# The Model Attestation Interface (MAI-1)

*A Normative Profile and Conformance Protocol for Foundation Model Governance*

Clause AI-5

Auburn Patent Family

Ryan Fields

UncleBroFields@proton.me
fieldsryanchristopher@gmail.com

February 2026

---

**Document Classification:** Normative Specification (Draft v1.0)

**Status:** Candidate Harmonized Profile / Procurement Reference

**Clause Designation:** AI-5

**Family:** Auburn Patent Family

---

*This document is the public specification.*

*Strategic implementation details, binding algorithms, and the reference implementation architecture are reserved in the private version.*

**Abstract**

The deployment of foundation models into regulated, safety-critical, and liability-bearing contexts has exposed a fundamental governance gap: no standardized, machine-verifiable mechanism exists to prove that a specific model was operating in a certified internal state at the moment of inference or training. Current governance artifacts—model cards, training logs, static benchmarks—are descriptive rather than prescriptive and fundamentally non-verifiable. This document defines the **Model Attestation Interface (MAI-1)**, a normative IETF RATS-conformant attestation profile that composes three verification layers—hardware-rooted platform attestation, continuous model state invariants, and cryptographic supply chain provenance—into a single verifiable artifact for foundation-scale models. MAI-1 builds upon the theoretical foundations of the Model State Attestation Framework (MSAF) and specifies: a canonical attestation endpoint, required evidence fields encoded as Entity Attestation Token (EAT) claims, cryptographic binding rules for decision receipts, and a three-level conformance test suite with strictly binary pass/fail semantics. The interface is mapped to specific evidence requirements across seven regulatory frameworks (EU AI Act, SR 11-7, FDA SaMD PCCP, OMB M-25-21, OMB M-26-04, FY2026 NDAA §1513, AI Insurance) and three harmonized standards currently in CEN/CENELEC Enquiry (prEN 18282, prEN 18229-2, prEN 18286). The honest framing: MAI-1 provides accountability infrastructure and probabilistic risk reduction, 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. The standards window is open. What is missing is the composition layer—the normative profile that binds these components into an enforceable, conformance-testable interface. MAI-1 is that profile.

# Contents

# 1 Executive Summary

This document defines the **Model Attestation Interface (MAI-1)**: a standardized, machine-verifiable interface for producing cryptographic evidence that a foundation model was operating in a **certified internal state** at the moment of inference or training.

MAI-1 transforms AI governance from descriptive documentation into **enforceable evidence**, enabling four operational outcomes that no existing artifact provides simultaneously:

(i) **Regulatory conformity assessment**: automated evidence generation satisfying EU AI Act Articles 11 and 15, SR 11-7 ongoing monitoring, FDA SaMD PCCP change traceability, and OMB M-25-21 high-impact AI requirements.

(ii) **Insurance underwriting**: cryptographic governance artifacts replacing self-reported questionnaires, enabling the transition from epistemic uncertainty to actuarial risk quantification.

(iii) **Procurement eligibility**: copy-pasteable contract language allowing contracting officers to mandate verifiable AI governance without bespoke technical evaluation.

(iv) **Post-incident forensic attribution**: immutable decision receipts binding specific outputs to specific model states on verified platforms, establishing the evidentiary chain that courts and regulators increasingly demand.

The interface is constructed from existing, individually validated components: the IETF Entity Attestation Token (RFC 9711), the `measured-component` extension for software and model measurements, Concise Reference Integrity Manifests (CoRIM) for reference value distribution, and Supply Chain Integrity, Transparency, and Trust (SCITT) receipts for provenance binding. Each component exists in isolation. No system yet composes them into a unified, conformance-testable attestation artifact for foundation-scale models.

MAI-1 is that composition layer.

**Requirement 1.1** (Normative Claim)**.** *Any AI system deployed in a regulated, safety-critical, or liability-bearing context **MUST** expose an MAI-1 compliant attestation endpoint. Systems unable to provide valid MAI-1 attestation receipts are non-compliant by construction.*

This specification defines: the canonical interface (§6), the required evidence fields (§6), the mandatory invariants and thresholds (§7), the cryptographic binding rules (§8), the conformance test suite (§9), and the regulatory evidence mapping (§10). The theoretical foundations, including the five internal model invariants and the three-tier attestation architecture, are established in the Model State Attestation Framework (MSAF; Fields, 2026).

# 2 Scope and Applicability

## 2.1 Systems in Scope

MAI-1 applies to any AI system whose outputs can trigger legal, financial, medical, safety, or civil-rights consequences. The following system classes are explicitly within scope:

1. **Foundation models** with parameter counts $\geq 10^9$, including both dense architectures and Mixture-of-Experts (MoE) configurations, whether deployed for inference, fine-tuning, or continuous learning.

2. **Autonomous decision systems** whose outputs serve as a principal basis for decisions with legal, material, or significant effect on individuals—including creditworthiness assessment, diagnostic pathways, public resource allocation, employment screening, and insurance underwriting.

3. **AI systems subject to regulatory obligation**, including but not limited to:
   - EU AI Act high-risk systems (Annex III classifications) and General-Purpose AI models (Article 53);
   - US financial models subject to SR 11-7 / OCC 2011-12 model risk management;
   - FDA-regulated Software as a Medical Device (SaMD) under Predetermined Change Control Plans (PCCP);
   - US federal AI systems classified as high-impact under OMB M-25-21;
   - Large Language Models procured under OMB M-26-04 transparency requirements;
   - Department of Defense AI systems subject to FY2026 NDAA §1513 cybersecurity framework requirements.
4. **AI systems subject to insurance coverage**, including systems underwritten by specialist AI insurers (e.g., Armilla AI, Munich Re aiSure, Testudo, AIUC-1 certified systems) and systems excluded from general commercial liability policies under Verisk/ISO generative AI endorsements effective January 1, 2026.
5. **AI systems in procurement pipelines** where contracting authorities require verifiable governance evidence as a condition of vendor eligibility.

### 2.2 Systems Explicitly Out of Scope

The following system classes are not subject to MAI-1 requirements:

1. Research prototypes operating exclusively in sandboxed experimental environments with no external-facing outputs.
2. Offline experimentation and model development conducted prior to any deployment decision.
3. Consumer-only entertainment systems whose outputs carry no legal, financial, medical, or safety consequences and are not subject to regulatory obligation.

