The Model State Attestation Framework

Evidence-Based Governance for Foundation Models

Ryan Fields

UncleBroFields@proton.me

fieldsryanchristopher@gmail.com

February 2026

Abstract

The deployment of foundation models into critical infrastructure—creditworthiness assessment, diagnostic pathways, public resource allocation—has created a fundamental governance paradox: organizations cannot definitively demonstrate that a specific output was generated by a model version in a compliant internal state. Current governance artifacts (model cards, training logs, static benchmarks) are descriptive rather than prescriptive and fundamentally non-verifiable. This paper proposes the Model State Attestation Framework (MSAF), a layered architecture that binds internal model invariants to cryptographic proofs rooted in hardware trust primitives. The framework composes three verification tiers: TEE-backed platform attestation for execution environment integrity, signed continuous metrics for training and inference health monitoring, and selective zero-knowledge proofs for high-value inference verification. A survey of 200+ papers (2022–2026) confirms that while each component exists individually, no system yet composes them into a unified attestation artifact for foundation-scale models. We map the composed artifact to specific evidence requirements across five regulatory frameworks (EU AI Act, FDA SaMD, SR 11-7, US Federal Procurement, AI Insurance) and identify the theoretical ceilings—Rice's theorem, the Impossibility Sandwich, TEE physical side-channel vulnerabilities—that bound what attestation can and cannot guarantee. The honest framing: MSAF provides probabilistic risk reduction and accountability infrastructure, not behavioral safety guarantees. This is analogous to financial auditing, which certifies process compliance without guaranteeing future solvency. The building blocks exist, the regulatory pull is strong, and the standards window is open.

Document Classification: Academic Review Version

This document presents the theoretical architecture and regulatory mapping of the MSAF. Composition protocols, cryptographic specifications, and implementation details are reserved.

Contents

Part I: The Problem

1 The Governance Paradox

Foundation models now occupy positions of consequential authority. They adjudicate creditworthiness for millions of loan applicants, recommend diagnostic pathways in clinical settings, allocate public resources across competing demands, and execute autonomous trading strategies in financial markets. Yet these systems operate as black boxes whose internal integrity is *assumed* rather than *proven*.

The current governance toolkit—model cards, voluntary training logs, static evaluation benchmarks, red-teaming reports—describes what a model *was* at a point in time. It does not prove what the model *is* at the moment of decision. An organization deploying a foundation model in a regulated sector faces a tripartite evidentiary gap:

1. **Identity Gap.** No cryptographic proof binds a specific inference output to a specific model version with specific weights.

2. **Health Gap.** No verifiable certificate attests that the model's internal dynamics—gradient stability, entropy distribution, representational coherence—remained within safe operating parameters during execution.

3. **Provenance Gap.** No tamper-evident record proves that the training data, fine-tuning procedures, and deployment configuration match what was documented.

This situation is analogous to pre-Sarbanes-Oxley financial reporting, where organizations self-attested to the accuracy of their financial statements without independent verification of the underlying processes. The passage of SOX in 2002 did not guarantee that no company would ever commit fraud; it created an accountability infrastructure that made fraud detectable, attributable, and legally consequential. The AI governance field requires an equivalent transition: from trust-based to evidence-based deployment.

1.1 The Inadequacy of Descriptive Governance

Model cards, introduced by Mitchell et al. (2019), were a necessary first step toward transparency. However, they are static documents that become stale the moment a model is updated, fine-tuned, or deployed in a new context. They describe intended use but cannot verify actual use. They report aggregate performance metrics but cannot attest to per-inference health. They are, by design, self-reported and unverifiable.

The same limitation applies to training logs, evaluation benchmarks, and red-teaming reports. Each is a *claim* about the model's properties at a specific point in time. None is a *proof* that those properties held at the moment of a specific decision. In any domain where decisions carry legal, financial, or medical consequences, the distinction between a claim and a proof is the distinction between trust and evidence.

1.2 The Evidence-Based Alternative

The Model State Attestation Framework proposes a different paradigm: every consequential AI decision should be accompanied by a cryptographic attestation artifact that binds three elements into a single tamper-evident digest:

1. The **identity** of the model (a hash of the weights and configuration),

2. The **health** of the model (signed measurements of internal invariants at the time of inference), and

3. The **provenance** of the model (a hash-chained record linking the current state to its training data and development history).

This artifact serves as a "decision receipt"—a tamper-evident record that can be presented in legal proceedings, regulatory audits, insurance claims, or procurement evaluations. Its existence converts AI governance from a documentation exercise into an engineering discipline.

2 Internal Invariants as Health Certificates

The first requirement of the MSAF is the identification of *internal invariants*—measurable properties of the model's weights, activations, and operational environment that must hold true to certify its "health." These invariants serve as a multidimensional certificate of integrity, analogous to vital signs in clinical medicine: no single metric is sufficient, but the combination provides a reliable diagnostic signal.

The framework identifies five invariant classes, each grounded in established mathematical theory and supported by recent empirical research.

2.1 Structural Coherence via Dirichlet Energy

Research has identified a pervasive "representation-use gap" in large language models: models learn accurate internal maps of context but fail to deploy them for downstream tasks, leading to a "fossilized" state where internal structure crystallizes without remaining actionable. Dirichlet Energy—the smoothness of representations over a graph topology—provides a mathematical measure of this phenomenon.

Definition 2.1 (Structural Coherence Invariant)**.** *Let $G = (V, E)$ be the graph induced by the model's representational topology, and let $f : V \to \mathbb{R}^d$ be the learned representation function. The Dirichlet Energy is:*

$$\mathcal{E}_D(f) = \sum_{(i,j)\in E} \|f(v_i) - f(v_j)\|^2 \tag{1}$$

The structural coherence invariant requires that the eigenvalues λ_k of the associated graph Laplacian satisfy:

$$\boxed{0.8 \leq \lambda_k \leq 1.2 \quad \forall\, k \in \{1, \ldots, K\}} \tag{2}$$

where K is the number of monitored eigenvalues (typically the top 10–20 principal components).

A system operating below the lower bound ($\lambda_k < 0.8$) has entered a fossilized regime where representations are structurally coherent but functionally inert. A system exceeding the upper bound ($\lambda_k > 1.2$) is drifting toward chaotic fragmentation where representations lose mutual consistency. Either condition constitutes an attestation fault.

2.2 Distributional Health via Entropy Floor

As models are iteratively trained on their own outputs—a practice increasingly common in RLHF pipelines and self-play systems—they risk "model collapse": a rapid decline in output diversity and the erosion of the distributional tails that encode rare but critical phenomena. In medical generative AI, self-referential training causes the "vanishing" of life-threatening findings like effusions or pneumothorax from synthetic documentation, even as the model maintains fluency and false diagnostic confidence.

The MSAF defines an entropy floor using the Entropy-Reservoir Bregman Projection (ERBP) framework, which models the training loop as a sequence of stochastic projections in information-geometric space.

Definition 2.2 (Entropy Floor Invariant)**.** *Let p_θ denote the model's output distribution at parameters θ, and let $H(p_\theta)$ denote its Shannon entropy. The ERBP framework identifies a necessary condition for distributional stability: the existence of an external high-entropy reservoir $\mathcal{R}$ maintaining controllable entropy flux Φ_H. The invariant requires:*

$$\boxed{H(p_\theta(t)) \geq H_{\min} \quad \forall\, t \geq 0} \tag{3}$$

where $H_{\min}$ is calibrated to preserve the diagnostic tails of the target distribution.

The attestation artifact includes a signed measurement proving that entropy remained above the floor throughout the attested period. A breach triggers an audit fault and initiates a re-training protocol drawing from the verified high-entropy reservoir.

2.3 Gradient Integrity via Stability Monitoring

Gradient health is the most production-ready invariant class. Global L_2 norm clipping (typically at 1.0) is nearly ubiquitous in LLM training, but clipping alone is not attestable—it is a local intervention, not a verifiable certificate.

Recent work has formalized gradient stability via spectral norms of sub-layer Jacobian matrices, proving two sufficient conditions: small sub-layer parameters and large shortcut standard deviations (≈ 1.0). The AGGC framework (January 2026) partitions parameters into functional groups and regulates each via exponential moving average, simultaneously mitigating explosion and vanishing. The SPAM detector identifies gradient spikes reaching magnitudes far beyond typical values and resets momentum.

Definition 2.3 (Gradient Integrity Invariant)**.** *Let $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ denote the gradient at step t, and let $\bar{g}_t$ denote its exponential moving average. The gradient integrity invariant requires:*

$$\boxed{\frac{\|g_t\|_2}{\bar{g}_t + \epsilon} \leq \kappa_{\max} \quad \forall\, t \in [0, T]} \tag{4}$$

where $\kappa_{\max}$ is the maximum permitted spike ratio (empirically, $\kappa_{\max} \approx 10$).

A gradient norm exceeding 10× its EMA baseline is a well-established anomaly signal. The attestation artifact records per-step ratios and flags any breach as an audit event.

2.4 Stability Envelope via Lyapunov Functions

For models deployed in closed-loop systems—autonomous infrastructure, automated trading, robotics—proving the stability of the neural network feedback loop is a core safety requirement. The MSAF incorporates Lyapunov-based methods that construct invariant sublevel sets of a quadratic function, enabling estimation of the Region of Attraction (ROA).

Definition 2.4 (Lyapunov Stability Invariant)**.** *Let $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a candidate Lyapunov function for the closed-loop system $\dot{x} = f(x, \pi_\theta(x))$, where π_θ is the neural network controller. The stability invariant requires:*

$$\boxed{\dot{V}(x) \leq -\alpha V(x) \quad \forall\, x \in \Omega_c} \tag{5}$$

where $\alpha > 0$ is the exponential decay rate and $\Omega_c = \{x : V(x) \leq c\}$ is the certified ROA.

Novel sector bounds propagated layer-by-layer through the network allow these certificates to be computed without the conservatism of global Lipschitz bounds. The attestation includes the ROA volume and the decay rate, proving that controller outputs will not lead to divergent behavior within the defined operating envelope.

2.5 Representational Drift via KL Divergence

In nonstationary deployment environments, the data distribution shifts over time. A model trained on one distribution and deployed against another will silently degrade. The MSAF monitors this drift using KL divergence modeled through Fokker-Planck probability flow.

Definition 2.5 (Drift Monitoring Invariant)**.** *Let p_0 denote the reference distribution at validation time, and let p_t denote the observed input distribution at time t. The drift invariant requires:*

$$\boxed{D_{KL}(p_t \,\|\, p_0) \leq D_{\max} \quad \forall\, t \geq 0} \tag{6}$$

where $D_{\max}$ is calibrated to the model's empirically measured robustness envelope.

