AI Under Seal: End-to-End Cryptographic Attestation for Critical AI Systems

Why proving that a model has not been tampered with is becoming the common language of trust — from the lab to the embedded sensor.

2026 SECURITY IMPERATIVE

CRITICAL INFRASTRUCTURE

The Stakes

The Race for Provability Has Begun

The previous wave of AI engineering measured progress in parameters, FLOPs, and benchmarks. A second, more consequential race begins in 2026: **the race for provability**.

In critical domains — energy grids, defense systems, financial infrastructure, medical devices — silent error is no longer tolerable. Whether caused by hallucination, training-time poisoning, or model substitution at deployment, an undetected alteration can propagate silently through mission-critical decision loops.

Core thesis: Sealing AI is not an engineering option. It is a condition of operability. This presentation maps the full cryptographic stack that makes it possible — and the regulatory horizon that makes it mandatory.

Where This Fits

This is the symmetric counterpart to **calibrated abstention**. Where abstention removes unreliable answers from the output, sealing removes the very possibility of silent alteration from the system — not by preventing every attack, but by making any alteration **detectable, proven, and legally opposable**.

From the lab to the embedded sensor — one unbroken chain of cryptographic proof.

Problem Statement

When "Trust" Becomes Technical Debt

1,400+

Malicious Models

Identified on Hugging Face since 2024, including fine-tuned backdoors and compromised LoRA adapters.

18,000+

Malicious OSS Packages

Catalogued by Sonatype in Q1 2025, many targeting the AI ecosystem — PyTorch, TensorFlow, Hugging Face.

€15M

AI Act Penalty

Or 3% of global revenue — the regulatory cost of invisible technical debt, enforceable from August 2026.

The Proof of Concept Already Exists

PoisonGPT (2023) demonstrated that a model could be surgically altered using ROME editing to return false factual claims on specific triggers — while passing all standard benchmarks. Today's threat is more sophisticated: poisoned LoRA adapters, compromised PyPI packages exfiltrating credentials during installation, and OWASP's elevation of **Model and Data Supply Chain Compromise to LLM03:2025**.

The Operational Question

> How do you prove that the model running in your critical system is the one that was qualified, certified, and signed at the factory — and that it has not been substituted, altered, or subverted?

Threat Mapping

Four Attack Surfaces, One Architectural Response

Every attack surface across the AI lifecycle maps to a distinct attestation layer. The 2026 breakthrough is that these layers no longer exist in isolation — they compose into a coherent cryptographic stack.

Training & Fine-Tuning Poisoning

Manipulating datasets or pre-trained components to insert a behavioral backdoor invisible to classic benchmarks. ROME surgical editing can alter factual recall in a single gradient step — undetectable without signed artifact provenance.

Weights at Rest

Tampering during transit through registries, CI/CD pipelines, and mirrors. SafeTensors mitigates arbitrary code execution but does not cover full-chain integrity without a signed AIBOM binding every artifact hash.

Runtime Manipulation

Substituting the loaded model or reading weights directly from VRAM via a compromised hypervisor. Without confidential computing, privileged host processes can access weights and inputs in cleartext.

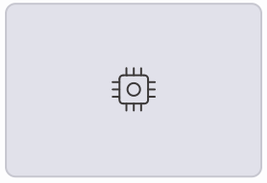
Software Supply Chain Compromise

A single malicious package in the dependency tree can exfiltrate credentials, environment variables, and model weights. The torchtriton PyPI incident (2022) remains the canonical example.

Architecture Overview

The Attestation Stack: Four Layers, One Chain of Proof

A verifiable chain spanning the full AI lifecycle — from raw training data to edge inference — where every link is bound to a cryptographically verifiable attestation artifact.

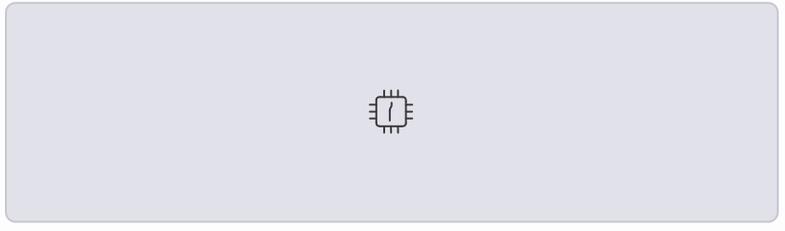


Layer 4 — Edge Attestation

Frugal signing and embedded TEEs for drones, sensors, and constrained field systems operating outside the data center.

Layer 3 — Zero-Knowledge Inference Proofs

zkLLM, zkGPT — cryptographic proofs that inference Y resulted from model X on input Z, without revealing weights or inputs.



Layer 2 — Confidential Computing

Intel TDX, AMD SEV-SNP, NVIDIA H100/Blackwell/Vera Rubin — hardware-encrypted TEEs with remote attestation for runtime execution integrity.

Layer 1 — Provenance & Supply Chain (AIBOM)

CycloneDX ML-BOM, in-toto, SLSA, Sigstore — the cryptographically signed inventory of every artifact from data to deployment.

