

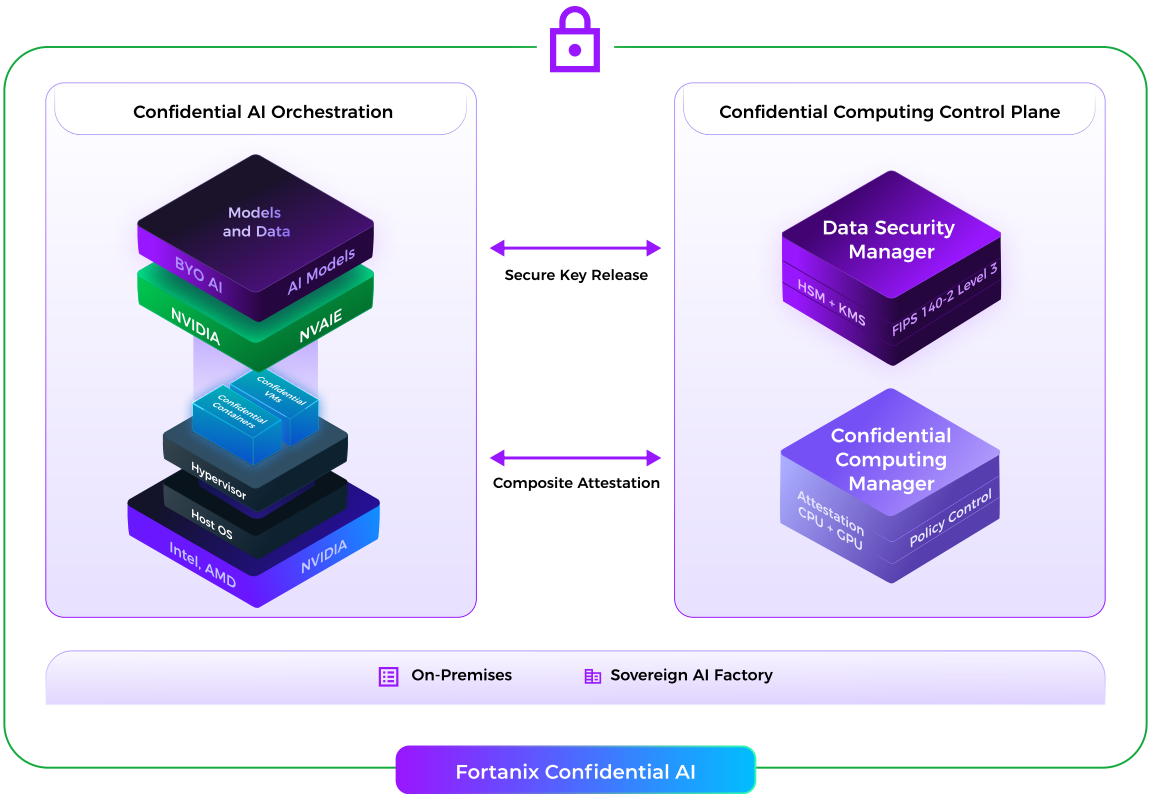
Fortanix Confidential AI

Protect AI Models and Sensitive Data Throughout the Entire AI Lifecycle

Data exposure and intellectual property theft are blocking AI adoption across enterprises and nations. Over half of enterprise IT leaders cite data security and compliance risks as a fundamental obstacle to AI deployment, while AI providers hesitate to deploy their proprietary models in third-party environments, whether AI Factories, cloud platforms, or edge infrastructure. The problem is data in use. Encryption protects data at rest and in transit, but the moment AI starts processing it, during training or inference, sensitive data and model IP are exposed in memory. Anyone with administrative access to the underlying infrastructure, from cloud operators to insiders to compromised hosts, can potentially see both. For regulated enterprises, AI providers, and sovereign organizations alike, that gap is often why promising AI projects never make it past the pilot.