When drift exceeds the threshold, the attestation chain records the breach and the model's outputs carry a reduced-confidence flag until revalidation is completed.

2.6 Invariant Summary

Internal Invariant	**Metric**	**Mathematical Foundation**	**Regulatory Relevance**
Structural Coherence	Dirichlet Energy	Spectral Graph Theory	Detects fossilized states and loss of actionability
Distributional Health	Entropy Floor	Bregman Projection	Prevents model collapse and diagnostic tail erosion
Gradient Integrity	Spike Ratio κ	EMA-normalized L_2 Norm	Proves training stability; detects corruption and poisoning
Stability Envelope	Lyapunov Function V	Nonlinear Control Theory	Certifies local exponential stability in closed-loop systems
Representational Drift	KL Divergence	Fokker-Planck Flow	Monitors distribution shift in nonstationary environments

Table 1: The five invariant classes of the Model State Attestation Framework.

The specific monitoring intervals per invariant class, computational overhead budgets, minimum detection probabilities, multi-invariant correlation analysis protocols, and the connection between these invariants and per-clause enforcement triggers are specified in the private version of the framework.

Part II: The Verification Stack

3 Hardware-Rooted Trust: TEEs as Platform Integrity Layer

A Model State Attestation Framework is only as secure as the hardware upon which it executes. The invariants defined in §2 are meaningless if the code measuring them can be tampered with, if the measurements can be forged, or if the execution environment itself is compromised. The primary mechanism for binding internal state measurements to cryptographic proofs is the use of Trusted Execution Environments (TEEs) and hardware roots of trust (RoT) found in modern enterprise accelerators.

This section traces the architectural evolution from the Hopper generation's performance-limiting bounce buffers through Blackwell's inline encryption resolution, evaluates composite CPU–GPU attestation, and identifies the operational realities that constrain deployment.

3.1 The Measured Boot Sequence

The trust chain begins at power-on. During the measured boot sequence, the GPU verifies its firmware against digital signatures burned into on-die fuses and locks the firmware to prevent runtime tampering. Key components of the system state are measured and stored in Runtime Measurement Registers (RTMRs) or Platform Configuration Registers (PCRs). These measurements form the basis of a *quote*—a cryptographically signed report of the system's identity and configuration.

In the MSAF, this quote is extended to include the hash of the model weights and the specific invariant measurements computed during inference or training. The quote thus becomes a composite certificate: it proves not only that the platform is genuine, but that the model running on it is the model that was validated, and that the health metrics recorded during execution were computed by trusted code in a verified environment.

3.2 The Hopper Generation: Capability and Constraint

NVIDIA's H100/H200 GPUs introduced the first production-grade confidential computing architecture for accelerators. Each H100 contains an on-die hardware root of trust with a device-unique private identity key burned into fuses. In Confidential Computing mode (CC-On), the GPU generates cryptographically signed attestation reports covering firmware integrity, device identity, and VBIOS configuration, verifiable against NVIDIA's Certificate Authority via the Remote Attestation Service (NRAS).

The H100 establishes an encrypted channel with the CPU TEE using the Security Protocol and Data Model (SPDM). All traffic between the CPU and GPU traverses encrypted bounce buffers in shared memory, utilizing AES-GCM encryption with rotating initialization vectors to prevent interception or replay.

3.2.1 The Bounce Buffer Bottleneck

The bounce buffer architecture introduces a multi-step data path: the CPU reads data from the confidential VM's encrypted memory, decrypts it, copies it into an unencrypted shared region, where it is re-encrypted for GPU consumption via SPDM. This "double-copy" operation serializes what should be parallel DMA transfers, creating a throughput ceiling of approximately 4 GB/s on the CPU–GPU interconnect.

For single-GPU inference, this bottleneck is modest. Independent benchmarks on H100 NVL hardware (Zhu et al., September 2024) measured the following throughput overheads:

Model	Throughput Overhead	TTFT Overhead	Analysis
Llama-3.1-8B	6.85%	∼19%	I/O-bound; bounce buffer visible
Phi-3-14B	4.58%	Moderate	Mixed compute/I/O
Llama-3.1-70B (4-bit)	∼0%	∼0%	Fully compute-bound; overhead masked

Table 2: Single-GPU inference overhead on NVIDIA H100 in Confidential Computing mode.

The pattern is clear: larger models that spend proportionally more time on GPU-resident computation experience negligible overhead, because the bounce buffer latency is masked by Tensor Core utilization. For 70B-class models, confidential inference is effectively free.

However, multi-GPU distributed training tells a different story entirely. Lee and Wang (January 2025) measured that distributed data-parallel training with four GPU TEEs averages **8× slower**, with worst-case scenarios reaching **41.6× slower**. The root cause is that every gradient synchronization step in ring all-reduce requires $4\times(n-1)$ encryption and authentication operations across GPUs, all flowing through software bounce buffers. Even after aggressive tuning, GPT-2-XL with four GPUs remained 3× slower than native execution. Multi-tenant model swapping scenarios show 45–70% throughput loss and 20–30% latency increase due to encrypted model loading and unloading.

This performance profile defines the Hopper-era constraint: *confidential inference is viable; confidential distributed training is not.*

3.3 The Blackwell Resolution: Inline Encryption via TDISP/IDE

The NVIDIA Blackwell architecture (B200/B300, fully available 2025–2026) resolves the bounce buffer bottleneck through native hardware support for the Trusted Device Interface Security Protocol (TDISP) and Integrity and Data Encryption (IDE), developed in conjunction with PCI-SIG.

3.3.1 The Zero-Copy Architecture

Blackwell GPUs feature DMA engines capable of performing AES-GCM encryption and decryption at full PCIe Gen6 and NVLink line rates. This enables a fundamentally different data path:

1. The CPU TEE and GPU TEE negotiate a session key via TDISP.
2. The IDE engine at the CPU Root Port encrypts data as it enters the PCIe fabric—at wire speed.
3. The GPU's internal IDE engine decrypts data directly into HBM3e or L2 cache, bypassing any bounce buffer entirely.

The "privacy tax" reduces to the nanosecond-scale latency of pipelined AES-GCM logic gates. NVIDIA's Unified Virtual Memory (UVM) extension handles the TDISP state machine transparently, meaning developers use standard `cudaMallocManaged` APIs without modification—a "lift and shift" capability critical for adoption.

3.3.2 Estimated Blackwell Overhead

Performance data synthesized from MLPerf Training v5.1 results and Blackwell architecture documentation indicates that compute-bound workloads now incur minimal overhead:

Workload	Model	Overhead	Bottleneck
LLM Pre-training	Llama 3.1 405B	∼1.0%	Compute-bound; encryption hidden by GEMM latency
LLM Fine-tuning	Llama 2 70B LoRA	∼2.5%	Mixed; slight checkpoint overhead
Computer Vision	RetinaNet	∼4.3%	Data-bound; image loading triggers IDE latency
Recommender	DLRM-DCNv2	∼5.6%	Memory-bound; random access stresses encryption

Table 3: Estimated Confidential Computing overhead on NVIDIA Blackwell (GB200 NVL72).

Critically, the Vera Rubin NVL72 rack extends the trust domain to 72 Blackwell GPUs and 36 CPUs connected via an encrypted NVLink fabric at 1.8 TB/s. Model parallelism—sharding a model like Llama 3.1 405B across dozens of GPUs—occurs entirely within the encrypted domain. The heavy all-to-all communication required for attention heads and mixture-of-experts routing happens securely without exiting the TEE to the network layer.

Important caveat: As of February 2026, no independent third-party benchmarks with specific percentages have been published for Blackwell confidential computing. The overhead figures above are synthesized from NVIDIA documentation and MLPerf results; independent verification is pending. No MLPerf submission has ever been run in confidential computing mode.

3.4 Intel TDX Composite Attestation

While the GPU performs inference and training, the CPU orchestrates the execution environment. Intel Trust Domain Extensions (TDX) provides the host-side TEE, creating hardware-isolated virtual machines (Trust Domains) by removing the hypervisor from the Trusted Computing Base.

3.4.1 The Split-Verifier Architecture

Composite attestation—jointly verifying CPU and GPU trust domains—operates through a split-verifier model:

1. The workload inside the Trust Domain collects a TDX DCAP quote (signed by the SGX-based TD Quoting Enclave) and an NVIDIA GPU attestation report (obtained via SPDM).
2. Both are submitted to Intel Trust Authority (ITA).
3. ITA verifies the TDX quote directly and forwards the GPU evidence to NVIDIA's Remote Attestation Service (NRAS).
4. NRAS checks the GPU report against its Reference Integrity Manifest (golden measurements).

5. ITA combines both results into a composite JWT token containing verified claims about CPU state (MRTD, RTMRs, module versions) and GPU state (driver version, firmware integrity).

6. Policies are expressed in Rego (OPA) syntax, enabling fine-grained composite verification rules.

3.4.2 Operational Fragility

This architecture introduces significant operational risks. Verification latency is the sum of network round-trip times to ITA *plus* ITA-to-NRAS. Intel documentation explicitly warns that "lengthy service delays" from NRAS can cause verifier nonces to expire, requiring full attestation restart. A cloud outage at either Intel or NVIDIA prevents the launch of new confidential workloads globally—a "dual-dependency" that enterprise architects have flagged as unacceptable for mission-critical deployments.

Additionally, TDX attestation fundamentally depends on SGX infrastructure: the TD Quoting Enclave is an SGX enclave, and the entire certificate chain flows through SGX's Provisioning Certification Enclave. Any vulnerability compromising SGX attestation keys cascades to break TDX attestation and, by extension, the composite GPU attestation chain.

For I/O-intensive workloads, TDX overhead is non-trivial: academic benchmarks measure 28.6% average overhead for I/O-heavy workloads, and network throughput can drop up to 60% under heavy TCP traffic—a concern for ML data loading and distributed training communication.

3.5 Hardware Primitives Summary

Primitive	Technology	Role in MSAF Attestation
RTMR / PCR	TPM / Intel TDX	Stores cumulative hashes of boot and runtime states
SPDM Session	NVIDIA CC / Hopper+	Facilitates secure key exchange between CPU and GPU TEEs
TDISP / IDE	Blackwell / PCIe Gen6	Inline encryption eliminating bounce buffer overhead
HPC / PMU	Intel / ARM / RISC-V	Collects hardware performance counters to detect anomalies
FIPS 140-2 HSM	Cloud / Hybrid HSM	Provably releases keys only to attested workload identities
NVLink Fabric	Vera Rubin NVL72	Rack-scale encrypted communication for model parallelism

Table 4: Hardware primitives leveraged by the MSAF for tamper-proof measurement.