### 2.3 Applicability Principle

**Requirement 2.1** (Applicability Determination)**.** *The applicability of MAI-1* ***SHALL*** *be determined by the* ***consequence profile*** *of the system's outputs, not by the technical architecture of the system itself. If an AI system's output can trigger legal, financial, medical, or safety consequences—whether directly or as a principal input to a human decision—MAI-1 compliance is mandatory. The burden of demonstrating that a system falls outside the scope defined in §2 rests with the deploying entity.*

This consequence-based scoping ensures that MAI-1 remains technology-neutral and architecture-agnostic. The interface does not prescribe how invariants are computed, how models are trained, or which hardware platforms are used. It prescribes what evidence **MUST** be produced and how that evidence **MUST** be structured, signed, and verified. This separation of interface from implementation is the design principle that allows MAI-1 to accommodate both current and future architectures without revision to the protocol itself.

# 3 The Governance Failure This Interface Solves

The AI governance landscape as of early 2026 is defined by a paradox: regulatory, judicial, and insurance enforcement is accelerating rapidly, while the technical infrastructure to demonstrate compliance remains fundamentally inadequate. Total financial exposure across settlements, fines, and verdicts now exceeds $2 billion. The enforcement apparatus is mature. The evidence infrastructure is not.

This section identifies the three structural gaps that MAI-1 resolves. Each gap represents a class of governance failure that no existing artifact—model cards, training logs, static benchmarks, or self-reported questionnaires—can address.

## 3.1 The Identity Gap

No existing governance artifact cryptographically binds a **specific output** to a **specific model** running in a **specific internal state** on a **verified platform**.

The consequences of this gap are no longer theoretical. In *Benavides v. Tesla* (S.D. Fla., August 2025), a federal jury returned a verdict of up to $329 million after Tesla's Autopilot system caused a fatal collision. The evidentiary challenge at trial was straightforward: which version of the model, in which configuration, produced the decision that killed a pedestrian? Tesla's governance artifacts could not answer this question with cryptographic certainty. In *Mobley v. Workday* (N.D. Cal.), the court certified a nationwide collective action for all applicants over 40 screened by Workday's AI since September 2020, establishing that AI vendors face direct liability as agents for employment discrimination. The discovery demands in these cases—prompts, outputs, metadata, model identifiers, audit logs—define a new evidentiary standard that descriptive governance artifacts cannot satisfy.

Courts have now issued what is believed to be the first order requiring an AI company to preserve all user logs, including those users had deleted. "AI discovery" is becoming the de facto model governance audit, and organizations without cryptographic identity binding between outputs and model states face unlimited evidentiary exposure.

## 3.2 The Health Gap

No regulator, insurer, or auditor can currently verify that a model's internal health metrics—gradient stability, entropy reserves, representational coherence, distribution drift—were within certified bounds **at execution time**.

This gap has direct financial consequences in the insurance market. Verisk/ISO filed multistate endorsements effective January 1, 2026, allowing carriers to exclude generative AI exposures from commercial general liability policies. At least nine insurance groups—including WR Berkley, AIG, Great American, Cincinnati Financial, and Philadelphia Insurance—filed to adopt these exclusions. WR Berkley's "Absolute" AI exclusion eliminates coverage for "any actual or alleged use, deployment, or development of Artificial Intelligence," including inadequate AI governance. The structural shift from silent coverage to explicit exclusion means that organizations without demonstrable, verifiable AI health monitoring may soon find themselves **uninsurable**.

Specialist AI insurers have emerged to fill the gap—Armilla AI (Lloyd's-backed, launched April 2025), Munich Re's aiSure (performance guarantees since 2018), Testudo (litigation-data-driven, launched January 2026), and AIUC-1 certified coverage—but every underwriter reports the same constraint: deployers lack the continuous, cryptographically signed health metrics that would enable actuarial pricing. Armilla's CEO has stated that "very few data scientists tend to undertake rigorous stress testing for business or regulatory requirements." Munich Re's four-step technical due diligence process requires historical performance monitoring data that most organizations simply do not possess.

The health gap is not a documentation problem. It is an **infrastructure** problem. No amount of retrospective paperwork can substitute for continuous, signed, hardware-rooted health attestation generated at execution time.

### 3.3 The Provenance Gap

No chain of custody exists from training data through fine-tuning through deployment through inference that is both cryptographically verifiable and machine-auditable.

The provenance gap has been most visibly exploited in copyright litigation. The $1.5 billion *Bartz v. Anthropic* settlement (September 2025)—the largest publicly reported copyright recovery in US history—turned on whether training data was lawfully acquired. The certified class covers approximately 500,000 works. In *Thomson Reuters v. ROSS Intelligence* (February 2025), the court found that Westlaw headnotes are copyrightable and ROSS's fair use defense failed. Discovery in these cases demanded detailed training data provenance and model development process documentation—precisely the artifacts that cryptographic provenance binding would provide automatically.

The provenance gap extends beyond copyright. The Federal Reserve's SR 11-7 requires "conceptual soundness"—models must be justified by theory and documented with sufficient detail for independent review. The FDA's PCCP framework requires that modifications stay within pre-authorized parameters, with evidence of data management practices, performance evaluation, and cumulative drift monitoring. OMB M-26-04 requires model cards, system cards, and data cards for all federally procured LLMs. Each of these requirements demands provenance evidence. None specifies how that evidence should be cryptographically bound to the model in production.

### 3.4 The Compound Effect

These three gaps—identity, health, and provenance—are not independent failures. They compound. A system that cannot prove *which model* produced an output (identity), cannot demonstrate that the model was *healthy* when it did so (health), and cannot trace the model's *lineage* back to validated training data (provenance) offers **zero** verifiable governance. The current state of AI governance is analogous to financial markets before standardized auditing: the paperwork exists, but the evidence does not.

The enforcement landscape confirms the urgency. The SEC has brought seven AI-washing enforcement actions since March 2024. The FTC's Operation AI Comply yielded five simultaneous actions in September 2024. State attorneys general have emerged as the most consequential enforcers on algorithmic bias—the Massachusetts AG's $2.5 million settlement with Earnest Operations (July 2025) required a comprehensive AI governance structure including annual model inventories and disparate impact testing. Researcher Damien Charlotin's database tracks over 905 cases of AI-hallucinated legal citations globally, with frequency accelerating from roughly two per week before April 2025 to two to three per day. AI-related securities class actions exceed 53 filings between March 2020 and June 2025.

**Proposition 3.1** (Governance by Construction)**.** *Without a standardized attestation interface that cryptographically binds model identity, model health, and model provenance into a single verifiable artifact, AI governance is* ***unenforceable by construction****. Descriptive artifacts (model cards, training logs, static benchmarks) are necessary for documentation but insufficient for verification. MAI-1 closes this gap.*

# 4 Standards Landscape and the Composition Gap

MAI-1 does not invent new cryptographic primitives, new attestation token formats, or new trust architectures. It **composes** existing, individually validated standards into a normative profile that no existing system provides. This section surveys the relevant standards landscape as of February 2026 and identifies the specific composition gap that MAI-1 fills.

