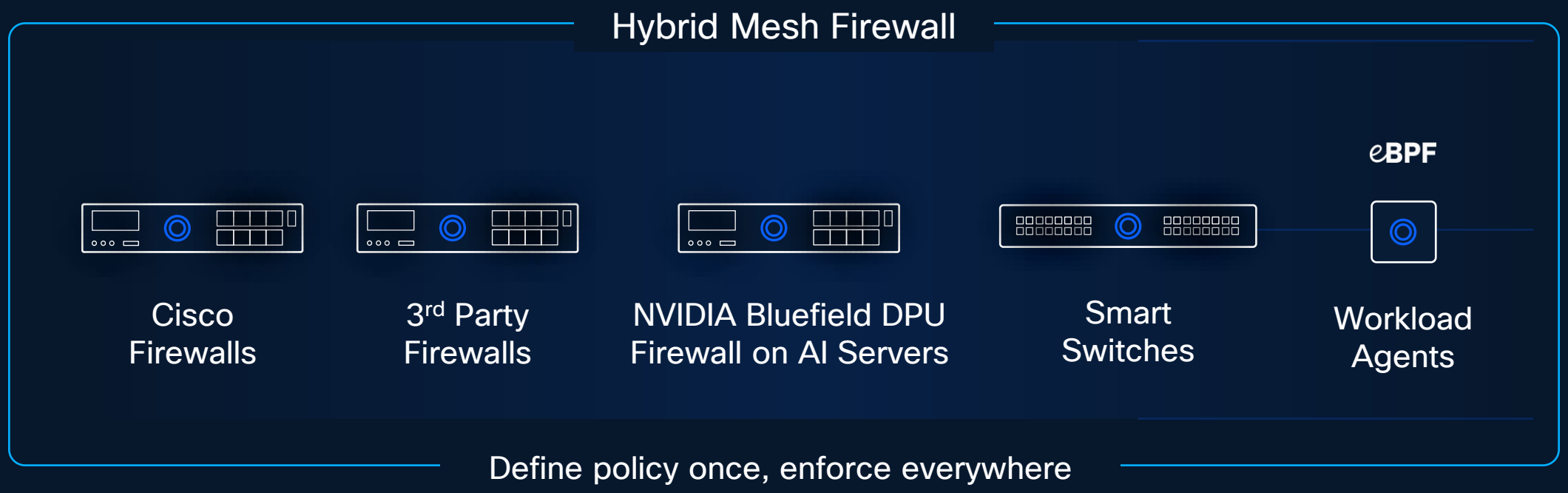
- Kubernetes networking
- Load balancing
- Kubernetes services
- Identity-based security
- L7 policies

```
apiVersion: "cilium.io/v2"
kind: CiliumNetworkPolicy
metadata:
 name: "agent-rule"
spec:
 endpointSelector:
  matchLabels:
   role: agent
 ingress:
 - fromEndpoints:
  - matchLabels:
    role: frontend
```



Hybrid Mesh Firewall:

Reduce attack surface and stop lateral movement



Securing a Document Processing & Q/A Assistant

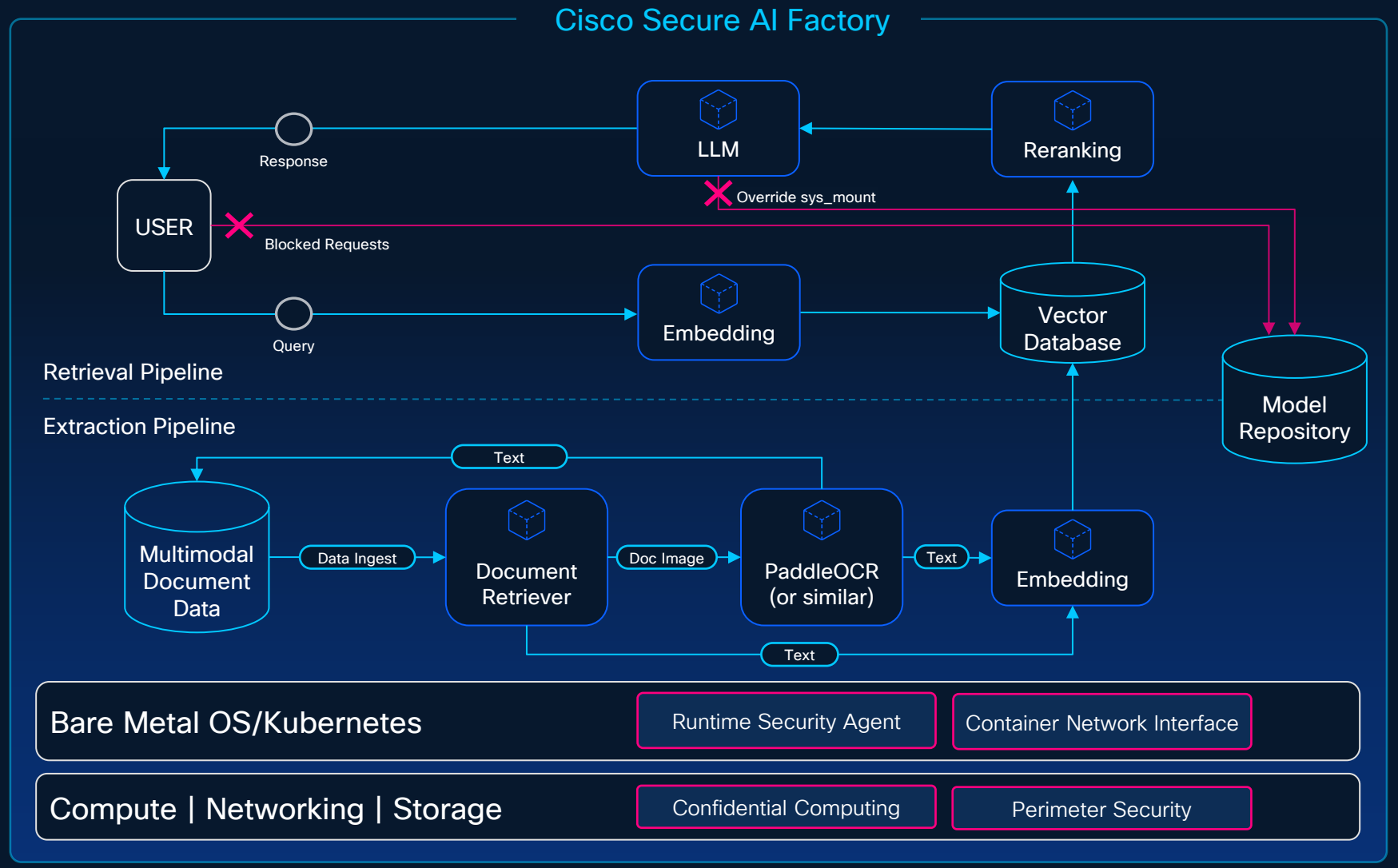
An example

Platform Runtime Security
Mitigate exploitation of known weaknesses in software and operating systems. Assure compliance with File Integrity Monitoring

Secure Container Networking
Layer 3, 4, and 7 network policy policing ingress, egress, and inter-NIM communication