3.6 Cloud Provider Landscape

Three cloud providers offer confidential GPU computing as of early 2026, with dramatically different capabilities:

Azure NCC H100 v5 (GA September 2024) pairs AMD SEV-SNP with NVIDIA H100. Production customers include OpenAI (Whisper inference), Royal Bank of Canada, ServiceNow, and Bosch. Pricing is approximately $7.16/hour in limited regions.

Google Cloud A3 instances combine Intel TDX with NVIDIA H100, available in three zones. Google emphasizes "no code changes needed for most AI and ML workloads." Xiaomi deployed this for its HyperOS 2 Private Cloud Compute feature.

AWS Nitro Enclaves remain CPU-only with no GPU support, communicating only via vsock—fundamentally unsuitable for GPU-accelerated ML.

A critical gap across all providers: only single-GPU configurations are currently available. No cloud provider offers multi-GPU confidential computing, leaving large-model distributed training outside the reach of confidential cloud offerings. CVM boot times exceed 2× standard VM boot times, and initial memory allocation shows up to 92% overhead from CVM-specific operations—significant when loading large models.

4 Cryptographic Proofs: zkML as Computational Assurance

While hardware TEEs provide *environment-level* assurance (proving that code ran on genuine, unmodified hardware), Zero-Knowledge Machine Learning (zkML) provides *computational-level* assurance: a cryptographic proof that a specific inference was computed correctly according to a committed model, without revealing the model's weights or the user's private inputs.

The zkML landscape has undergone a phase transition in the eighteen months preceding February 2026. Proof-generation overhead has collapsed from approximately $10^6\times$ native computation in 2022 to roughly $10^4\times$ by late 2025, and continues to shrink. Multiple systems now verify inference of billion-parameter models in minutes. Training verification, however, remains 4–5 orders of magnitude behind—establishing the sharpest boundary in the current feasibility landscape.

4.1 The Dominant Paradigm: GKR + Sumcheck + Lookup Arguments

The single most important architectural shift in zkML has been the displacement of Halo2/PLONKish proof systems by the GKR (Goldwasser-Kalai-Rothblum) protocol combined with sumcheck arguments and lookup tables. This trifecta exploits a structural insight: neural networks are layered arithmetic circuits, and the GKR protocol was designed precisely for layered computation.

GKR achieves $O(n)$ prover complexity versus $O(n \cdot \log n)$ for Groth16/PLONKish systems, and avoids intermediate layer commitments. The **lookup argument** innovation—formalized by the tlookup protocol—addresses the true bottleneck in zkML: non-linear operations (GELU, SwiGLU, Softmax, LayerNorm). By mapping quantized activation function inputs and outputs to precomputed tables, lookup arguments eliminate the need for expensive arithmetic circuit representations of non-linearities, reducing their constraint footprint by orders of magnitude.

4.2 The Inference Verification Frontier

Three systems define the current frontier for inference verification, each occupying a distinct niche:

4.2.1 zkLLM: The Billion-Parameter Blueprint

zkLLM (Sun, Li, and Zhang; University of Waterloo; ACM CCS 2024) remains the foundational contribution for large language model verification. It demonstrated that **13-billion-parameter models (LLaMa-2-13B) can be verified in 1–15 minutes** on a single NVIDIA RTX A6000 GPU, producing proofs under 200 kB with verification in 1–3 seconds. This represented a 50× speedup and 10× model-size increase over prior generic approaches.

zkLLM introduced two widely adopted innovations: **tlookup** (parallelized tensor lookup argument for non-linear operations) and **zkAttn** (a dedicated protocol for attention mechanisms

that decomposes Softmax verification through shift-invariance and base-b digit factorization). Built on BLS12-381 elliptic curve operations and Pedersen commitments, zkLLM provides genuine zero-knowledge privacy—model weights remain hidden from the verifier.

4.2.2 DeepProve: From Research to Production

DeepProve (Lagrange Labs, $21.5M from Founders Fund and 1kx) is the most production-deployed zkML system, having generated over **11 million zero-knowledge proofs** with more than 3 million AI inferences verified. It uses a GKR-style sumcheck protocol paired with LogUp GKR lookup arguments, ingesting ONNX and GGUF model files directly.

DeepProve achieves **54–158× faster proof generation** and up to **671× faster verification** compared to EZKL. It supports dense layers, convolutions, multi-head and group query attention, RMSNorm, LayerNorm, Softmax, and multiple activation functions. Critically, DeepProve achieves *sublinear* proving time scaling: when model size grows n-fold, verification time grows slower than linearly.

A significant distinction: DeepProve does *not* provide true zero-knowledge privacy. Proofs are succinctly verifiable, but model weights are not hidden. Lagrange positions this as verifiable inference rather than private inference.

4.2.3 EZKL: The Developer Ecosystem

EZKL (Zkonduit Inc.) occupies the broadest developer-accessibility niche. Built on a modified Halo2 proof system with KZG polynomial commitments, it converts any ONNX-format model into a ZK-SNARK circuit. Version 23.0.3 (October 2025) is the **only audited zkML framework** (Trail of Bits). EZKL supports CNNs, transformers, RNNs/LSTMs, decision trees, random forests, XGBoost, SVMs, and linear models—the widest architecture coverage of any framework.

The primary constraint is scale: EZKL's Halo2-based approach struggles above approximately 30 million parameters due to memory overhead. For large transformer models, the specialized GKR-based systems have decisively overtaken EZKL's throughput.

4.3 The Feasibility Matrix

Scale	Model	System	Proof Time	Hardware	Privacy
∼250K	nanoGPT	EZKL	237s	CPU	Yes
∼117M	GPT-2	zkGPT	<25s	CPU	Partial
∼6B	GPT-J	ZKTorch	∼20 min	64 CPU	Yes
∼8B	Llama-3	zkPyTorch	∼150s/tok	CPU	No
∼13B	LLaMa-2-13B	zkLLM	<15 min	A6000 GPU	Yes

Table 5: zkML proof times across model scales (inference verification).

The **360× gap** between generic approaches (>90 hours for 1.5B parameters via Modulus Labs) and specialized protocols (<15 minutes for 13B parameters via zkLLM) demonstrates that architectural specialization—not raw compute—determines feasibility.

4.4 The Training Verification Wall

The sharpest boundary in zkML sits between inference and training verification. While inference proofs reach 13 billion parameters, full-training proofs top out at roughly **10 million parameters**—a gap of three orders of magnitude.

The state of the art for full gradient-descent verification is Kaizen (CCS 2024): VGG-11 (10M parameters, 11 layers) with proof time of **15 minutes per gradient step**, proof size of 1.63 MB, and verification in 130 milliseconds. Since a typical training run involves tens of thousands of steps, each approximately 100× more expensive than inference, the total proving cost for training even a 10M-parameter model measures in months of continuous computation.

LoRA fine-tuning has emerged as the practical bridge. VeriLoRA (August 2025) demonstrated end-to-end ZK-verifiable LoRA fine-tuning of LLaMA-2 up to **13 billion parameters** using GPU acceleration, covering forward propagation, backward propagation, and parameter updates. ZKLoRA (ICML 2025 Workshop) verifies LoRA module compatibility in 1–2 seconds per module. This low-rank workaround is currently the only path to verifiable fine-tuning at frontier scale.

Full pre-training of modern LLMs in zero-knowledge remains **completely infeasible**—requiring what would amount to years of proof-generation compute at current overhead levels.

4.5 Hardware Acceleration and the Overhead Trajectory

The proof-generation overhead trajectory is driven by three converging hardware developments:

GPU-native proving dominates current deployment. Ingonyama's ICICLE library delivers 8–10× MSM speedups and 3–5× NTT speedups across multiple proof systems. A critical finding (ZKProphet, IEEE IISWC 2025): after years of MSM optimization, **NTT now accounts for up to 90% of proof generation latency**—the optimization bottleneck has shifted.

Dedicated ZK-ASICs are shipping. Cysic's ZK Air claims 100× faster than software; Fabric Cryptography's VPU chip achieves 900% more big-integer operations than a GPU; Accseal's Leo ASIC claims 80× lower cost than GPU. Academic designs (UniZK at ASPLOS 2025: 46× faster than GPU; NoCap at MICRO 2024: 586× over 32-core CPU) suggest purpose-built hardware could reduce the 10,000× overhead to 100× or less.

Streaming proofs on NVIDIA's Rubin architecture enable token-by-token proof generation for LLM sequences, reducing memory complexity from $O(N)$ to $O(1)$. This is critical for verifying chain-of-thought reasoning in "thinking" models—the proof verifies that the logical steps $A \to B \to C$ were actually executed, preventing "reasoning hallucination" where a model outputs a correct answer for incorrect reasons.

A milestone result: Succinct's SP1 Hypercube proved 99.7% of Ethereum blocks in under 12 seconds using only 16 NVIDIA RTX 5090 GPUs (November 2025), down from approximately 200 RTX 4090s six months earlier—a 12× hardware efficiency gain in six months.

4.6 Implications for the MSAF

The zkML feasibility frontier defines what the MSAF can cryptographically prove at each tier:

- **Inference verification** for models up to 13B parameters is practical today for asynchronous audit (not real-time interactive use). This covers the majority of deployed enterprise models.
- **Fine-tuning verification** via LoRA is practical at 13B parameters, providing cryptographic proof that a specific fine-tuning procedure was applied to a specific base model with specific data.

- **Pre-training verification** remains infeasible at frontier scale. The MSAF addresses this gap through TEE-backed provenance chains (Tier 1 and Tier 2 attestation) rather than computational proofs.
- **Frontier models** (>100B parameters) require hybrid approaches: OpML (optimistic verification with fraud proofs) or selective layer verification, where ZK proofs cover only the "dense" decision-critical layers.

The specific integration architecture connecting zkML proof generation to the MSAF's Tier 3 attestation artifacts—including proof scheduling policies, selective verification criteria, the OpML dispute resolution protocol, and the cryptographic binding between zkML receipts and TEE-rooted attestation chains—is specified in the private version of the framework.

5 Data Provenance and Contamination Control

Verifying the internal state of a model is meaningless without verifying the data used to train or refine that state. A model with perfect gradient health and stable entropy can still produce catastrophically biased, legally non-compliant, or medically dangerous outputs if its training data was contaminated, unrepresentative, or improperly sourced. The MSAF treats data provenance as a first-class attestation requirement, not an afterthought documented in a model card.

5.1 Cryptographic Provenance via Multiset Hashing

Large-scale training runs sample data randomly from terabyte-scale corpora. Traditional sequential hashing is incompatible with this access pattern: it requires processing the dataset in a fixed order, which is neither how training works nor how datasets are stored in distributed systems.

The MSAF utilizes incremental multiset hashing over memory-mapped datasets. Multiset hashing is commutative—data pages can be hashed in any order as they are accessed, decoupling measurement time from dataset size. This allows the framework to produce a cryptographic proof of provenance even when data is sampled stochastically during training. The resulting hash serves as a fingerprint: any modification to the training corpus (addition, deletion, or alteration of a single document) changes the hash, making unauthorized data manipulation detectable.