## 4.1 IETF RATS: The Attestation Infrastructure

The Internet Engineering Task Force's Remote ATtestation procedureS (RATS) working group has produced the most comprehensive framework for vendor-neutral attestation. Three published standards and several advanced drafts form the infrastructure upon which MAI-1 builds.

**RFC 9334** (RATS Architecture, January 2023) defines the foundational roles—Attester, Verifier, Relying Party, Endorser—and explicitly introduces the **Composite Device** concept (Section 3.3): a device composed of multiple sub-entities, each with its own Attesting Environment, where trustworthiness depends on appraising all sub-entities. A "Lead Attester" collects evidence from component attesters and conveys it to the verifier. This directly models the CPU+GPU attestation scenario required for foundation model inference.

**RFC 9711** (Entity Attestation Token, April 2025) provides the token format. EAT defines a flexible claims set serializable in CBOR (via CWT) or JSON (via JWT), with a **submodule mechanism** (Section 4.2.18) that accommodates composite devices through nested tokens. For foundation model attestation, the CBOR encoding is preferred due to its compactness and efficient handling of binary data. Key claims include `ueid` (universal entity ID, hardware-backed), `oemid` (manufacturer ID), `security-level` (isolation properties), `boot-seed` (ephemeral freshness), and `nonce` (replay prevention).

**RFC 9782** (EAT Media Types, May 2025) standardizes MIME types (`application/eat+cwt`, `application/eat+jwt`) enabling AI orchestration systems to automatically recognize and route attestation tokens.

The `measured-component` extension (`draft-ietf-rats-eat-measured-component`, IESG Evaluation stage, early 2026) introduces structured reporting of distinct software or data components within an Attester—replacing the file-system-centric view of prior standards with a **memory-centric view** critical for models that may never touch a disk in a confidential VM. Each measured component carries an identifier, a cryptographic measurement (hash or Merkle root), an `authorities` array linking runtime integrity to build-time signing authority, and profile-specific flags (encrypted, quantized, LoRA active, debug enabled).

**Concise Reference Integrity Manifests (CoRIM)** (`draft-ietf-rats-corim`, WGLC) define the format for distributing reference values—the "golden" measurements that verifiers use to appraise evidence. For foundation models, a CoRIM file contains reference triples mapping hardware environment, model identity, and allowed measurement hashes.

**Conceptual Message Wrappers (CMW)** (IESG-approved, December 2025) define **CMW Collections** explicitly for the composite attester use case—aggregating evidence from CPU TEE, GPU TEE, and accelerator TEE into a single self-describing message.

**EAT Attestation Results (EAR)** (`draft-ietf-rats-ear`, v01, July 2025) define how verifiers encode appraisal results, with a `submods` map giving each separately appraised attester its own entry and a **trustworthiness vector** covering eight appraisal facets (via AR4SI, `draft-ietf-rats-ar4si`, v09).

## 4.2 SCITT: Supply Chain Provenance

The IETF Supply Chain Integrity, Transparency, and Trust (SCITT) working group (`draft-ietf-scitt-archi` WGLC) defines "Transparent Statements"—signed claims attached to cryptographic receipts from append-only transparency logs. In the foundation model context, SCITT provides the

mechanism to bind runtime attestation evidence to build-time provenance: training data manifests, AI Bills of Materials (CycloneDX or SPDX format), and model signing certificates are registered in a SCITT ledger, and the resulting receipt is embedded in the EAT token as proof that the running model was logged by a trusted entity in the supply chain.

SCITT receipts enable **offline verification**—a relying party need not query the ledger in real time, as the receipt itself is cryptographic proof of registration. This is critical for air-gapped deployments in defense and critical infrastructure sectors.

### 4.3 CEN/CENELEC Harmonized Standards: The EU Regulatory Engine

The European standards engine is operating under extreme velocity. CEN/CENELEC Joint Technical Committee 21 (JTC 21), tasked under Standardization Request M/593 (amended by M/613), is delivering the harmonized standards that provide "presumption of conformity" for EU AI Act high-risk systems. An unprecedented "acceleration package" adopted in October 2025 authorized JTC 21 to bypass the Formal Vote for six priority drafts, compressing timelines by five to eight months. The drafts entering Enquiry in January–February 2026 are, for practical purposes, the final regulatory text.

**prEN 18282** (Cybersecurity Specifications, WG5, Enquiry active February 2026) operationalizes Article 15's cybersecurity requirements through a lifecycle-centric security model. The standard incorporates substantial content from the OWASP AI Exchange through a formal Liaison integration—the OWASP AI Exchange founder serves as Co-Editor. Controls address training phase integrity (data poisoning detection, provenance tracking), operational phase resilience (adversarial input detection, model extraction prevention), and alignment with the Cyber Resilience Act horizontal standards. MAI-1's Layer 1 (platform attestation) and Layer 3 (provenance binding) map directly to prEN 18282's requirements.

**prEN 18229-2** (Accuracy and Robustness, WG4, Enquiry launch February 11, 2026) standardizes the *methodology* of validation rather than prescribing universal accuracy thresholds. The standard requires justified metric selection, statistically disjoint test data, and systematic perturbation testing quantifying robustness as the accuracy delta between clean and perturbed inputs. MAI-1's Layer 2 (model state invariants) provides the continuous, cryptographically signed metrics that prEN 18229-2's methodology framework requires as evidence.

**prEN 18286** (Quality Management Systems, WG2, Enquiry closed December 2025) serves as the "connective tissue" linking all Article 11 documentation requirements. Its Clause 4.4 mandates a Technical File enforcing traceability from legal requirement to technical specification to verification evidence. MAI-1's attestation artifacts serve as the verification evidence layer in this traceability chain.

### 4.4 Composite CPU–GPU Attestation: The Hardware Gap

Unified attestation of CPU and GPU trusted execution environments remains fragmented across vendor silos. The closest production solution—Intel Trust Authority's composite JWT bundling TDX and NVIDIA GPU claims—operates only on Intel platforms and relies on two independent verification services (Intel Trust Authority + NVIDIA NRAS) rather than a truly unified cryptographic proof. AMD SEV-SNP platforms lack any equivalent composite attestation service.

Five architectural barriers prevent true unified attestation: **incompatible PKI hierarchies** (Intel, AMD, and NVIDIA maintain separate certificate chains with no cross-signing), **TOCTOU gaps** between CPU and GPU evidence collection (Intel TDX collateral has a 30-day validity window; NVIDIA NRAS JWTs have a 24-hour TTL), **absence of deployed TDISP** (current Hopper hardware uses software bounce buffers rather than hardware-secured PCIe channels), **asymmetric trust** (GPUs cannot verify the CPU TEE's state), and **verifier complexity** scaling poorly with heterogeneous environments.