The deliverable: a single auditor-facing artifact bundle — signed training data, attested environment, signed weights, dependency manifest, GPU attestation report, and cryptographic inference proof — all bound into one verifiable chain.

Layer 1

AIBOM & Supply Chain Provenance

What an AIBOM Traces

The SBOM lesson — *you cannot manage what you have not inventoried* — extends to AI through the **AI Bill of Materials**. An AIBOM captures the full artifact graph:

- Datasets: origin, license, preprocessing steps, known biases
- Pre-trained models: architecture, hyperparameters, performance, energy footprint
- Training pipelines: environment, configuration, metrics
- Fine-tuning artifacts: LoRA adapters, PEFT configurations
- Applied controls, guardrails, and evaluation results

Reference Standards (2026)

- **CycloneDX ML-BOM v1.7** — de facto industry standard
- **OWASP AIBOM Project** — launched 2025
- **SPDX AI & Dataset Profiles**
- **in-toto + SLSA** — each pipeline stage emits a signed provenance record bound to its output artifacts

2026 Breakthrough: AIBoMGen (arXiv:2601.05703) — a neutral observer platform that hashes all artifacts and binds them via in-toto. Any modification is detected on re-verification. From declarative inventory to cryptographically verifiable proof.

Layer 2

Confidential Computing — Sealing Inference

The Principle

Execute code and data inside a Trusted Execution Environment — a hardware-encrypted enclave isolated from the host OS, hypervisor, and cloud provider, whose integrity can be remotely attested via a signed hardware report.

CPU TEEs (Mature)

Intel TDX, AMD SEV-SNP, ARM CCA — production-ready. Composite attestation via Intel Trust Authority delivers a single JWT proving CPU TEE and GPU TEE integrity simultaneously.

GPU Breakthrough

NVIDIA H100 (2023): first GPU TEE rooted in silicon. **Blackwell B200:** extended coverage. **Vera Rubin NVL72 (2026):** rack-scale confidential computing with encrypted NVLink — multi-GPU without performance compromise.

Performance Reality

Benchmarking on H100 shows **<5–7% average overhead** for typical LLM workloads — near-zero on larger models with long sequences (arXiv:2409.03992). The performance tax is no longer an operational objection.

Layer 3

Zero-Knowledge Proofs of Inference

The Problem Confidential Computing Doesn't Fully Solve

Confidential computing protects execution — but the verifier must still trust the TEE hardware vendor. When IP sensitivity, defense secrecy, or sovereignty demands a stronger guarantee, **zk-SNARKs** offer a mathematically independent answer.

A prover cryptographically demonstrates that *inference Y is the output of model X on input Z* — without revealing model weights or, if required, the input itself. No hardware vendor trust required.

State of the Art

- **zkLLM** (Sun et al., CCS 2024) — 13B-parameter models, proofs in **under 15 minutes**
- **zkGPT** (Qu et al., USENIX Security 2025) — GPT-2 inference proofs in **under 25 seconds**

Practical Deployment Strategy

⚠ Real-time inference on very large models remains orders of magnitude away from mission tempo. Full zk proofs are not yet viable for live production systems at scale.

Two Pragmatic Mitigations

Asynchronous Proofs

Inference runs in real time inside an attested TEE. The cryptographic proof is generated post-hoc for audit and regulatory purposes — satisfying most compliance requirements without breaking latency budgets.

Partial Proofs

Only the critical segments are cryptographically proven — the final decision boundary, threshold crossing, or target classification — not the full inference graph. Efficient and audit-defensible.

Layer 4

Edge Attestation — The Most Strategic Layer

The Data Center Is Not the Frontier

Confidential GPU computing solves the data center problem. It does not solve the field problem — and the field is precisely where critical AI is massively deployed:

- Drones, radar, sonar, and signal processing units
- Embedded computers in armored and autonomous vehicles
- Industrial IoT sensors in energy and manufacturing
- Edge perception, data fusion, and tactical decision-support AI

This is the concrete difference between a lab AI and an operable AI in constrained, sovereign, and contested environments.

The Industrial Challenge: Frugality

An embedded radar sensor has a few watts of power budget — not an H100's 700W TDP. Signing, hashing, and remote attestation must fit within that power and latency envelope without degrading real-time mission response.

OWASP Guidance (LLM03:2025)

> Encrypt models deployed at the edge with integrity controls and use vendor attestation APIs to prevent tampered apps and models, and to terminate applications with unrecognized firmware.

Strategic point: edge attestation is where sovereign and defense technology suppliers will build their decisive competitive moat — not in data center efficiency, but in **trustworthy AI at the constrained edge**.

Regulatory Horizon

Convergence: From Engineering Option to Compliance Vocabulary

What was an architectural choice in 2024 becomes the **common vocabulary of at least four European regulators** by 2026–2027.