This capability directly addresses Article 10 of the EU AI Act, which mandates that high-risk AI training data be "representative, free of errors, and complete." The multiset hash does not prove these properties directly—no hash can verify semantic quality—but it proves that the data used in production matches the data that was audited and approved, closing the gap between what was documented and what was deployed.

5.2 Synthetic Data Contamination

A significant and growing challenge in foundation models is synthetic data contamination: the recursive training of models on data generated by previous iterations of themselves or other AI systems. This phenomenon creates a systematic erosion of content fidelity, particularly for rare but critical "tail" events that appear infrequently in natural data but carry outsized consequences.

The medical domain provides the starkest illustration. Self-referential training in clinical AI causes the "vanishing" of life-threatening findings—effusions, pneumothorax, rare malignancies—from synthetic documentation. The model maintains high fluency and apparent diagnostic confidence while its internal representation of rare pathology degrades toward zero. A clinician relying on such a model receives outputs that *sound* authoritative but have silently lost the ability to flag the conditions most likely to kill the patient.

Contamination detection has matured significantly. Min-K% Prob (ICLR 2024) identifies outlier low-probability tokens in unseen examples, achieving 7.4% improvement over prior methods without requiring knowledge of the pretraining corpus. CDD (ACL 2024) analyzes output distribution patterns, finding approximately 40% contamination in ChatGPT's HumanEval examples. However, a fundamental tradeoff has been proven: strategies that increase contamination resistance distort the benchmark task itself, creating an irreducible tension between detection power and task fidelity.

5.3 The Provenance Ecosystem: Atlas, AICert, AIBoMGen

Three emerging frameworks address ML lifecycle attestation, each with a fundamentally different trust architecture. Their comparison illustrates the design space available to the MSAF.

Atlas (Intel Labs, February 2025) represents the deepest TEE integration. Built as a sidecar container for Kubeflow pipelines, Atlas uses Intel TDX for both runtime integrity enforcement and as a hardware root of trust for provenance metadata. It generates C2PA (Coalition for Content Provenance and Authenticity) manifests containing cryptographic hashes of all ML artifacts—datasets, models, checkpoints—along with TDX attestation evidence and linked transformation records. Training overhead is under 8%. The Rust-based CLI (v0.1.0) is open-source under Apache 2.0, but is explicitly not production-ready.

AICert (Mithril Security, v1.0, September 2024) uses virtual TPMs rather than enclave-style TEEs. Running on Azure confidential VMs, AICert measures the entire boot chain into TPM Platform Configuration Registers, then binds training inputs (PCR 14) and outputs (PCR 8) to a TPM-signed attestation quote. Training executes in an isolated container with no outside network access, preventing model substitution. AICert was motivated by Mithril's PoisonGPT demonstration showing surgical model modification that evades standard benchmarks. Current limitations: supports only fine-tuning (via Axolotl framework), Azure-only, and does not audit training code or data quality.

AIBoMGen (Ghent University–imec, January 2026) deliberately avoids TEEs, treating the training platform as a neutral third-party observer. It generates AI Bills of Materials using in-toto attestations—cryptographically signed records of each pipeline step—with all artifacts hashed and platform-signed. The authors explicitly considered and rejected TEE-first architectures as "tying the approach too tightly to specific hardware assumptions." AIBoMGen achieves 100% detection accuracy for artifact and metadata mutations with approximately 0.38s overhead independent of workload size, using CycloneDX AI Bill of Materials format. Accepted for presentation at CAIN 2026.

Framework	Trust Root	Artifact Format	Key Limitation
Atlas	Intel TDX (hardware)	C2PA manifest	v0.1.0; not production-ready
AICert	Virtual TPM	TPM quote	Fine-tuning only; Azure-only
AIBoMGen	Platform signature	CycloneDX AIBOM	Assumes honest platform

Table 6: Provenance frameworks and their trust architectures.

The critical observation: each framework addresses a *subset* of the provenance problem. Atlas tracks transformations but does not audit data quality. AICert binds code to weights but does not track the full training pipeline. AIBoMGen captures the complete pipeline but trusts the platform. The MSAF's three-tier architecture (§6) is designed to compose these

capabilities—using the strongest available trust root at each layer—into a unified provenance chain.

5.4 Provenance Requirements Summary

Provenance Feature	**Mechanism**	**Mitigation Goal**
Persistent Tagging	Metadata in EHR / C2PA	Prevents uncurated downstream contamination
Multiset Hashing	Commutative cryptographic hash	Ensures large-dataset integrity during random sampling
Membership Queries	Privacy-preserving access	Verifies if specific IP or personal data was used in training
Quality-Aware Filtering	Entropy / similarity metrics	Preserves diagnostic tails and rare pathology
Contamination Bounds	Synthetic content ratio tracking	Maintains minimum proportion of verified human-authored data

Table 7: Provenance features, mechanisms, and mitigation goals.

The specific contamination ratio thresholds, provenance chain cryptographic binding protocols, the integration architecture connecting multiset hashes to TEE-rooted attestation quotes, and the privacy-preserving membership query implementation are specified in the private version of the framework.

Part III: The Composed Artifact

6 The Three-Tier Attestation Architecture

The preceding sections established the individual components: internal invariants (§2), hardware-rooted trust (§3), cryptographic proofs (§4), and data provenance (§5). Each exists in isolation. The central contribution of the Model State Attestation Framework is their *composition* into a unified attestation artifact that no existing system provides.

The composition problem is the framework's core technical challenge. Heterogeneous invariants—continuous-valued gradient norms, statistical contamination scores, binary hardware fault signals, probabilistic zkML receipts—must be unified into a single verifiable artifact that is compact enough for high-throughput logging, rich enough for regulatory audit, and cryptographically bound to a hardware root of trust.

6.1 The Compositional Challenge

Existing work on compositional neural certificates provides the closest theoretical template. Zhang et al. (MIT REALM, L4DC 2023) decomposed networked dynamical systems into subsystems, learned Input-to-State Stability (ISS) Lyapunov functions for each, then composed them into a global stability proof via a small-gain condition. The "Certificates in AI: Learn but Verify" paper (Communications of the ACM, 2025) argues for a Certified Machine Learning principle: every AI output should come with a certificate witnessing correctness, checkable by an independent verifier using a learner-verifier (CEGIS) architecture.

On the cryptographic side, the "Constant-Size Cryptographic Evidence Structures" paper (arXiv 2511.17118) proposes fixed-size tuples per event, composable with hash-chained logs, Merkle-tree anchoring, and TEE binding—with per-event overhead compatible with high-throughput workloads. The IETF RATS (Remote ATtestation procedureS) architecture specifies COSE-formatted attestation tokens containing device claims (header, payload, signature). Google's DICE specification creates hardware Root of Trust measurement chains.

The MSAF draws on these foundations to define a three-tier architecture where each tier operates at a different temporal frequency, trust level, and computational cost.

6.2 Tier 1: Platform Attestation (Hardware Root of Trust)

Tier 1 establishes that the execution environment is genuine and unmodified. It operates at boot time and at platform state changes.

Content: A TEE platform quote comprising the DICE/TPM root of trust measurement, firmware integrity verification, software stack hash, and device identity signed by the manufacturer's certificate authority.

Trust Root: On-die hardware key (NVIDIA GPU fuses, Intel TDX Module, TPM endorsement key). The trust chain is anchored in silicon that cannot be modified by software, firmware, or the machine owner.

Frequency: Generated at boot, regenerated on any platform state change (firmware update, driver change, configuration modification). Typically once per session or deployment.

What it proves: The code measuring the model's health invariants ran on genuine hardware in an unmodified environment. An auditor verifying a Tier 1 quote can confirm that the measurements reported in Tier 2 were computed by trusted code, not fabricated by a compromised host.

What it does not prove: Nothing about the model itself. A genuine platform running a poisoned model produces a valid Tier 1 quote. Platform integrity is necessary but not sufficient.

6.3 Tier 2: Continuous Invariant Metrics (Signed Runtime Measurements)

Tier 2 provides the real-time health signal. It operates continuously during training and inference, at cadences ranging from per-token to per-batch depending on the invariant class.

Content: Signed, timestamped measurements of the five invariant classes defined in §2—gradient norms, loss trajectory, activation entropy, weight checksums, Lyapunov stability estimates, and drift divergence scores. Each measurement is computed within the TEE and cryptographically bound to the Tier 1 platform quote via a hash chain.

Trust Root: Derived from Tier 1. The signing key is held within the TEE and is released only after successful platform attestation. A measurement signed by this key carries an implicit guarantee: "this value was computed by code C running on platform P, where P passed attestation."

Frequency: Continuous. Gradient integrity is monitored per training step. Entropy and drift are monitored per batch or per N inference requests. Structural coherence is monitored periodically (e.g., every 1,000 steps). Each measurement is appended to a hash-chained log, creating a tamper-evident time series.

What it proves: The model's internal dynamics remained within attested bounds during the period covered by the log. An auditor can verify that gradient spikes did not exceed $\kappa_{\max}$, that entropy did not fall below $H_{\min}$, and that drift did not exceed $D_{\max}$—all with cryptographic certainty that the measurements were not fabricated after the fact.

What it does not prove: That the monitored metrics are *sufficient* to characterize the model's behavior. Tier 2 measurements are low-dimensional projections of an astronomically high-dimensional parameter space. This limitation is addressed honestly in §8.

6.4 Tier 3: Periodic Audit Anchors (Deep Verification)

Tier 3 provides the deepest assurance at the highest computational cost. It operates periodically—at model release, after fine-tuning, at regulatory audit intervals, or when triggered by Tier 2 anomalies.

Content: Contamination detection scores, benchmark evaluation results, fairness and bias metrics, and—where feasible—zkML inference proofs or LoRA fine-tuning verification receipts. These results are anchored to the Tier 1/Tier 2 chain via Merkle roots, creating a single hash that commits to the entire audit history.

Trust Root: Combination of Tier 1 platform attestation (proving the audit code ran on genuine hardware), Tier 2 continuity (proving no unattested gap exists between the last audit and the current one), and—for zkML components—the cryptographic soundness of the proof system itself.

Frequency: Periodic. Triggered by model updates, fine-tuning events, regulatory deadlines, Tier 2 threshold breaches, or scheduled audit intervals. A typical cadence might be monthly for routine monitoring, with immediate triggering on any Tier 2 violation.

What it proves: At the deepest level available, the model's outputs are consistent with its documented training, its performance meets validated benchmarks, and—for zkML-attested inferences—a specific output was computed correctly by the committed model.