The IETF RATS stack provides the data model for vendor-neutral composite tokens—CMW Collections for evidence aggregation, EAR `submods` for composite results, AR4SI for normalized trustworthiness vectors—but no finalized standard composes these into a conformance-testable profile for foundation model attestation. An individual draft (`draft-deshpande-rats-multi-verifier`, v03, October 2025, from Arm/Huawei/Fraunhofer SIT) directly addresses multi-TEE composite verification and explicitly names GPUs and Neural Processor Units as multi-vendor components, but has not yet been adopted as a working group document.

The Confidential Computing Consortium's Attestation SIG focuses on interoperable RA-TLS and coordinates with Veraison (the CCC-hosted open-source verification service), but explicitly considers "unification of attestation token formats" **out of scope**.

### 4.5 The AI Governance and Accountability Protocol (AIGA)

The most advanced application of RATS concepts to AI governance visible in the 2026 landscape is the AI Governance and Accountability Protocol (`draft-aylward-aiga`, active). AIGA operationalizes RATS into a coherent governance protocol featuring an "Immutable Kernel" (a TCB enforcing policy that the agent cannot modify), "Liveness Tokens" (short-lived attestation results serving as heartbeats), and tiered risk governance scaling from basic software attestation (Tier 1) to Multi-Vendor TEE Attestation (Tier 4, requiring simultaneous evidence from two different hardware TEEs).

AIGA demonstrates that the RATS infrastructure is sufficient for AI governance. However, AIGA defines a *protocol* for agent lifecycle management, not a *conformance-testable interface* for regulatory evidence generation. MAI-1 and AIGA are complementary: MAI-1 defines *what evidence must be produced and how it must be structured*, while AIGA defines *how agents are managed using that evidence*.

### 4.6 The Composition Gap

Table 1 summarizes the standards landscape and the specific gap each leaves open.

Table 1: Standards landscape and the composition gap MAI-1 fills.

| Standard/Framework | What It Provides | What It Does Not Provide |
|---|---|---|
| RFC 9711 (EAT) | Token format, claims, submodules | No ML-specific profile or required fields |
| `measured-component` | Component measurement structure | No composition into unified artifact |
| CoRIM | Reference value distribution | No model-specific reference triples |
| SCITT | Supply chain transparency receipts | No binding to runtime attestation |
| CMW Collections | Composite evidence aggregation | No conformance test suite |
| EAR / AR4SI | Composite attestation results | No regulatory evidence mapping |
| Intel Trust Authority | Production composite JWT | Vendor-specific; Intel-only |
| AIGA | Agent lifecycle governance | No conformance-testable interface |
| prEN 18282 | Cybersecurity controls | No attestation artifact specification |
| prEN 18229-2 | Accuracy/robustness methodology | No cryptographic evidence binding |
| MSAF (Fields, 2026) | Three-tier attestation architecture | No canonical endpoint or conformance suite |
| **MAI-1** | **Composes all of the above into a single, conformance-testable, regulatory-mapp** | |

The composition gap is the central observation of this specification. Every component required for foundation model attestation exists individually. No system composes them. MAI-1 is the normative profile that performs this composition—defining the required fields, the encoding format, the cryptographic binding rules, the conformance levels, and the regulatory evidence

mapping—so that implementers, regulators, insurers, and procurement officers can reference a single, authoritative interface.

**Proposition 4.1** (The Composition Thesis)**.** *Let $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ be the set of individually validated attestation components (EAT, measured-component, CoRIM, SCITT, CMW, EAR). The governance value of $\mathcal{C}$ deployed independently is bounded by the weakest link in the evidence chain. The governance value of $\mathcal{C}$ composed under a normative profile $\pi$ with conformance testing $\tau$ is multiplicative:*

$$V(\pi, \tau) \gg \sum_{i=1}^{n} V(C_i)$$

*MAI-1 defines $\pi$ and $\tau$.*

# 5 Normative Architecture Overview

MAI-1 composes three mandatory attestation layers, each operating at a different temporal frequency, trust level, and computational cost. The architecture maps directly to the three-tier structure defined in the Model State Attestation Framework (MSAF; Fields, 2026, §6) and to the RATS Composite Device model (RFC 9334, §3.3). No layer is optional. An attestation artifact missing any layer is invalid.

## 5.1 Layer 1: Platform Attestation (Hardware Root of Trust)

Layer 1 establishes that the execution environment is genuine and unmodified. It answers the question: *Is the hardware platform trustworthy?*

**Definition 5.1** (Platform Attestation)**.** *A cryptographic proof, rooted in on-die hardware keys, that the execution environment—CPU TEE, GPU TEE, firmware, and driver stack—is genuine, has booted a measured software image, and has not been tampered with since boot.*

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

**Content:** A TEE platform quote comprising the DICE/TPM root of trust measurement, firmware integrity verification, software stack hash, device identity signed by the manufacturer's certificate authority, and—for GPU-accelerated inference—the GPU attestation report (SPDM 1.1 measurement format) verified against the GPU manufacturer's Remote Attestation Service.

**Frequency:** Generated at boot, regenerated on any platform state change (firmware update, driver change, configuration modification).

**RATS Mapping:** Layer 1 evidence is encoded as EAT claims `ueid` (universal entity ID), `oemid` (manufacturer ID), `security-level`, and `boot-seed`, with the full platform quote carried as a `measured-component` entry for the TCB. In composite CPU+GPU deployments, Layer 1 produces a CMW Collection aggregating both TEE evidence bundles.

**MSAF Mapping:** MSAF Tier 1 (Fields, 2026, §6.2).

## 5.2 Layer 2: Model State Invariants (Continuous Health)

Layer 2 establishes that the model's internal health metrics are within certified bounds at execution time. It answers the question: *Is the model healthy right now?*

**Definition 5.2** (Model State Invariant)**.** *A continuously measured, cryptographically signed metric reflecting an internal property of the model that **MUST** remain within a certified envelope during operation. Breach of any invariant triggers automatic compliance status degradation.*

**Trust Root:** The measurement code executes within the TEE boundary established by Layer 1. Invariant values are computed by trusted code on verified hardware, preventing forgery.

**Content:** A vector of signed health metrics including (at minimum) entropy, gradient stability, distribution drift, and structural coherence. The specific invariants and their certified thresholds are defined in §7. Each metric is encoded as a `measured-component` entry with the `authorities` field binding the measurement to the governance authority that certified the thresholds.

**Frequency:** Computed at each inference event for MAI-C2 compliance; at configurable intervals (minimum once per hour) for MAI-C1 compliance. The frequency requirement ensures that health attestation is contemporaneous with the outputs it governs.