AI Act — Articles 11, 15 & Annex IV

Article 15 mandates cybersecurity controls against data poisoning, model poisoning, adversarial inputs, and confidentiality attacks. Annex IV lists exactly what a signed AIBOM materializes. Sanctions: **up to €15M or 3% of global revenue**. Enforceable August 2026 (or December 2027 under omnibus).

Cyber Resilience Act (CRA)

Imposes secure-by-design on all products with digital elements on the EU market. AI embedded in critical systems is in scope. Essential requirements include integrity protection and SBOM/AIBOM provision to authorities upon incident.

NIS2 & DORA

Operational resilience for essential service operators (NIS2) and financial entities (DORA) now explicitly includes the AI supply chain. Which models run on which systems, with which dependencies, becomes an incident-reporting obligation.

ISO/IEC 42001:2023

AI management system standard already adopted by CISOs and DPOs. AIBOM is its natural operational translation — the machine-readable artifact that fulfills its documentation and traceability requirements in practice.

Honest Assessment

Limits, Criticality Matrix & the Sovereign Moat

Four Honest Limits

→ Cost

Realistic **2× to 4× inference TCO** for full sealing — weighed against AI Act sanctions and incident liability.

→ Operational Complexity

PKI, validation services, revocation, and key rotation at fleet scale is a non-trivial trust anchor problem.

→ TEEs Are Not Infallible

25.3% of TEE-using projects have coding practices that void the hardware guarantee. **BadRAM (2024)** and **TEE.fail (2025, under $1,000)** forged valid attestations on AMD SEV-SNP, Intel TDX, and NVIDIA on up-to-date hardware.

→ Ecosystem Maturity

Only **19% of organizations** have full visibility on their AI usage (Cycode, 2026). You cannot seal what you have not inventoried.

Criticality Matrix — Seal What Must Be Sealed

Tier	Use Case	Stack	Risk Covered
Low	Internal chatbot, non-sensitive summary	SafeTensors + Sigstore + basic monitoring	~80% risk for 5–10% of cost
Medium	Enterprise RAG, regulated copilot	CycloneDX ML-BOM + SLSA Build L2 + NIS2/DORA logging	Compliance-ready
High	AI Act high-risk, NIS2 essential operator	Confidential GPU + composite attestation + signed in-toto AIBOM	Audit-defensible
Extreme	Defense, national security, strategic IP	Async zk proofs + edge attestation + sovereign qualification	Adversary-resistant

Honest Roadmap

A Direction, Not a Destination

The Realistic Industrialization Roadmap

Don't Confuse the Compass with the Finish Line

AI Under Seal as described here — full chain, composite attestation, cryptographic proofs, attested edge — is a target, not a near-term achievable state for most organizations. The building blocks exist. Their industrialization will face very real friction for months, possibly years:

Hardware ruptures

BadRAM, TEE.fail — valid attestations forged on up-to-date AMD SEV-SNP, Intel TDX, and NVIDIA hardware for under $1,000.

Tooling immaturity

AIBOM tooling remains partially production-ready; CycloneDX ML-BOM adoption is still nascent.

PKI at scale

Key rotation, revocation, and trust anchor management at fleet scale is a non-trivial operational burden.

Skills gap

Engineers who master both MLOps and applied cryptography are rare. Most ML pipelines were never designed to produce proofs.

Legacy debt

Existing ML pipelines carry years of technical decisions incompatible with attestation-by-design.

The Realistic Trajectory: A Staircase, Not a Big Bang

Model signing first

SafeTensors + Sigstore. Immediate risk reduction, low cost.

AIBOM next

CycloneDX ML-BOM + in-toto. Supply chain visibility before compliance deadlines.

Confidential computing

on the most sensitive workloads only. Prioritize by criticality tier.

Composite attestation

for the most demanding certifications (AI Act high-risk, NIS2 essential operators).

Attested edge

last, for field-deployed systems where the data center model breaks down.

Each step reduces an attack surface and a regulatory risk — without waiting for the full stack to be in place.

> The right question is not 'are we there yet?' — it is 'do we know where we are going, why we are going there, and can we take the first steps today?' The gap between those who will gain advantage and those who won't will be decided less by sophistication at the top layers than by the consistency with which they climb the first ones.

Closing Thesis

The European Sovereign Moat

The competitive advantage of the next generation of critical-AI suppliers will not be measured by raw model performance. It will be measured by the ability to present, on demand, **a verifiable end-to-end chain of attestation** — from training data to live inference.

The building blocks already exist and are production-ready:

SafeTensors + Sigstore

in-toto + CycloneDX ML-BOM

NVIDIA CC + Intel TDX + AMD SEV-SNP

zkLLM + ARM CCA

The work ahead is no longer to invent these tools. It is to industrialize them.

Those who transform this requirement into a **distinctive industrial competence** will hold, by 2027–2028, a defensive moat few non-European competitors can replicate quickly. Regulatory convergence — AI Act, CRA, NIS2, DORA — is not a burden. It is a **structural barrier to entry** that rewards early movers.

> Attestation is not the cost of compliance. It is the architecture of trust — and the foundation of the next competitive era in critical AI.