What it does not prove: That the model will behave safely for all future inputs. Rice's theorem (§8) guarantees this is undecidable.

6.5 Composition: The Unified Artifact

The three tiers compose into a single attestation artifact structured as a Merkle tree:

- The **root hash** commits to the entire attestation state—platform identity, continuous health metrics, and periodic audit results—in a single fixed-size value.

- Any individual component (a specific gradient measurement, a specific benchmark result) can be verified against the root via a Merkle proof without revealing the full tree.
- The root hash is signed by the TEE-held attestation key and timestamped, creating a tamper-evident snapshot of the model's complete governance state at a specific moment.

This structure satisfies two competing requirements: *compactness* (the root hash is fixed-size regardless of how many measurements it commits to) and *selective disclosure* (an auditor can verify a specific claim without accessing the entire attestation history). For high-throughput workloads, the per-event overhead of Merkle-tree updates is compatible with production inference latencies.

The specific composition protocol—including the Merkle tree construction algorithm, inter-tier cryptographic binding mechanisms, the artifact schema (field definitions, encoding format, versioning), hash-chained log structure, the selective disclosure protocol for regulatory auditors, and the integration with existing attestation formats (COSE, C2PA, CycloneDX)—is specified in the private version of the framework.

7 Attesting Dense vs. Mixture-of-Experts Architectures

The MSAF must be compatible with both dense models and the increasingly prevalent Mixture-of-Experts (MoE) architectures. While dense models execute the entire network for every input, MoE models utilize sparse conditional computation, routing each token to a subset of specialized "expert" sub-networks. This sparsity introduces a fundamental attestation challenge: different tokens follow different computational paths, making the model's internal state non-deterministic at the token level.

7.1 The Stochasticity Problem

In a dense model, every input traverses the same computational graph. The attestation artifact can commit to the full set of weights and verify that a specific input-output pair is consistent with those weights. In an MoE model, a gating network (router) selects which experts process each token. If the router's decisions are non-deterministic (e.g., due to load-balancing noise or top-k sampling with ties), then two identical inputs may activate different experts and produce different outputs—both legitimately.

This creates a tension between operational efficiency (stochastic routing improves load balancing and throughput) and attestation requirements (auditors need to verify that a specific output was produced by the attested model). If the routing decision itself is unverifiable, an adversary could claim that any output is "consistent with the model" by asserting an appropriate routing pattern.

7.2 Deterministic Routing for Audit Reproducibility

The MSAF addresses this through a requirement for *deterministic routing* during attested inference. When an inference is subject to attestation (not necessarily every inference—this is a policy decision), the router must operate in a deterministic mode where the expert selection for each token is a pure function of the input and the router weights, with no stochastic load-balancing noise.

This requirement is compatible with modern MoE implementations. Expert parallelism allows the system to predict which experts will be activated for a given input, enabling pre-staging of expert weights and eliminating the non-determinism introduced by dynamic scheduling. The performance cost of deterministic routing is modest for attested inferences, as the primary overhead is the loss of load-balancing flexibility rather than additional computation.

7.3 Fabric-Level Attestation for Distributed Experts

Large MoE models (e.g., the 1.6-trillion-parameter Switch Transformer) distribute experts across multiple GPUs, requiring "all-to-all" communication during the routing phase. Each token must be dispatched to its selected expert (potentially on a different GPU), processed, and returned. This communication phase is a potential attack surface: an adversary controlling the interconnect could redirect tokens to substitute experts.

The NVIDIA Blackwell and Vera Rubin platforms address this through fabric-level attestation via NVLink 4.0. Data-in-transit encryption secures the communication between experts distributed across multiple GPUs, ensuring that the all-to-all communication phase remains verifiable and protected from "victim flows" or "head-of-line blocking." The NVLink fabric participates in the confidential computing scheme, with fabric encryption keys managed by attested management processors.

7.4 Gating Network Integrity

Beyond routing determinism, the MSAF requires attestation of the gating network itself. The router weights determine which experts are activated for each input class, making them a high-value target for adversarial manipulation. A subtle modification to router weights could redirect specific input patterns (e.g., queries about a particular topic, or inputs from a particular demographic) to a compromised expert while leaving all other behavior unchanged.

The attestation artifact for an MoE model therefore includes: (a) the hash of the complete weight set including all experts and the router; (b) the deterministic routing log for attested inferences, mapping each token to its selected expert(s); and (c) the Tier 2 invariant measurements computed *per expert* (not just globally), enabling detection of individual expert degradation or starvation.

The specific MoE attestation protocol—including per-expert invariant monitoring schedules, the routing log format and compression scheme, the connection between MoE attestation and the Gradient Starvation Envelope (Auburn Clause AI-3′), and the fabric-level attestation verification procedure—is specified in the private version of the framework.

Part IV: Honest Limitations

8 Failure Modes, Theoretical Ceilings, and Epistemic Honesty

The design of any governance framework that does not confront its own limitations is an exercise in marketing, not engineering. The MSAF faces fundamental constraints at every layer—from hardware vulnerabilities to mathematical impossibility results—that bound what attestation can promise. This section catalogs those constraints honestly, because the framework's credibility depends on acknowledging what it cannot do as clearly as it describes what it can.

8.1 Rice's Theorem: The Ultimate Ceiling

Rice's theorem (1953) establishes that all non-trivial semantic properties of programs are undecidable. Determining whether a neural network produces "safe" outputs for all possible inputs is a semantic property. Therefore, no algorithm can decide this for all neural networks. No finite set of test cases, no monitoring scheme, no attestation architecture can guarantee behavioral safety for arbitrary future inputs.

This is not a limitation of the MSAF specifically—it is a limitation of computation itself. Attestation of the training process (correct code, correct data, metrics in bounds) does *not* guarantee behavioral properties of the resulting model. The gap between process attestation and behavioral guarantees is irreducible.

Proposition 8.1 (The Process-Behavior Gap)**.** *Let $\mathcal{M}$ be a model produced by training process $\mathcal{P}$ on dataset $\mathcal{D}$. Let ϕ be any non-trivial behavioral property (e.g., "$\mathcal{M}$ never produces harmful output"). Then:*

$$\mathit{Attest}(\mathcal{P}, \mathcal{D}) \not\Rightarrow \phi(\mathcal{M}) \tag{7}$$

Process attestation is necessary for accountability but never sufficient for behavioral guarantees.

8.2 The Impossibility Sandwich

The paper "On the Mathematical Impossibility of Safe Universal Approximators" (arXiv 2507.03031, 2025) formalizes a more specific constraint: the *minimum* complexity required for an AI system to be useful exceeds the *maximum* complexity for which safety can be formally verified. Any model powerful enough to perform the tasks we demand of it is too complex for its safety to be provably certified.

This result does not invalidate attestation—it contextualizes it. The MSAF operates in the space between "provably safe" (impossible for useful models) and "completely unmonitored" (the status quo). The framework provides probabilistic risk reduction: a model that passes all five invariant checks is *more likely* to be operating correctly than one that does not, even though the invariants cannot *guarantee* correctness.

8.3 TEE Vulnerabilities: Hardware Trust Has Physical Limits

The TEE trust model assumes that silicon is trustworthy even when software is not. A cluster of 2024–2025 attacks has demonstrated that this assumption has physical limits.

8.3.1 Physical Memory Interposition

TEE.Fail (Georgia Tech and Purdue, late 2025) uses an approximately $1,000 DDR5 interposer to capture all memory bus traffic, exploiting deterministic AES-XTS tweaking to extract

ECDSA attestation keys from Intel's Provisioning Certification Enclave. This breaks SGX attestation, TDX attestation (which depends on SGX), and—critically—NVIDIA GPU Confidential Computing's entire attestation chain, since GPU attestation is anchored in the CPU TEE. Researchers demonstrated that "extracted attestation keys can be used to compromise NVIDIA's GPU Confidential Computing, allowing attackers to run AI workloads without any TEE protections" while producing valid attestation reports.

Battering RAM (KU Leuven) achieves similar results on DDR4 with a $50 interposer, enabling falsified SEV-SNP attestation by replaying encrypted launch digests.

Both Intel and AMD classify physical access attacks as outside their threat models—a position that creates fundamental tension for organizations trusting TEE attestation in shared data center infrastructure where they do not control physical access.

8.3.2 ML-Specific Side Channels

Machine learning workloads exhibit distinctive, data-dependent memory access patterns that are inherently observable through microarchitectural side channels, even inside TEEs:

- **HyperTheft** (CCS 2024): Trains hyper-networks on ciphertext collision patterns from AMD SEV-SNP to directly output DNN model weights, achieving 77–97% test accuracy on surrogate models from passive observation of a single inference execution—no queries to the victim model needed.
- **CipherSteal** (IEEE S&P 2025, Distinguished Paper Award): Steals input data from TEE-protected neural networks by exploiting data-dependent memory access patterns during GEMM operations.
- **DeepCache** (CCS 2024): Uses contrastive learning on cache side-channel traces to recover DNN architectures from compiled executables.

8.3.3 Firmware and Microcode Vulnerabilities

CVE-2024-56161 (CVSS 7.2) revealed that AMD's microcode signature verification used an insecure hash function, allowing local administrators to load malicious microcode that completely undermines all SEV-SNP guarantees across Zen 1 through Zen 5 processors. **TDXploit** (USENIX Security 2025) revived single-stepping attacks against Intel TDX with >99.99% accuracy even with Intel's mitigations active, demonstrating full AES key recovery. An independent security analysis of NVIDIA's GPU CC architecture (IBM Research, July 2025) revealed that GPU memory is *not encrypted at runtime*—relying solely on access control firewalls—and identified residual information exposure through 1,042 non-zeroed BAR0 register fields.

8.3.4 Implication for the MSAF

TEE attestation should be treated as **necessary but not sufficient** for ML workload protection. It significantly raises the cost of attacks, creates tamper-evident audit trails, and enables after-the-fact forensics. It does not provide cryptographic-level assurance against state-level adversaries with physical access. The MSAF's layered architecture is designed to degrade gracefully: even if Tier 1 is compromised, Tier 2 continuous monitoring and Tier 3 zkML proofs provide independent verification channels.

8.4 Partial Observability and Goodhart's Law

The five invariants defined in §2 are scalar summaries of an astronomically high-dimensional parameter space. Two models with identical gradient norms can have radically different behaviors. Two models with identical entropy can encode entirely different biases. The monitored

metrics are projections—useful projections, empirically validated projections, but projections nonetheless.

Goodhart's Law applies directly: if adversaries know which metrics are monitored, they can optimize to satisfy those metrics while pursuing objectives invisible to the monitoring scheme. This is not hypothetical. The GCB backdoor attack (ICLR 2025) achieves less than 1% clean accuracy drop while evading all known defenses. HPMI injects backdoors by replacing a single attention head with >99.55% attack success while bypassing state-of-the-art detection.