**RATS Mapping:** Layer 2 evidence is encoded as a set of `measured-component` entries within the EAT, each carrying a component identifier (invariant name), a measurement value (the metric), and an authorities array (the certifying entity). Profile-specific flags indicate operational mode (inference-only, fine-tuning, continuous learning).

**MSAF Mapping:** MSAF Tier 2 (Fields, 2026, §6.3).

### 5.3 Layer 3: Provenance Binding (Supply Chain Integrity)

Layer 3 establishes an unbroken, cryptographically verifiable chain of custody from training data through fine-tuning through deployment through the specific inference event. It answers the question: *Where did this model come from, and can I prove it?*

**Definition 5.3** (Provenance Binding)**.** *A cryptographic receipt, issued by a transparency service, proving that the model's identity, lineage, and associated metadata (AI Bill of Materials, training data manifest, signing certificates) were registered in an append-only ledger prior to deployment.*

**Trust Root:** The SCITT transparency service and its append-only ledger. The receipt is a COSE_Sign1 object serving as an inclusion proof—a Merkle tree path demonstrating that the model's signed statement was logged.

**Content:** A SCITT receipt binding the model's weights hash, configuration hash, and AI Bill of Materials (CycloneDX or SPDX format) to the transparency ledger. The receipt proves not merely that the model hash is "present" in memory, but that it was "registered" by a trusted entity in the supply chain. For models with Predetermined Change Control Plans (FDA PCCP), the receipt additionally binds the delta between the current model state and the last validated baseline.

**Frequency:** Generated at model registration; refreshed at each model update, fine-tuning event, or weight modification. The receipt is embedded in every MAI-1 attestation artifact as a static reference, enabling offline verification without real-time ledger queries.

**RATS Mapping:** Layer 3 evidence is encoded as a `scitt-receipt` claim within the EAT, containing the COSE_Sign1 inclusion proof. Reference values for the model hash are distributed via CoRIM, enabling verifiers to appraise the evidence against the registered "golden" measurement.

**MSAF Mapping:** MSAF Tier 3 (Fields, 2026, §6.4).

### 5.4 The Composition Principle

The three layers are not alternatives. They are not a menu from which implementers select. They are mandatory, concurrent, and cryptographically bound.

**Requirement 5.1** (Layer Composition)**.** *A valid MAI-1 attestation artifact **MUST** contain evidence from all three layers, bound under a single cryptographic root. An artifact missing any layer is **invalid** and **MUST NOT** be accepted by any conforming verifier, relying party, regulatory authority, or insurance underwriter.*

The composition is achieved through the EAT submodule mechanism (RFC 9711, §4.2.18): Layer 1, Layer 2, and Layer 3 evidence are encoded as submodules within a top-level EAT token, bound by a Merkle tree whose root hash is signed by the TEE-held attestation key and timestamped. This structure satisfies two competing requirements: **compactness** (the root hash is fixed-size regardless of the number of measurements) and **selective disclosure** (an auditor can verify a specific claim—e.g., a single invariant—without accessing the entire attestation history).

The specific composition protocol—including the Merkle tree construction algorithm, the inter-layer cryptographic binding mechanism, the submodule nesting structure, and the selective disclosure protocol for regulatory auditors—is specified in the private version of this framework.

# 6 The Model Attestation Interface (Normative)

This section defines the canonical MAI-1 attestation endpoint, the required evidence payload, and the encoding rules. This is the normative core of the specification. Implementations that do not conform to this section are non-compliant.

## 6.1 Canonical Endpoint

Every MAI-1 compliant system **MUST** expose a single attestation endpoint accepting evidence requests and returning signed attestation artifacts.

Listing 1: MAI-1 Canonical Endpoint

```
POST /v1/mai/attest
Content-Type: application/eat+cwt
Accept: application/eat+cwt

Request Body:
{
    nonce: bstr, ; Verifier-supplied freshness nonce
    requested_layers: uint, ; Bitmask: 0x01=L1, 0x02=L2, 0x04=L3
                            ; Default 0x07 (all layers)
    selective_disclosure: [ ; Optional: specific claims requested
        component_id: tstr
    ]
}
```

The endpoint **MUST** return a COSE-signed EAT token (`application/eat+cwt`) containing evidence from all requested layers. For MAI-C1 and MAI-C2 compliance (§9), the default request **MUST** include all three layers (bitmask `0x07`). Selective disclosure requests allow auditors to retrieve specific invariant measurements without requiring the full artifact, supporting bandwidth-constrained and privacy-preserving verification scenarios.

## 6.2 Required Response Payload

The MAI-1 response payload is a CBOR-encoded EAT token containing the following mandatory fields. The payload structure is defined using CDDL (Concise Data Definition Language) notation for precision, with a human-readable summary following each block.

### 6.2.1 Layer 1 Fields: Platform Attestation

Listing 2: Layer 1: Platform Attestation Fields

```
mai-platform-attestation = {
```

```
    ; Hardware identity
    ueid: bstr, ; RFC 9711 claim key 8
    oemid: bstr, ; RFC 9711 claim key 13
    security-level: security-level-type,

    ; TEE evidence
    tee-type: tee-type-enum,
    hardware-quote: bstr, ; Raw TEE platform quote
    firmware-measurements: [+ hash-entry],
    boot-seed: bstr, ; RFC 9711

    ; Composite GPU attestation (when applicable)
    ? gpu-attestation: {
        gpu-oemid: bstr,
        gpu-quote: bstr, ; SPDM 1.1 measurement
        gpu-cert-chain: [+ bstr], ; Certificate chain to mfr CA
        cc-mode: bool ; Confidential Computing enabled
    }
}

tee-type-enum = &(
    TDX: 1,
    SEV-SNP: 2,
    CC-GPU: 3,
    CCA: 4 ; Arm Confidential Compute Architecture
)

hash-entry = [
    algorithm-id: int, ; COSE algorithm registry
    hash-value: bstr
]
```

Layer 1 fields establish the hardware trust anchor. The `ueid` provides a permanent, hardware-backed identifier for the specific silicon executing the model. The `hardware-quote` is the raw TEE attestation report, verifiable against the manufacturer's certificate authority. When GPU-accelerated inference is used, the `gpu-attestation` substructure carries the GPU's independent attestation evidence, enabling composite verification even in the absence of a vendor-neutral composite attestation standard.