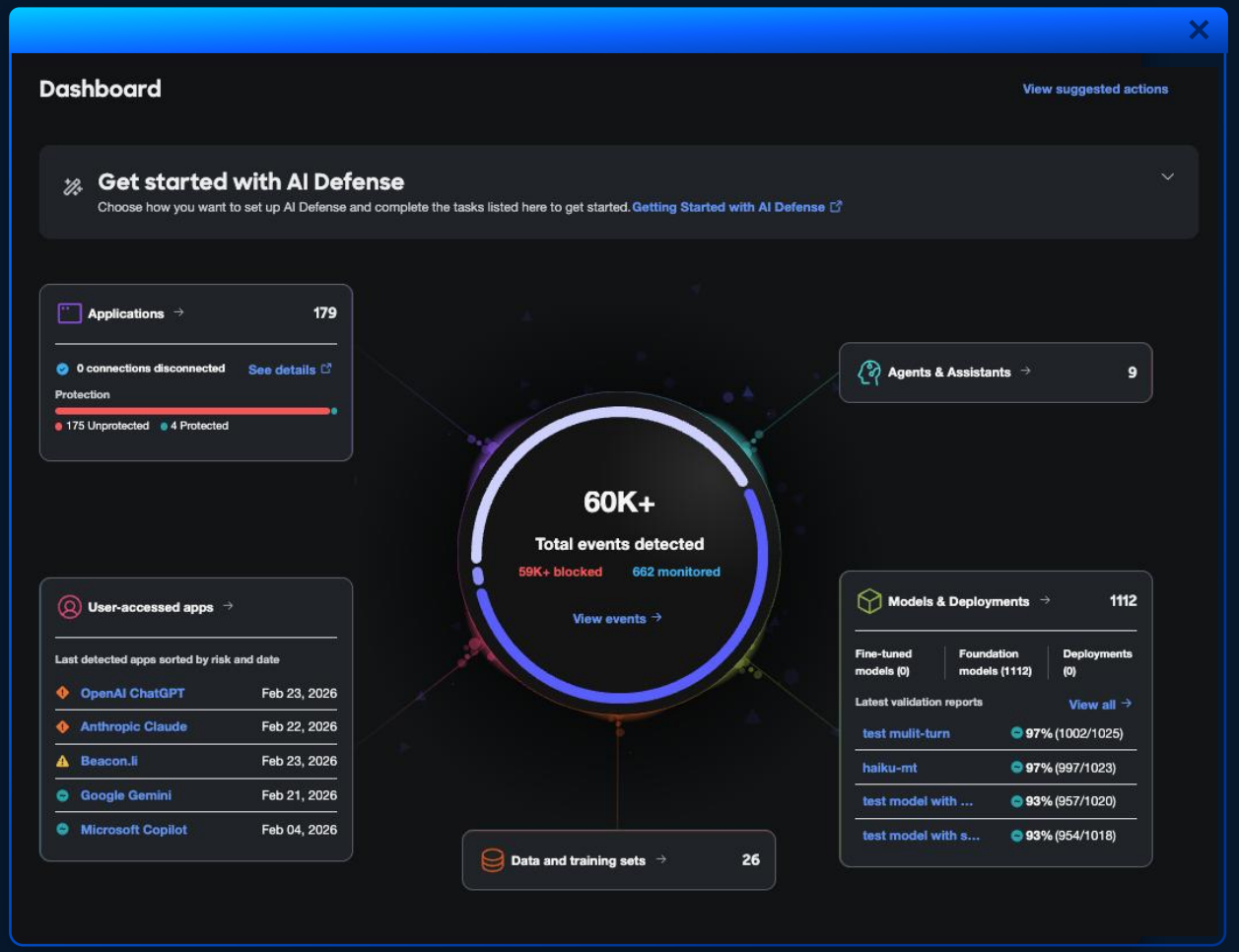
Confidential Computing
Encrypted execution for containers/VMs

Perimeter Security
Broker and police access to the supporting AI cluster. Mitigating DDoS and enforcement multi-tenant segmentation.



Cisco AI Defense

AI-native model and agentic security



Discovery

Identify AI assets

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

Detection

Scan for threats and detect vulnerabilities

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

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

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.