The MSAF mitigates this through two mechanisms: (a) the invariant set is designed to be *multi-dimensional*, making it difficult to satisfy all five invariant classes simultaneously while maintaining a hidden adversarial objective; and (b) the Tier 3 periodic audits employ different evaluation methodologies than the Tier 2 continuous monitors, reducing the probability that an adversary optimized against one tier will evade both. These are probabilistic mitigations, not guarantees.

8.5 GPU Floating-Point Non-Determinism

A structural challenge for any attestation framework operating on GPU hardware is floating-point non-determinism. PyTorch documentation explicitly states: "Completely reproducible results are not guaranteed across PyTorch releases, individual commits, or different platforms." Non-deterministic GPU kernels (`atomicAdd`, `scatter_add`, `embedding_bag`) produce non-associative floating-point reductions whose order varies across runs.

Research (arXiv 2408.05148, 2024) found that 1,000 models trained with identical inputs are "completely non-reproducible." This means two identical training runs can produce different attestation measurements. The MSAF addresses this by defining *tolerance bands* rather than exact-match requirements for all continuous-valued invariants. The width of these bands is calibrated to the measured non-determinism of the target hardware platform.

This creates a known tradeoff: wider tolerance bands accommodate legitimate non-determinism but provide slack that a sophisticated adversary could exploit. The MSAF documents this tradeoff explicitly rather than ignoring it, and recommends that organizations deploying in high-assurance environments enforce deterministic execution modes (at significant performance cost) for attested workloads.

8.6 The Honest Framing

The MSAF provides **probabilistic risk reduction and accountability infrastructure**, not safety guarantees. This framing is not a weakness—it is an accurate description of what governance can achieve in any complex domain:

- **Financial auditing** certifies process compliance without guaranteeing future solvency. Enron passed its audits. Sarbanes-Oxley did not prevent the 2008 financial crisis. But the existence of auditing infrastructure makes fraud detectable, attributable, and legally consequential.
- **Medical device regulation** certifies the manufacturing process without guaranteeing therapeutic outcomes for every patient. A device can pass all FDA requirements and still harm individual patients. But the regulatory framework ensures accountability.
- **Food safety inspection** certifies production conditions without guaranteeing the absence of all contaminants. But inspection infrastructure makes contamination events traceable and correctable.

Each of these domains operates successfully with probabilistic, process-based assurance. The AI governance field should adopt the same epistemic stance: attestation as accountability in-

frastructure that enables evidence-based deployment, not as a safety guarantee that eliminates uncertainty.

Part V: The Demand Signal

9 Regulatory Mapping Across Five Domains

For the MSAF to be useful, it must generate artifacts that map directly to the evidence requirements of global regulatory frameworks. This section provides a detailed mapping across five domains, identifying both the specific requirements that MSAF attestation satisfies and the strategic windows where attestation recognition could be embedded into emerging standards.

9.1 EU AI Act: The Standards Window Is Open

The EU AI Act mandates that high-risk AI systems maintain detailed technical documentation and audit logs for up to 10 years. As of February 2026, **zero harmonized standards have been published or cited in the Official Journal of the EU**, meaning no standard yet provides the legal "presumption of conformity." The original delivery deadline of April 30, 2025, and the amended deadline of August 31, 2025, were both missed. CEN/CENELEC's Technical Boards adopted exceptional acceleration measures in October 2025, targeting all standards for availability by Q4 2026.

This delay creates a strategic window. The cybersecurity standard **prEN 18282** (mapping to Article 15(1)(5)) is undergoing comprehensive redrafting after European Commission review. The MSAF's three-tier architecture maps directly to the anti-tampering requirements of this standard:

- **Article 15(5)** requires resilience against unauthorized third-party tampering—specifically citing data poisoning, adversarial examples, model manipulation, and confidentiality breaches. Tier 1 TEE attestation detects unauthorized platform modifications. Tier 2 continuous monitoring detects model manipulation via invariant violations. Tier 3 provenance hashing detects data poisoning.
- **Article 9** mandates "regular systematic review and updating" throughout the lifecycle. Tier 2's hash-chained continuous metrics provide tamper-evident records of ongoing review.
- **Article 72** requires documented post-market monitoring. The MSAF's Tier 2 log constitutes a cryptographically verifiable post-market monitoring record.
- **Article 10** (Data Governance) requires representative, error-free training data. Tier 3's multiset hash of the training corpus proves that the deployed model was trained on the audited dataset.

The Commission's **Digital Omnibus proposal** (November 2025) links high-risk obligations to standards availability: Annex III systems would apply no later than December 2, 2027, and Annex I product-embedded systems no later than August 2, 2028. If this proposal passes, organizations have a 12–18 month window to influence the content of harmonized standards before they become binding. The MSAF's modular design—standardize the cryptographic protocol, let the health invariants evolve—aligns with CEN/CENELEC's own stated preference for performance-based rather than prescriptive standards.

9.2 FDA SaMD: Cryptographic PCCPs

The FDA's Predetermined Change Control Plan (PCCP) framework, finalized December 4, 2024 and updated August 18, 2025, authorizes pre-approved modification pathways for AI/ML-enabled medical devices. A compliant PCCP requires three components: a description of modifications, a modification protocol (data management, retraining practices, performance evaluation, update procedures), and an impact assessment.

Cryptographic attestation creates a natural fit: the MSAF can prove that modifications stayed within pre-authorized parameters via signed state transitions. Specifically:

- **Data Management Evidence:** Tier 3 multiset hashes prove that retraining data matches the documented corpus, sourced from the required multiple clinical sites.
- **Performance Evaluation:** Tier 2 invariant measurements provide continuous evidence that model health remained within validated parameters after each modification.
- **Cumulative Drift:** Tier 2 KL divergence monitoring detects when iterative modifications have cumulatively shifted the model beyond its validated operating envelope.
- **Audit Trail:** The complete MSAF attestation chain serves as the QMS documentation required under the QMSR transition to ISO 13485 (effective February 2, 2026).

Adoption remains low: of over 3,000 510(k) submissions in 2024, only 41 ($\sim$1%) included PCCPs. The MSAF lowers the barrier to PCCP adoption by automating the evidence generation that currently requires manual documentation.

9.3 SR 11-7: Cryptographic "Effective Challenge"

The Federal Reserve's SR 11-7 (April 2011) remains the primary model risk management framework for US banking despite never being formally updated to address AI/ML. Its three pillars—model development documentation, model validation, and ongoing monitoring—all benefit from cryptographic attestation:

- **Conceptual Soundness:** Tier 3 attestation artifacts provide verifiable evidence of architecture decisions, training procedures, and the empirical basis for model selection—the "design traceability" that examiners expect.
- **Ongoing Monitoring:** Tier 2 continuous invariant measurements replace periodic manual review with cryptographically signed metric streams. Drift detection via KL divergence monitoring and gradient integrity checks provide the "stability metrics" (PSI and equivalents) that validation reports must include.
- **Effective Challenge:** SR 11-7's requirement for independent, objective parties to review model decisions becomes more tractable when challengers can verify attestation artifacts rather than requesting proprietary access to model weights. An external validator can confirm that a model's attested health metrics remained within bounds without ever seeing the model itself.
- **Third-Party Vendor Validation:** Banks relying on third-party AI solutions face the "vendor evidence paradox"—responsible for validating vendor models but unable to access proprietary weights. zkML inference proofs (Tier 3) allow vendors to prove computational correctness without revealing weights, directly addressing this gap.

A significant open question: whether generative AI outputs constitute "quantitative estimates" under SR 11-7's model definition remains unresolved. Text and image outputs may not fit the traditional definition, potentially placing GenAI outside the framework's formal scope. The MSAF's evidence artifacts remain useful regardless of how this definitional question is resolved, as they satisfy the broader prudential expectation of documented governance.

9.4 US Federal Procurement: The Compliance Fork

The federal AI procurement landscape has undergone wholesale replacement. OMB M-24-18 was rescinded on April 3, 2025, and replaced by M-25-22. M-24-10 was replaced by M-25-21. Both replacements implement EO 14179 (January 23, 2025). Additionally, **M-26-04** (December 11, 2025) establishes enforceable requirements for Large Language Models around two principles: **truth-seeking** (historical accuracy, scientific objectivity, acknowledgment of uncertainty) and **ideological neutrality** (nonpartisan outputs, no encoded partisan judgments).

This creates what the compliance community calls the "compliance fork": a model tuned to aggressively filter content to satisfy EU fundamental rights requirements might be flagged by a US federal agency as violating "ideological neutrality" if the filters are perceived as suppressing specific viewpoints. Organizations serving both markets face potentially contradictory evidence requirements.

The MSAF addresses the compliance fork through architectural modularity:

- Tier 1 and Tier 2 attestation artifacts (platform integrity, health invariants) are jurisdiction-neutral. Gradient stability and entropy floors are physics, not politics.
- Tier 3 audit artifacts can be configured domain-specifically: EU-targeted audits include fundamental rights impact assessments and bias testing against protected classes; US-targeted audits include truth-seeking evaluations and ideological neutrality assessments.
- The underlying attestation chain is shared, ensuring that domain-specific compliance artifacts are derived from the same verified governance substrate rather than maintained as separate, potentially inconsistent documentation systems.

Key near-term deadlines: agencies must revise procurement policies for M-26-04 compliance by March 11, 2026, and document implementation of minimum risk management practices for high-impact AI by April 3, 2026.

Meanwhile, M-26-05 (January 23, 2026) rescinded the previously mandatory Secure Software Development Attestation Form and made Software Bills of Materials optional. This deregulatory move in software supply chain security paradoxically increases the value of voluntary attestation: organizations that can demonstrate cryptographic governance evidence differentiate themselves in a market where the floor has been lowered.

9.5 AI Insurance: The De Facto Regulator

The AI insurance market in early 2026 mirrors the cyber insurance market of the late 1990s: traditional policies are excluding newly recognized risks precisely as specialist products emerge to fill the gap. **Verisk's ISO Core Lines introduced optional CGL endorsements excluding generative AI exposures effective January 1, 2026**, with multiple carriers filing their own exclusion wordings. Simultaneously, purpose-built AI coverage is arriving from three directions.

Munich Re (aiSure, since 2018) requires thorough technical due diligence including model performance documentation, training data provenance, drift monitoring systems, and deployment context analysis. Coverage triggers when AI fails against agreed performance benchmarks. The MSAF's Tier 2 continuous metrics and Tier 3 benchmark attestations map directly to Munich Re's evidence requirements.