### 6.2.2 Layer 2 Fields: Model State Invariants

Listing 3: Layer 2: Model State Invariant Fields

```
mai-model-state = {
    ; Model identity
    model-identity: {
        weights-hash: hash-entry, ; SHA-256 or SHA-384
        config-hash: hash-entry,
        architecture-id: tstr,
        ? parameter-count: uint,
        ? quantization: tstr ; e.g., "INT8", "FP16"
    },

    ; Invariant measurements (see Section 6)
    invariants: {
        entropy: float, ; H >= H_min
        gradient-stability: float, ; Lyapunov exponent
```

```
        drift-kl: float, ; KL divergence from baseline
        coherence-energy: float, ; Dirichlet energy
        ? thermal-integrity: float ; SRAM thermal bound
    },

    ; Certified thresholds (reference values)
    thresholds: {
        entropy-floor: float,
        gradient-stability-max: float,
        drift-kl-max: float,
        coherence-energy-band: [float, float], ; [min, max]
        ? thermal-ceiling: float
    },

    ; Compliance determination
    compliance-status: compliance-flag,
    ? breached-invariant: tstr,
    ? breach-severity: severity-enum,
    ? remediation-action: tstr
}

compliance-flag = &(
    GREEN: 0, ; All invariants within certified bounds
    YELLOW: 1, ; Warning threshold crossed; monitoring escalated
    RED: 2 ; Hard invariant breach; automatic intervention
)

severity-enum = &(
    WARNING: 0,
    CRITICAL: 1,
    FATAL: 2
)
```

Layer 2 fields are the operational core of MAI-1. The `model-identity` block cryptographically identifies the specific model version. The `invariants` block carries the health metrics measured at execution time. The `thresholds` block carries the certified bounds against which the invariants are evaluated. The `compliance-status` flag provides an immediate, machine-readable governance determination: GREEN (compliant), YELLOW (warning—monitoring escalated), or RED (breach—automatic intervention triggered).

The inclusion of both measured values *and* certified thresholds in the same artifact enables any verifier to independently confirm the compliance determination. This prevents the "self-grading" problem endemic to current governance artifacts, where the entity asserting compliance also defines the criteria.

### 6.2.3 Layer 3 Fields: Provenance Binding

Listing 4: Layer 3: Provenance Binding Fields

```
mai-provenance = {
    ; SCITT receipt (inclusion proof)
    scitt-receipt: COSE_Sign1,

    ; AI Bill of Materials reference
    ai-bom: {
        format: bom-format-enum,
        bom-hash: hash-entry,
        registry-uri: tstr,
```

```
        ? bom-version: tstr
    },

    ; Training provenance summary
    training-provenance: {
        data-hash: hash-entry, ; Multiset hash of corpus
        training-run-id: tstr,
        ? fine-tune-delta: hash-entry, ; Hash of delta from base
        ? pccp-baseline: hash-entry ; FDA PCCP reference state
    }
}

bom-format-enum = &(
    CycloneDX: 1,
    SPDX: 2
)
```

Layer 3 fields close the provenance chain. The `scitt-receipt` is the cryptographic proof that the model's identity and metadata were registered in a transparency ledger. The `ai-bom` block references the AI Bill of Materials, enabling auditors to trace the model's components, dependencies, and licensing. The `training-provenance` block provides the cryptographic summary of the training corpus—not the data itself, but a multiset hash sufficient to prove consistency with a registered baseline.

The `pccp-baseline` field is specifically designed for FDA-regulated SaMD: it carries the hash of the last validated model state, enabling automated verification that modifications stayed within the Predetermined Change Control Plan's authorized parameters.

#### 6.2.4 Execution Context and Signature

Listing 5: Execution Context and Signature Fields

```
mai-execution-context = {
    ; Temporal binding
    inference-id: uuid, ; Unique per inference event
    timestamp: time, ; RFC 3339
    duration-ms: uint,

    ; Freshness
    nonce: bstr, ; Echo of verifier-supplied nonce

    ; Profile identification
    mai-profile: tstr, ; "MAI-1-v1.0"
    conformance-level: conformance-enum
}

conformance-enum = &(
    MAI-C0: 0, ; Research / Development
    MAI-C1: 1, ; Commercial Deployment
    MAI-C2: 2 ; Regulated / Insured / Federal
)

; Top-level MAI-1 artifact
mai-1-token = COSE_Sign1<{
    mai-platform-attestation,
    mai-model-state,
    mai-provenance,
    mai-execution-context
```

```
}>
```

The execution context binds the attestation to a specific moment in time. The `inference-id` is a UUID unique to each inference event, enabling post-incident forensic tracing. The `nonce` echoes the verifier-supplied freshness value, preventing replay attacks. The `mai-profile` and `conformance-level` fields identify which version of the MAI-1 specification and which conformance tier the artifact claims to satisfy.

The entire payload is wrapped in a COSE_Sign1 envelope, signed by the TEE-held attestation key. This signature binds all four sections—platform, model state, provenance, and execution context—into a single tamper-evident artifact.

## 6.3 Normative Encoding Rules

**Requirement 6.1** (Encoding)**.** *MAI-1 attestation artifacts **MUST** be encoded in CBOR (RFC 8949) and signed using COSE_Sign1 (RFC 9052). The signing key **MUST** be held within the TEE boundary established by Layer 1. JSON encoding (via JWT) is permitted for transport and display purposes but **MUST NOT** be treated as the canonical form for verification or archival.*

**Requirement 6.2** (Field Completeness)**.** *All fields defined in §6.2.1–§6.2.4 that are not marked with the CDDL optional operator (`?`) are **mandatory**. If any mandatory field is missing, the attestation artifact is **invalid**. There is no partial compliance. A conforming verifier **MUST** reject any artifact with missing mandatory fields, regardless of the validity of the remaining content.*

**Requirement 6.3** (Signing Key Binding)**.** *The COSE_Sign1 signature **MUST** be generated by a key that is: (a) generated within the TEE boundary, (b) certified by the hardware manufacturer's certificate authority, and (c) bound to the platform attestation evidence in Layer 1. A signature generated by a key outside the TEE boundary is invalid, as it cannot guarantee that the enclosed measurements were computed by trusted code in a verified environment.*

## 6.4 RATS Profile Alignment

MAI-1 is designed as a **RATS profile**—a constrained application of the EAT standard (RFC 9711) that specifies which claims are mandatory, which encoding is canonical, and how composite evidence is structured. Table 2 maps MAI-1 fields to their RATS standard origins.