cisco

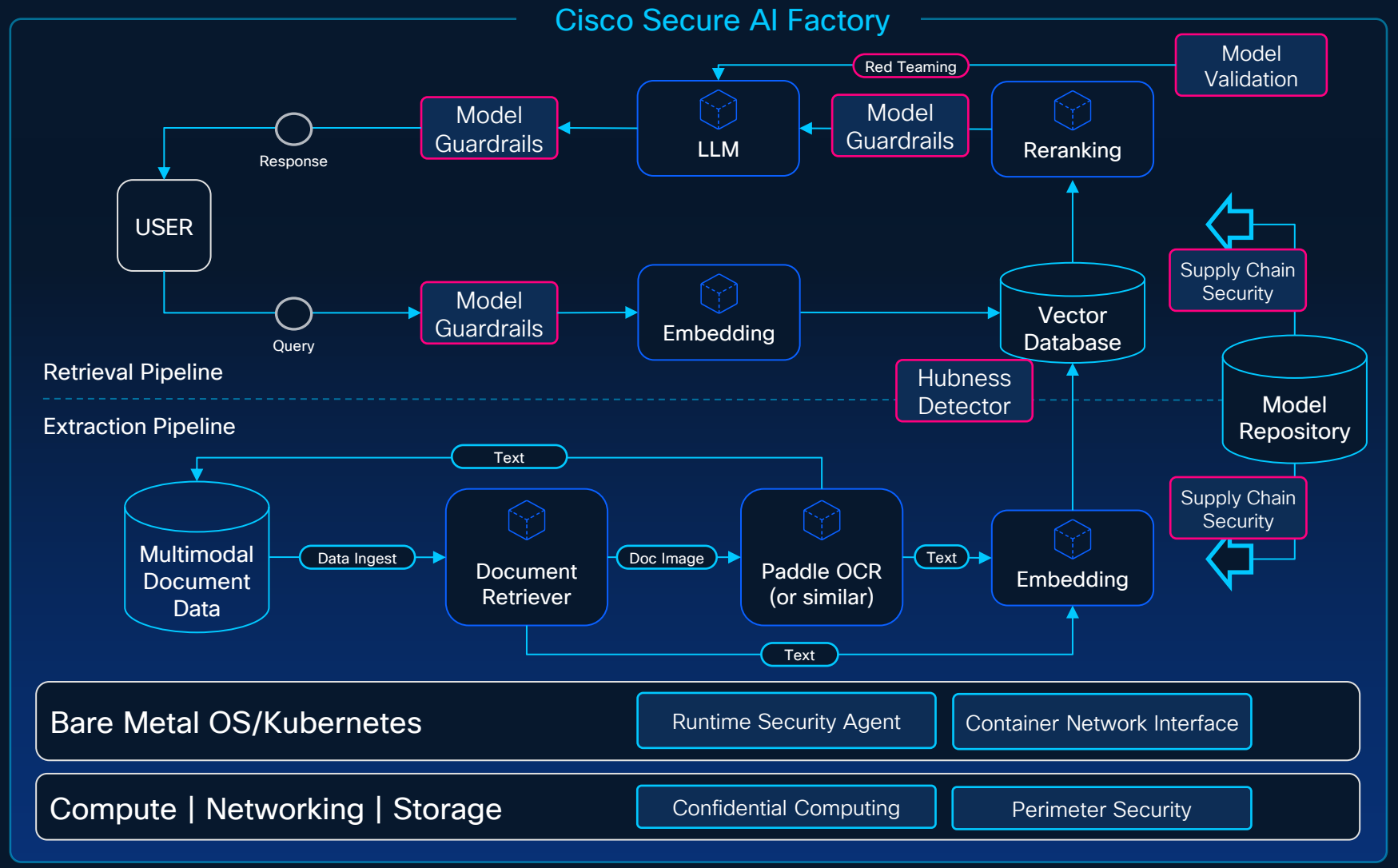
Securing a Document Processing & Q/A Assistant

An example

Model Guardrails
Police system input/outputs for LLM policy violation. Preventing IP loss and misuse.