Armilla AI (the world's first AI-only MGA, launched April 2025, underwritten by Lloyd's syndicates including Chaucer and AXIS Capital, backed by Swiss Re) conducts independent regulatory-grade model evaluations. Governance maturity signals—NIST AI RMF implementation, model testing records, bias controls, incident-response protocols—directly calibrate coverage terms. Better governance translates to better premiums, analogous to safe-driver discounts.

Testudo (launched January 2025) uses external risk assessment powered by a proprietary AI Risk Engine that ingests real-time litigation, regulatory, and incident data rather than requiring deep integration with applicants' AI systems.

A Geneva Association survey (October 2025) found **90% of 600 corporate insurance decision-makers expressed interest in AI insurance**, with two-thirds willing to pay at least 10% higher premiums. ISO/IEC 42001 certification has emerged as the "gold standard" for demonstrating insurability—the AI equivalent of what ISO 27001 became for cyber insurance. The MSAF's attestation artifacts provide the cryptographic backing that strengthens ISO 42001 certifications from self-reported governance documentation to verifiable evidence.

9.6 Regulatory Evidence Mapping

Framework	Specific Requirement	MSAF Evidence Artifact
EU AI Act	Art. 10 (Data Governance)	Tier 3: Multiset hash of training corpus; synthetic data contamination bounds
EU AI Act	Art. 15(5) (Anti-Tampering)	Tier 1: TEE platform quote; Tier 2: continuous integrity monitoring
EU AI Act	Art. 72 (Post-Market)	Tier 2: Hash-chained continuous monitoring log
FDA SaMD	PCCP Validation	Tier 2: Signed state transitions; Tier 3: benchmark attestation
FDA SaMD	QMSR Audit Trail	Full attestation chain as QMS documentation
SR 11-7	Ongoing Monitoring	Tier 2: Signed drift, stability, and gradient metrics
SR 11-7	Effective Challenge	Tier 3: zkML proofs enabling weight-private validation
OMB M-26-04	Transparency Documentation	Tier 3: Audit artifacts configured for truth-seeking evaluation
OMB M-25-21	High-Impact AI Risk Mgmt	Tier 2 + Tier 3: Pre-deployment and ongoing attestation
Insurance	Governance Maturity	Full MSAF attestation as underwriting evidence
Insurance	Performance Guarantees	Tier 2: Continuous metrics against contractual KPIs

Table 8: Mapping of MSAF attestation artifacts to specific regulatory evidence requirements.

10 The Economic Case: From Exclusion to Evidence-Based Pricing

The MSAF is not merely a technical architecture—it is an economic proposition. The convergence of regulatory enforcement timelines, insurance market dynamics, and procurement mandates creates immediate commercial demand for cryptographic governance evidence.

10.1 The Insurance Inflection

The AI-specific insurance market is projected to reach approximately **$4.8 billion in annual premiums by 2032** at roughly 80% CAGR. This growth is driven by a simultaneous contraction in traditional coverage (CGL exclusions for generative AI) and expansion in specialist products (Armilla, Testudo, Munich Re aiSure). Claims experience is accumulating rapidly: the Air Canada chatbot ruling (2024), Arup's $25M deepfake fraud, Google AI Overviews litigation ($110–210M claimed), and Coalition's finding that chatbots featured in 5% of all web privacy claims.

For organizations, the MSAF offers a direct path from "excluded from coverage" to "favorable terms." An organization that can produce a cryptographic attestation chain—proving platform integrity, continuous health monitoring, and validated provenance—presents a fundamentally lower risk profile than one relying on self-reported model cards. The attestation artifact functions as the AI equivalent of a building's fire safety certification: not a guarantee against fire, but evidence of systematic risk reduction that enables insurance underwriting.

10.2 The Tamper-Evident Decision Receipt

In the liability sector, the MSAF provides what no current governance artifact can: a tamper-evident record that a specific model was operating in a certified, healthy state at the moment of a specific decision. If an AI-driven medical diagnostic leads to a malpractice claim, or an automated credit system is accused of bias, the organization can produce a cryptographic attestation proving:

1. The model deployed was the model that was validated (Tier 1 weight hash).
2. The model's health metrics were within attested bounds at the time of the decision (Tier 2 signed measurements).
3. The model was trained on the documented, audited dataset (Tier 3 provenance hash).

This clarity facilitates more accurate risk pricing, more efficient procurement processes, and more defensible legal positions. Organizations can require "attestable model health" as a contractual term—a capability that transforms AI governance from a cost center into a competitive advantage.

10.3 Federal Procurement Pull

US federal agencies face an April 3, 2026 deadline to document implementation of minimum risk management practices for high-impact AI. AI use case inventories nearly doubled from 571 (2023) to 1,110 (2024), with generative AI use cases increasing ninefold. Yet GAO found only 4 of 35 recommendations from its December 2023 baseline report had been implemented, and 15 of 20 agencies had incomplete or inaccurate AI inventory data.

The MSAF's automated attestation generation addresses the gap between mandate and implementation. Rather than requiring each agency to build bespoke documentation processes, a standardized attestation framework produces the transparency reports, model cards, and risk

management evidence that M-25-21 and M-26-04 require—derived from the same cryptographic substrate that serves EU, FDA, and insurance requirements.

10.4 The Convergence Advantage

The most actionable insight across all five regulatory domains is architectural: organizations operating across multiple jurisdictions should design unified evidence management systems that generate domain-specific compliance artifacts from a common governance substrate. The MSAF provides this substrate. A single attestation chain—Tier 1 platform integrity, Tier 2 continuous health metrics, Tier 3 periodic audit anchors—can simultaneously serve EU conformity assessment, FDA PCCP documentation, SR 11-7 validation artifacts, federal procurement transparency disclosures, and insurance underwriting evidence.

Organizations that build this infrastructure now hold a structural advantage as every domain continues demanding more evidence, not less. The alternative—maintaining five separate documentation systems for five regulatory regimes—is neither scalable nor defensible.

Part VI: Conclusion

11 The Path Forward: Standards, Not Ossification

A critical sub-question for any governance framework is how to standardize without stifling innovation. The history of technology regulation is littered with standards that locked industries into outdated metrics: financial risk models anchored to Gaussian assumptions that failed catastrophically in 2008, medical device standards designed for deterministic software applied to probabilistic ML systems, and cybersecurity frameworks built for discrete-state machines applied to continuous-inference engines. The MSAF must avoid this trap.

11.1 The Modular Principle

The framework proposes a strict separation between *protocol* and *content*:

- **Standardize the protocol:** How a proof is generated, signed, transmitted, verified, and stored. The cryptographic primitives (hash functions, signature schemes, commitment schemes), the attestation token format (COSE, JWT, C2PA), the Merkle tree construction algorithm, and the inter-tier binding mechanism. These are infrastructure—they change slowly and benefit from interoperability.
- **Let the content evolve:** What is being proven. The specific health invariants (§2), the threshold values, the monitoring frequencies, the contamination detection algorithms, and the benchmark methodologies. These are science—they change rapidly as research advances, and premature standardization would freeze the framework at 2026-era understanding.

This separation ensures that when a new invariant class is discovered (or an existing one is shown to be insufficient), it can be incorporated into the MSAF without modifying the attestation infrastructure. The protocol remains stable; the content is versioned and updatable.

11.2 Standardization Venues

The MSAF's protocol layer is a natural candidate for standardization through multiple venues operating at different scopes:

CEN/CENELEC JTC 21 is developing harmonized standards for the EU AI Act. The cybersecurity standard prEN 18282 (Article 15) is undergoing comprehensive redrafting as of early 2026—an active opportunity to embed attestation protocol recognition. The accuracy and robustness standard prEN 18229-2 (Article 15(1)(3)(4)) entered enquiry in February 2026 and could reference attestation-based performance monitoring as an acceptable evidence methodology.

ISO/IEC JTC 1/SC 42 manages AI management standards, including ISO/IEC 42001 (AI Management System). The MSAF's three-tier architecture provides the cryptographic backing that converts ISO 42001's documentation requirements from self-reported governance claims into verifiable evidence. Integration would strengthen 42001 certifications without requiring changes to the standard's structure.

The IETF RATS working group (Remote ATtestation procedureS) has already defined the token format and verification architecture that the MSAF builds upon. Extension of RATS to include ML-specific claims (model weight hash, invariant measurements, provenance hash) is a natural evolution of the existing work.

The Confidential Computing Consortium is developing standardized composite attestation protocols. As of early 2026, no standardized protocol exists for joint CPU–GPU

attestation—the current split-verifier architecture is vendor-specific (Intel Trust Authority + NVIDIA NRAS). A vendor-neutral composite attestation standard would directly enable the MSAF's Tier 1 to operate across hardware platforms.

11.3 Open-Source Implementation Path

Standardization without accessible implementation is merely aspiration. The MSAF's protocol layer should be implementable through open-source toolkits that prevent vendor lock-in and enable independent audit. Several existing projects provide foundations:

- **Atlas** (Intel Labs, Apache 2.0): Provides the C2PA manifest generation and TDX attestation integration that maps to Tier 1 and Tier 3.
- **AIBoMGen** (Ghent University–imec): Provides the CycloneDX AIBOM format and in-toto attestation pipeline that maps to Tier 3 provenance.
- **EZKL** (Zkonduit, Trail of Bits audited): Provides the zkML proof generation pipeline that maps to Tier 3 inference verification.
- **Sigstore / Rekor**: Provides the transparency log infrastructure for publishing attestation root hashes to an immutable, publicly auditable ledger.

The gap between these existing tools and a complete MSAF implementation is the *composition layer*—the software that binds Tier 1, Tier 2, and Tier 3 into a unified artifact with a single root hash. This composition layer is the MSAF's core intellectual contribution and the primary subject of ongoing development.

The composition layer implementation—including the specific binding algorithms, the artifact schema, the inter-tier hash chain construction, the selective disclosure protocol, and the reference implementation architecture—is specified in the private version of the framework.

11.4 The Risk of Inaction

The alternative to proactive standardization is reactive regulation. If the AI governance community does not define what "attestable model health" means technically, regulators will define it non-technically—through prescriptive documentation requirements, rigid audit checklists, and compliance frameworks designed by lawyers rather than engineers. The result will be governance theater: organizations producing voluminous paperwork that satisfies legal requirements without providing genuine assurance.

The standards window is open now. CEN/CENELEC harmonized standards are being drafted. The EU AI Act's high-risk obligations are approaching. The FDA's PCCP framework is final but under-adopted. The AI insurance market is actively seeking underwriting criteria. The organizations and researchers who contribute to defining attestation standards in the next 12–18 months will shape the governance infrastructure that persists for decades.

12 Conclusion: Accountability Infrastructure for the Evidence-Based Era

The Model State Attestation Framework is not a binary question of feasibility but a spectrum of assurance levels achievable at different computational costs and trust assumptions.