Table 2: MAI-1 field mapping to RATS standards.

| MAI-1 Field | RATS Standard | Claim / Mechanism |
|---|---|---|
| `ueid` | RFC 9711 | Claim key 8 |
| `oemid` | RFC 9711 | Claim key 13 |
| `security-level` | RFC 9711 | Security level claim |
| `boot-seed` | RFC 9711 | Boot seed claim |
| `nonce` | RFC 9711 | Claim key 10 |
| `hardware-quote` | `measured-component` | Component measurement (TCB) |
| `model-identity` | `measured-component` | Component ID + measurement |
| `invariants` | `measured-component` | Measurement values (health metrics) |
| `thresholds` | CoRIM | Reference values |
| `scitt-receipt` | SCITT Architecture | COSE_Sign1 inclusion proof |
| `ai-bom` | SCITT / CycloneDX | Transparent statement payload |
| `gpu-attestation` | CMW Collection | Composite evidence submodule |
| Composite binding | EAT submodules | RFC 9711 §4.2.18 |
| Verification result | EAR / AR4SI | Trustworthiness vector |

This alignment ensures that MAI-1 artifacts are not a proprietary format but a **constrained profile** of international standards. Any RATS-conformant verifier can process MAI-1 tokens with a profile-specific configuration. This design decision prevents vendor lock-in and ensures that the interface can be implemented on any platform that supports the underlying RATS infrastructure—which, as of early 2026, includes all major cloud providers and TEE hardware vendors.

# 7 Mandatory Invariants and Thresholds

This section defines the model state invariants that every MAI-1 compliant system **MUST** measure, report, and enforce. Each invariant corresponds to an internal model property whose violation indicates degraded governance assurance. The invariants are drawn from the formal derivations in the Auburn Patent Family clause library; this section specifies *what must be reported*, not *how to compute it*. This separation is deliberate: it forces all implementations—regardless of architecture, framework, or vendor—through the same reporting interface, while preserving freedom in measurement methodology.

## 7.1 Invariant 1: Entropy Floor

**Definition 7.1** (Entropy Floor Invariant)**.** *Let $H(t)$ denote the Shannon entropy of the model's output distribution at time $t$, measured over a sliding window of $w$ inference events. The entropy floor invariant requires:*

$$H(t) \geq H_{\min} \quad \forall\, t \in [t_0, t_f]$$

*where $H_{\min}$ is the certified minimum entropy threshold for the model's intended operational domain.*

**Governance Rationale:** Entropy collapse indicates that the model has converged to a degenerate output distribution—producing repetitive, low-diversity responses regardless of input variation. In safety-critical contexts, entropy collapse means the model is no longer responsive to the input it is supposed to be processing. This invariant is formally derived in the Attention Thermodynamics framework (Clause AI-8, Auburn Patent Family; Fields, 2026).

**Breach Semantics:** If $H(t) < H_{\min}$, the MAI-1 compliance status **MUST** transition to RED. The `breached-invariant` field **MUST** report `"entropy-floor"` and the `remediation-action` field **MUST** specify the intervention triggered (e.g., inference suspension, fallback to verified checkpoint, human escalation).

**Threshold Calibration:** The value of $H_{\min}$ is domain-specific and **MUST** be established during model validation. A medical diagnostic system and a creative text generator will have different entropy floors. The certified threshold is carried in the `thresholds.entropy-floor` field of every MAI-1 artifact, enabling independent verification. The calibration methodology, including the statistical procedures for setting $H_{\min}$ from validation data, is specified in the private version of this framework.

## 7.2 Invariant 2: Gradient Stability

**Definition 7.2** (Gradient Stability Invariant)**.** *Let $\lambda_{\max}(t)$ denote the maximum Lyapunov exponent of the model's parameter dynamics at time $t$. The gradient stability invariant requires:*

$$\lambda_{\max}(t) \leq \lambda_{crit} \quad \forall\, t \in [t_0, t_f]$$

*where $\lambda_{crit}$ is the certified stability bound derived from the Lyapunov envelope analysis.*

**Governance Rationale:** A positive and growing Lyapunov exponent indicates that the model's parameter dynamics are diverging—small perturbations in input or state are being exponentially amplified rather than damped. In Mixture-of-Experts architectures, gradient instability manifests as expert starvation: individual expert sub-networks receive vanishing gradient signal, their parameters freeze, and routing collapses to a degenerate subset of the available capacity. This invariant is formally derived in the Gradient Starvation Envelope (Clause AI-2, Auburn Patent Family; Fields, 2026), which provides a Lyapunov-based stability proof satisfying the "conceptual soundness" requirement of SR 11-7.

**Breach Semantics:** If $\lambda_{\max}(t) > \lambda_{\text{crit}}$, the compliance status **MUST** transition to YELLOW (warning) or RED (critical), depending on the magnitude and duration of the breach. The Grönwall bound (Clause AI-2, Theorem 6.1) establishes the formal convergence guarantee: if the invariant is maintained, the system converges to a valid, equi-connected equilibrium.

**Threshold Calibration:** The value of $\lambda_{\text{crit}}$ is derived from the model's architecture and training configuration via the Lyapunov envelope analysis. For MoE architectures, the threshold incorporates the minimum acceptable expert utilization ratio $\delta$ and the recovery time constant $\tau$. The certified threshold is carried in `thresholds.gradient-stability-max`.

### 7.3 Invariant 3: Distribution Drift

**Definition 7.3** (Distribution Drift Invariant)**.** *Let $D_{KL}(P_t \| P_0)$ denote the Kullback–Leibler divergence between the model's current output distribution $P_t$ and its validated baseline distribution $P_0$. The drift invariant requires:*

$$D_{KL}(P_t \| P_0) \leq D_{\max} \quad \forall\, t \in [t_0, t_f]$$

*where $D_{\max}$ is the certified maximum permissible divergence from baseline.*

**Governance Rationale:** Distribution drift indicates that the model's behavior has shifted from its validated operating envelope. In continuously learning systems, drift may be intentional (adaptation to new data) or pathological (data poisoning, concept drift, catastrophic forgetting). In inference-only deployments, any significant drift from baseline indicates environmental change, adversarial manipulation, or hardware degradation. The drift invariant provides the continuous monitoring signal that SR 11-7 requires for ongoing model risk management and that FDA PCCP frameworks require for detecting when iterative modifications have cumulatively shifted the model beyond its validated parameters.

**Breach Semantics:** If $D_{\text{KL}}(P_t \| P_0) > D_{\max}$, the compliance status **MUST** transition to YELLOW. Sustained breach (exceeding $D_{\max}$ for a continuous period $\Delta t > \Delta t_{\text{crit}}$) triggers RED status. The `breached-invariant` field reports `"drift-kl"` and the artifact includes the current divergence value for forensic analysis.

**Threshold Calibration:** $D_{\max}$ is calibrated from the model's validation distribution using conformal prediction techniques to bound the expected natural variation. This approach aligns with Munich Re's aiSure methodology, which uses conformal prediction to quantify robustness in a model- and data-distribution-agnostic manner. The certified threshold is carried in `thresholds.drift-kl-max`.