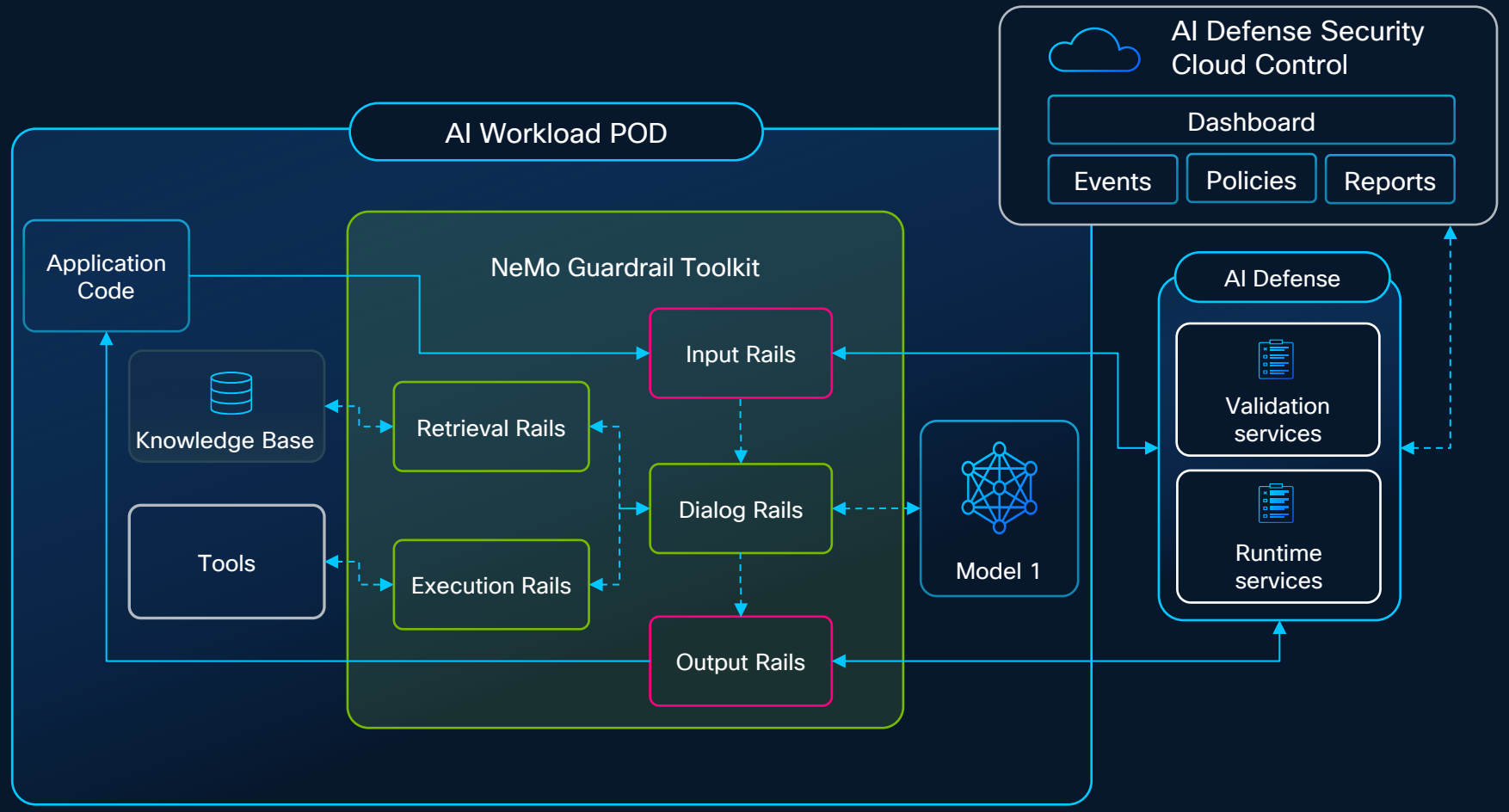
Model Validation
Red-teaming to understand which GenAI threats an LLM is susceptible to. Informs policy creation.

Supply Chain Security
AI BOM and provenance. Model files, Skills and MCP scanning.