At the strongest end, zkML provides cryptographic proof of inference correctness for models up to 13 billion parameters, but at approximately 10,000× computational overhead and with

training verification remaining intractable at frontier scale. At the pragmatic end, TEE-backed signed metrics provide tamper-evident attestation of monitored invariants for any model size at under 7% overhead on current hardware and under 2% on next-generation architectures, but depend on trusting hardware vendors and monitoring only a low-dimensional projection of model behavior.

The fundamental insight is that **the gap between process attestation and behavioral guarantees is irreducible**. Rice's theorem ensures that no finite monitoring scheme can certify arbitrary model safety. But this gap is not unique to AI. Financial auditing certifies process compliance without guaranteeing future solvency. Medical device regulation certifies the manufacturing process without guaranteeing therapeutic outcomes for every patient. Food safety inspection certifies production conditions without guaranteeing the absence of all contaminants. Each of these domains operates successfully with probabilistic, process-based assurance—and each would be unimaginable without it.

The AI governance field must adopt the same epistemic stance. The question is not whether attestation can guarantee safety—it cannot, and no honest framework should claim otherwise. The question is whether the absence of attestation is acceptable in a world where foundation models adjudicate credit, recommend diagnoses, allocate resources, and execute trades. The answer, across five regulatory domains, is increasingly and emphatically no.

The building blocks exist. TEE-backed platform attestation reaches production scale at modest overhead. zkML inference verification handles billion-parameter models in minutes. Provenance frameworks track training lineage with cryptographic fidelity. Continuous invariant monitoring detects gradient instability, entropy collapse, and distributional drift in real time. What does not yet exist is the composed system—the three-tier architecture that binds these components into a unified, auditable, regulatory-grade attestation artifact.

The MSAF is that composition. It converts AI deployment from trust-based to evidence-based. It provides the accountability infrastructure that enables regulated adoption of foundation models at scale. And it does so honestly—acknowledging its theoretical ceilings, documenting its failure modes, and positioning itself as what governance actually is: not a guarantee against failure, but a systematic reduction of the probability of undetected failure and a framework for accountability when failure occurs.

The regulatory pull is strong and growing. The standards window is open. The honest path forward is to build the layered stack, acknowledge its limits, and deploy it as the foundation for a regulatory regime that converts AI from trust-based to evidence-based—knowing that evidence reduces but never eliminates uncertainty.

References

[1] M. Mitchell et al., "Model Cards for Model Reporting," *Proc. FAT**, 2019.

[2] Z. Allen-Zhu and Y. Li, "Physics of Language Models: Knowledge Storage, Extraction, and Manipulation," *arXiv preprint*, arXiv:2309.14316, 2023.

[3] I. Shumailov et al., "The Curse of Recursion: Training on Generated Data Makes Models Forget," *arXiv preprint*, arXiv:2305.17493, 2023.

[4] T. Takase et al., "Spike No More: Stabilizing the Pre-training of Large Language Models," *COLM*, 2024.

[5] Y. Li et al., "AGGC: Adaptive Gradient Gradient Clipping for Training Deep Neural Networks," *arXiv preprint*, January 2026.

[6] "SPAM: Spike-Aware Adam with Momentum Reset for Stabilizing LLM Training," *arXiv preprint*, arXiv:2501.06842, 2025.

[7] M. Nazeri et al., "Entropy-based Adversarial Detection in Neural Network Activations," August 2025.

[8] W. Shi et al., "Detecting Pretraining Data from Large Language Models," *ICLR*, 2024.

[9] "CDD: Contamination Detection in Large Language Models," *ACL*, 2024.

[10] "LLM Vulnerability to GPU Soft Errors," *arXiv preprint*, January 2026.

[11] A. Barakat and P. Bianchi, "Convergence and Dynamical Behavior of the ADAM Algorithm for Nonconvex Stochastic Optimization," *SIAM J. Optim.*, vol. 31, no. 1, 2020.

[12] "PIDAO: PID-controller-based Optimization with Provable Convergence," *Nature Communications*, 2024.

[13] S. Zhang et al., "Compositional Neural Certificates for Networked Dynamical Systems," *L4DC*, 2023.

[14] "Certificates in AI: Learn but Verify," *Communications of the ACM*, 2025.

[15] "Constant-Size Cryptographic Evidence Structures," *arXiv preprint*, arXiv:2511.17118, 2025.

[16] Y. Zhu et al., "Benchmarking Confidential Computing for LLM Inference on NVIDIA H100," *arXiv preprint*, arXiv:2409.03992, September 2024.

[17] J. Lee and F. Wang, "Performance Analysis of Distributed Training under GPU TEEs," *arXiv preprint*, arXiv:2501.11771, January 2025.

[18] NVIDIA, "Blackwell Architecture Whitepaper: Confidential Computing," 2025.

[19] "TEE.Fail: Physical Side-Channel Extraction of TEE Attestation Keys via DDR5 Interposition," Georgia Tech and Purdue, 2025.

[20] "Battering RAM: Low-Cost Interposer Attacks on DDR4 Confidential VMs," KU Leuven, 2025.

[21] "HyperTheft: Thieving Model Weights from TEE-Shielded Neural Networks via Ciphertext Side Channels," *ACM CCS*, 2024.

[22] "CipherSteal: Stealing Input Data from TEE-Protected Neural Networks," *IEEE S&P*, 2025.

[23] CVE-2024-56161, "AMD Microcode Signature Verification Bypass," CVSS 7.2, 2024.

[24] "TDXploit: Single-Stepping Attacks against Intel TDX," *USENIX Security*, 2025.

[25] IBM Research, "Security Analysis of NVIDIA GPU Confidential Computing Architecture," *arXiv preprint*, arXiv:2507.02770, July 2025.

[26] Intel, "Intel Trust Domain Extensions (TDX) Module Architecture Specification," 2024.

[27] Intel, "Intel Trust Authority: Composite Attestation Documentation," 2025.

[28] H. Sun, J. Li, and T. Zhang, "zkLLM: Zero Knowledge Proofs for Large Language Models," *ACM CCS*, 2024.

[29] Lagrange Labs, "DeepProve: Verifiable AI Inference at Scale," Technical Report, 2025.

[30] J. Morton et al., "EZKL: Easy Zero-Knowledge Learning," Zkonduit Inc., v23.0.3, 2025. Trail of Bits Audit Report.

[31] "zkGPT: Efficient Zero-Knowledge Proofs for GPT-2 Inference," *USENIX Security*, 2025.

[32] "ZKTorch: Accumulation-Based Zero-Knowledge Proofs for Large Models," 2025.

[33] "zkPyTorch: Zero-Knowledge Inference for Llama-3 via Expander Engine," 2025.

[34] "Kaizen: Verifiable Training of VGG-11 via Recursive ZK Proofs," *ACM CCS*, 2024.

[35] "VeriLoRA: End-to-End ZK-Verifiable LoRA Fine-Tuning at 13B Scale," August 2025.

[36] "ZKLoRA: LoRA Module Compatibility Verification," *ICML Workshop*, 2025.

[37] Ingonyama, "ICICLE: GPU Acceleration for Zero-Knowledge Proofs," 2025.

[38] "ZKProphet: Profiling the zkML Proving Pipeline," *IEEE IISWC*, 2025.

[39] "UniZK: Unified ASIC Architecture for Zero-Knowledge Proofs," *ASPLOS*, 2025.

[40] "On the Mathematical Impossibility of Safe Universal Approximators," *arXiv preprint*, arXiv:2507.03031, 2025.

[41] H. G. Rice, "Classes of Recursively Enumerable Sets and Their Decision Problems," *Trans. AMS*, vol. 74, no. 2, 1953.

[42] "GCB: Gradient-Controlled Backdoor Attacks Evading All Known Defenses," *ICLR*, 2025.

[43] "HPMI: Single Attention Head Backdoor Injection," 2025.

[44] "On the Non-Reproducibility of GPU-Based Deep Learning Training," *arXiv preprint*, arXiv:2408.05148, 2024.

[45] Intel Labs, "Atlas: End-to-End Attestable ML Pipelines," v0.1.0, Apache 2.0, February 2025.

[46] Mithril Security, "AICert: Cryptographic Binding of Training Code and Model Weights," v1.0, September 2024.

[47] Ghent University–imec, "AIBoMGen: Automated AI Bill of Materials Generation," *CAIN*, 2026.

[48] European Parliament, "Regulation (EU) 2024/1689: The Artificial Intelligence Act," *Official Journal of the EU*, 2024.

[49] US Food and Drug Administration, "PCCP for AI/ML-Enabled Device Software Functions: Final Guidance," December 2024; updated August 2025.

[50] Board of Governors of the Federal Reserve System, "SR 11-7: Guidance on Model Risk Management," April 2011.

[51] Office of the Comptroller of the Currency, "Comptroller's Handbook: Model Risk Management," Bulletin 2021-39, 2021.

[52] Office of Management and Budget, "M-26-04: Increasing Public Trust in AI Through Unbiased AI Principles," December 2025.

[53] Office of Management and Budget, "M-25-21: Accelerating Federal Use of AI," 2025.

[54] International Organization for Standardization, "ISO/IEC 42001: Artificial Intelligence Management System," 2023.

[55] Geneva Association, "AI Insurance Survey: Corporate Decision-Maker Perspectives," October 2025.

[56] Deloitte, "AI Insurance Market Projections," 2025.

[57] Armilla AI, "AI Liability Insurance: Underwriting Methodology," Lloyd's of London, April 2025.

[58] "Attestable Audits of AI Systems via Three-Step Verification Protocol," *ICML*, 2025.

[59] NVIDIA, "Cryptographic Model Signing for NGC-Published Models," OpenSSF Model Signing Standard, March 2025.

[60] Succinct, "SP1 Hypercube: Proving 99.7% of Ethereum Blocks in Under 12 Seconds," November 2025.

Appendix: Intellectual Property Declaration

Auburn Patent Family Fields

Intellectual Property (IP) Declaration

The methods, logic structures, and "Certified Constant" registries contained in the associated works are the sole property of **Ryan Fields**.

Public License (Non-Commercial)

This work is licensed under the **Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)** license.

- **Academic Use:** Researchers may share and use this framework for non-commercial academic purposes, provided full attribution is given to Ryan Fields.
- **No Derivatives:** No modifications or "remixes" of the "Certified Constants" or logical proofs are permitted without express written consent.

Commercial Prohibition

Commercial use of this framework is strictly prohibited. This includes, but is not limited to:

- Use within proprietary high-frequency trading (HFT) risk models.
- Integration into commercial high-assurance AI governance software.
- Use by private financial institutions for "tail-risk" auditing of prime distribution variance.

Contact

Ryan Fields

`UncleBroFields@proton.me`
`fieldsryanchristopher@gmail.com`