Fortanix Confidential AI

Fortanix Confidential AI removes the barrier to AI adoption by protecting both proprietary AI models and sensitive data while in use, so organizations can accelerate AI innovation without compromising security, trust, or sovereignty. The joint NVIDIA and Fortanix solution is built on Confidential Computing GPUs that run AI workloads inside Trusted Execution Environments (TEEs)-- secure, hardware-isolated environments that are walled off from the rest of the hardware and operating system. Whether you deploy AI workloads in confidential virtual machines or through Confidential Containers (CoCo) on Kubernetes, Fortanix verifies every workload down to the CPU and GPU through composite attestation, and only then releases keys from a FIPS-validated HSM. The result is Zero Trust for AI, even while it runs, with cryptographic guarantees that sensitive data and proprietary models stay confidential, even from privileged insiders.



How Fortanix Confidential AI Works

Fortanix Confidential AI protects AI IP and data in use through composite attestation for CPUs and GPUs and secure key release that happens only after the environment and policy are verified. These controls block unauthorized access, even from privileged administrators, so organizations can run the AI they want and need, wherever their data lives.



Secure Enclave Management

Create and manage secure enclaves with Fortanix Confidential Computing Manager (CCM), isolating AI workloads from the OS, hypervisor, and admins. Bring your own model or a frontier model, encrypted at the source and deployed in confidential VMs or Confidential Containers (CoCo).



Composite Attestation

Before anything is decrypted, CCM verifies the exact CPU and GPU hardware and software environment as a single chain of trust, confirming the enclave is genuine and untampered. Every workload gets verifiable, auditable proof of its security.



Secure Key Release

Fortanix Data Security Manager (DSM), a FIPS 140-2 Level 3 HSM with built-in Key Management, releases keys only to verified TEEs, enforcing strict key custody and role-based access controls. Data and models are unlocked only inside the enclave, so no admin, cloud operator, or privileged insider can access them, even while in use.

Supported Workloads:

- Confidential VMs and Confidential Containers (CoCo)
- GPU: NVIDIA Hopper & Blackwell (Confidential Computing)
- CPU: Intel TDX & AMD SEV-SNP
- Deployments: On-premises, Air-Gapped, Sovereign AI Factory, Public Cloud

Benefits of Comprehensive AI Lifecycle Protection

- ✓ **Protect Model IP:** Deploy AI models to third-party environments, cloud platforms, or edge locations without revealing IP
- ✓ **Confidential Inference:** Process sensitive inputs and outputs without exposure to infrastructure providers
- ✓ **Prevent Data Leakage:** Ensure data used in the AI workload remains encrypted even while being processed
- ✓ **Multi-Party Collaboration:** Enable data sharing and model access across organizations without exposing underlying assets
- ✓ **Zero-Trust Architecture:** Protection is backed by attested cryptographic verification for CPU and GPU silicon, not software policies, eliminating the need to trust cloud providers, infrastructure operators, or administrators
- ✓ **Compliance-Ready:** Meet regulatory requirements for Cloud Act, PII, PHI, and sensitive data protection

Use Cases

Secure Distribution of Proprietary Models

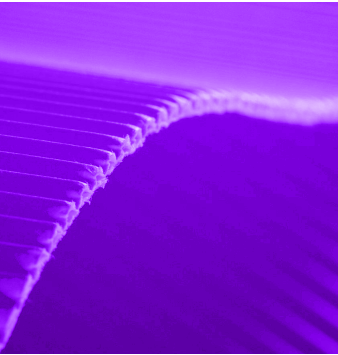
Model weights are encrypted at the source and decrypted only inside attested TEEs, so AI labs keep control of their IP and can reach enterprises that require on-premises or sovereign deployments. Enterprises run best-in-class models on their own data without exposing prompts or outputs.

Multi-Tenancy and Segmentation for AI Factories

Each tenant runs in its own confidential VM or container on shared accelerated infrastructure, with hardware-enforced isolation and keys released only to that tenant. AI Factory operators get the economics of shared infrastructure, and every tenant gets the trust of a dedicated one.

Sovereign AI Clouds

Governments, financial services, healthcare, and defense enterprise can now run AI on previously inaccessible sensitive data in fully attested AI environments within a target jurisdiction, with cryptographic, auditor-grade proof.



“Our models represent years of proprietary research and engineering - protecting that IP while expanding access is a core tension in enterprise deployment. Working with Fortanix on NVIDIA Confidential Computing-backed infrastructure lets us resolve that directly, giving organizations in government, healthcare, and finance the ability to run our Text to Speech models on their own servers, on their own data.”

Kuba Abramczyk
Forward Deployed Engineer at ElevenLabs