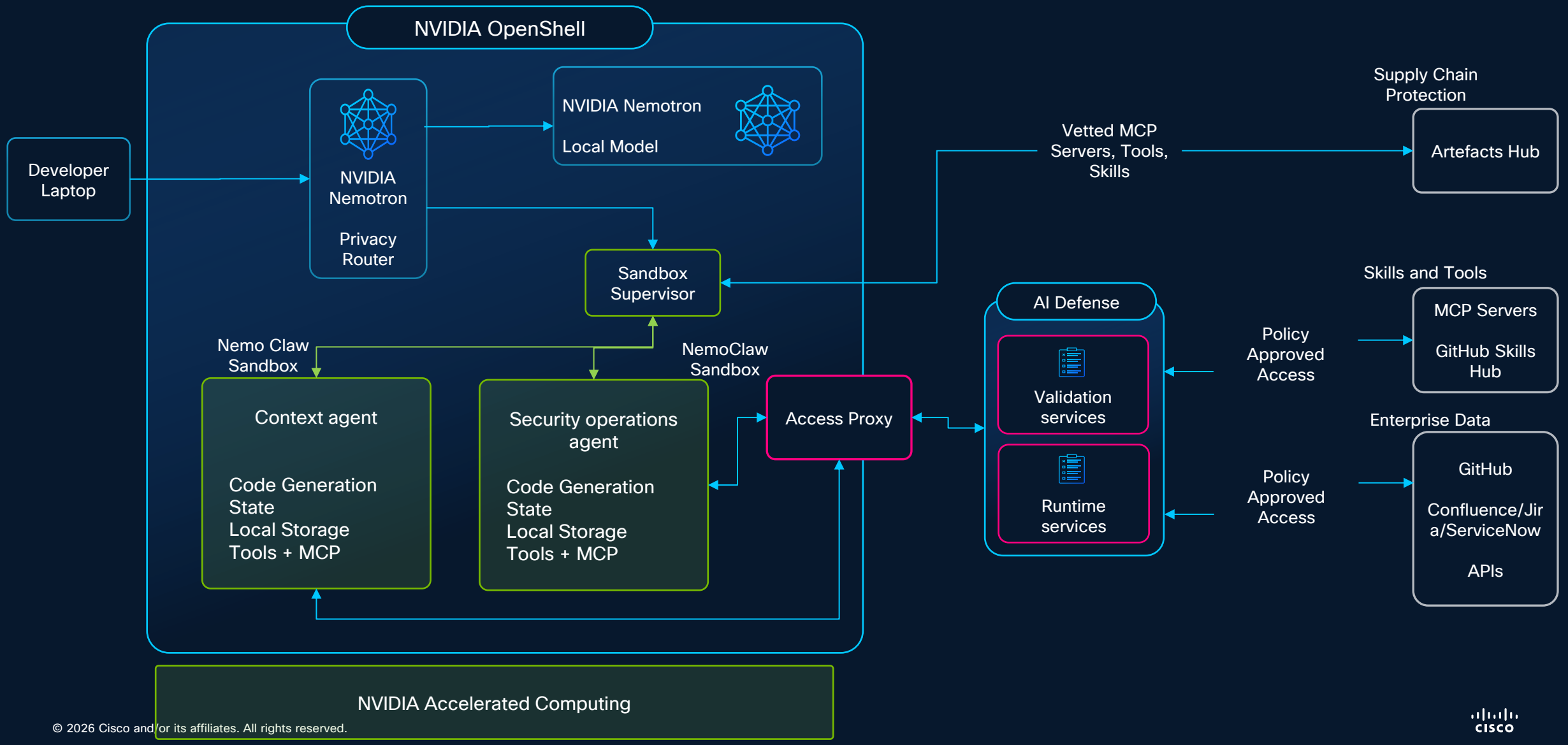


AI Defense & NeMo Guardrails Integration

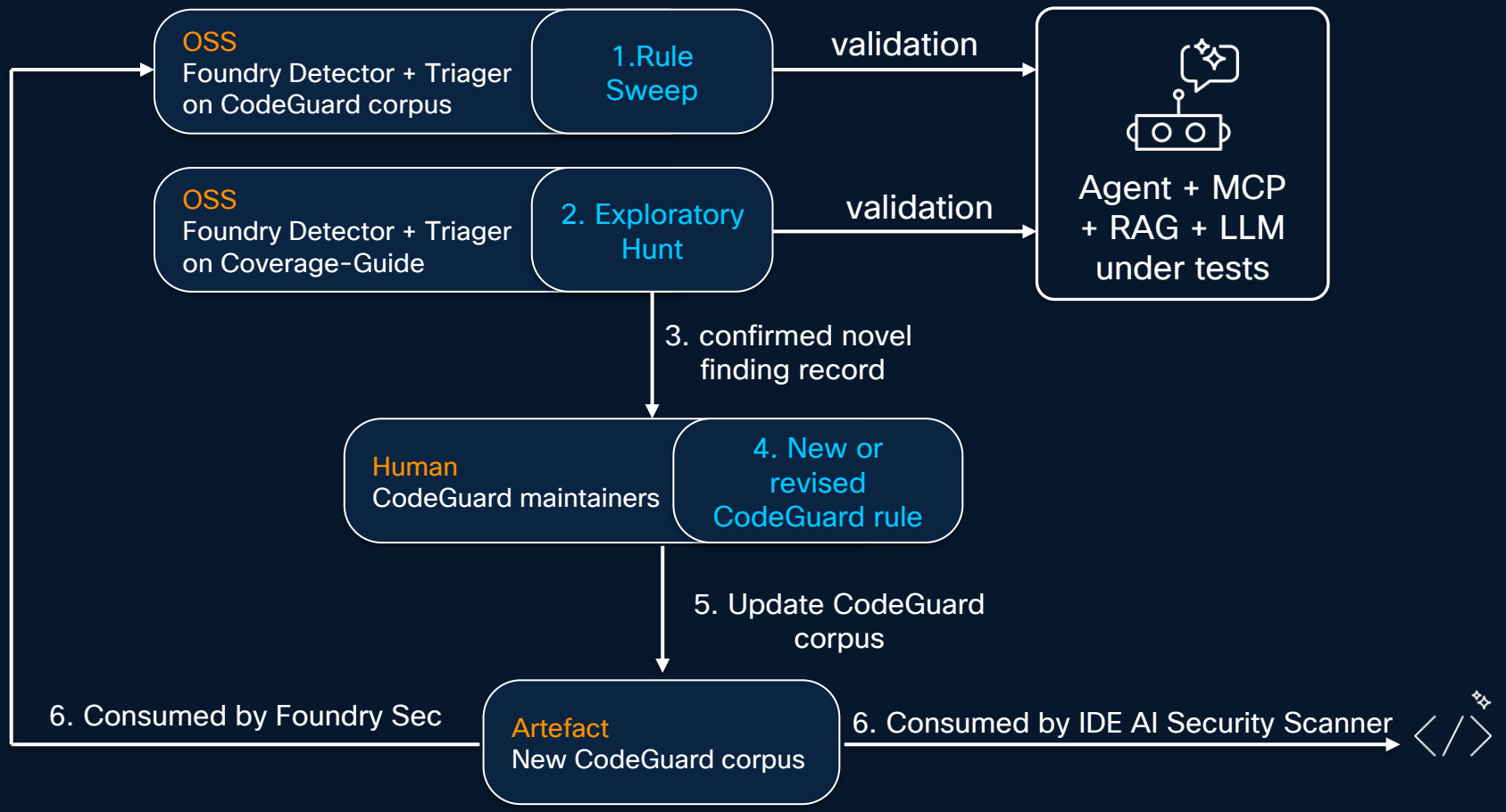
- AI Defense provides input & output guardrails via API disposition
- Common guardrail policy for on-premises and cloud deployments.
- Supports additional guardrail types included in the NeMo Guardrail Toolkit
- Nvidia documentation link



AI Defense & NemoClaw integration with DefenseClaw



Adversarial Symmetry – Detection ↔ Prevention flywheel



A shift toward embedded active defense is required with emerging AI-enabled threats

Ability to update protections independently of major software or hardware refresh cycles

Live Protect

Vulnerability shielding for Cisco AI networking devices

Cisco Secure AI Factory with NVIDIA

Eliminating friction with a modular reference architecture