### 7.4 Invariant 4: Structural Coherence

**Definition 7.4** (Structural Coherence Invariant)**.** *Let $E_D(t)$ denote the Dirichlet energy of the model's internal representational geometry at time $t$, measured over the activation manifold. The structural coherence invariant requires:*

$$E_D^{\min} \leq E_D(t) \leq E_D^{\max} \quad \forall\, t \in [t_0, t_f]$$

*where $[E_D^{\min}, E_D^{\max}]$ is the certified coherence band.*

**Governance Rationale:** Dirichlet energy quantifies the smoothness of the model's internal representations. Energy below the certified floor indicates representational collapse—the model has lost the capacity to distinguish between meaningfully different inputs. Energy above the certified ceiling indicates representational fragmentation—the model's internal geometry has become chaotically sensitive to small perturbations. Both conditions indicate that the model is operating outside the regime where its validated performance characteristics hold.

This invariant captures a dimension of model health that entropy, gradient stability, and drift do not: the *geometric structure* of the model's internal representations. A model can maintain stable entropy, bounded gradients, and low drift while its representational geometry degrades—particularly under adversarial attack or during extended fine-tuning on out-of-distribution data.

**Breach Semantics:** If $E_D(t) \notin [E_D^{\min}, E_D^{\max}]$, the compliance status **MUST** transition to YELLOW (boundary approach) or RED (hard breach). The `breached-invariant` field reports `"coherence-energy"`.

**Threshold Calibration:** The coherence band is established during model validation by measuring Dirichlet energy across the validation dataset and computing tolerance bounds. The certified band is carried in `thresholds.coherence-energy-band` as a two-element array $[E_D^{\min}, E_D^{\max}]$.

### 7.5 Invariant 5: SRAM Thermal Integrity (Conditional)

**Definition 7.5** (SRAM Thermal Integrity Invariant)**.** *Let $T_{SRAM}(t)$ denote the maximum junction temperature of the GPU's SRAM cache hierarchy at time $t$. The thermal integrity invariant requires:*

$$T_{SRAM}(t) \leq T_{ceil} \quad \forall\, t \in [t_0, t_f]$$

*where $T_{ceil}$ is the certified thermal ceiling derived from JEDEC and manufacturer specifications for the deployed silicon.*

**Governance Rationale:** SRAM thermal excursions cause silent bit-flip errors in the cache hierarchy, corrupting the very weight values and activation tensors that all other invariants measure. A model that reports healthy entropy, stable gradients, and low drift is providing *meaningless* attestation if the silicon computing those measurements is thermally compromised. The thermal integrity invariant is the "invariant of invariants"—it validates the physical substrate upon which all other measurements depend. This invariant is formally derived in the SRAM Thermal Integrity Bound (Clause AI-4, Auburn Patent Family; Fields, 2026), which establishes the relationship between junction temperature, bit-flip probability, and attestation validity.

**Conditionality:** This invariant is **mandatory** for MAI-C2 (regulated/insured/federal) conformance and **recommended** for MAI-C1 (commercial deployment). It is not required for MAI-C0 (research/development). The conditional status reflects the current state of hardware instrumentation: not all deployment environments provide real-time SRAM junction temperature telemetry at the granularity required for continuous attestation.

**Breach Semantics:** If $T_{\text{SRAM}}(t) > T_{\text{ceil}}$, the compliance status **MUST** transition to RED immediately. Thermal breach invalidates all concurrent invariant measurements. The `breached-invariant` field reports `"thermal-integrity"` and the `breach-severity` **MUST** be FATAL.

**Threshold Calibration:** $T_{\text{ceil}}$ is derived from JEDEC Standard JESD79-5B (DDR5 SDRAM) thermal specifications and manufacturer-specific reliability data for the deployed GPU silicon. The certified threshold is carried in `thresholds.thermal-ceiling`.

### 7.6 Invariant Summary

Table 3 summarizes the five mandatory invariants, their formal conditions, breach semantics, and Auburn Clause references.

Table 3: MAI-1 mandatory invariants summary.

| Invariant | Condition | Breach | Clause | Governance Function |
|---|---|---|---|---|
| Entropy Floor | $H(t) \geq H_{\min}$ | RED | AI-8 | Detects output distribution collapse |
| Gradient Stability | $\lambda_{\max} \leq \lambda_{\text{crit}}$ | Y/R | AI-2 | Detects parameter divergence and expert starvation |
| Distribution Drift | $D_{\text{KL}} \leq D_{\max}$ | Y/R | — | Detects behavioral shift from validated baseline |
| Structural Coherence | $E_D \in [E_D^{\min}, E_D^{\max}]$ | Y/R | — | Detects representational geometry degradation |
| Thermal Integrity | $T_{\text{SRAM}} \leq T_{\text{ceil}}$ | RED | AI-4 | Validates physical substrate of all measurements |

**Requirement 7.1** (Invariant Reporting)**.** *Every MAI-1 attestation artifact **MUST** report the current measured value of each applicable invariant and its certified threshold. The compliance determination **MUST** be independently verifiable by any party possessing the artifact. Attestation artifacts that report compliance status without including the underlying measurements and thresholds are invalid.*

This requirement eliminates the "trust the auditor" problem. A regulator, insurer, or court does not need to trust the deployer's compliance assertion. They can verify it themselves from the signed measurements and certified thresholds contained in the artifact.

## 8 Cryptographic Binding Rules

The attestation artifact defined in §6 establishes the model's governance state at a specific moment. This section defines how that attestation is **bound** to the model's outputs, creating an immutable evidentiary chain from inference event to governance proof. Without this binding, the attestation floats free of the outputs it is supposed to govern—a signed health certificate with no connection to the patient.

### 8.1 Decision Receipt Construction

**Definition 8.1** (Decision Receipt)**.** *A Decision Receipt is a compact, tamper-evident artifact that binds a specific model output to the MAI-1 attestation artifact that was valid at the time the output was produced. The receipt enables any holder to verify, after the fact, that a specific output was generated by an identified model in a certified internal state on a verified platform.*

Each inference output **MUST** be accompanied by a Decision Receipt containing the following fields:

Listing 6: Decision Receipt Structure

```
mai-decision-receipt = {
    ; Output binding
    output-hash: hash-entry, ; Hash of the model output
    inference-id: uuid, ; Matches attestation artifact

    ; Attestation reference
```

```
    attestation-root: hash-entry, ; Merkle root of MAI-1 artifact
    attestation-uri: tstr, ; Retrieval location

    ; Temporal binding
    timestamp: time, ; RFC 3339
    nonce: bstr, ; Freshness proof

    ; Compliance snapshot
    compliance-status: compliance-flag,

    ; Signature
    signature: COSE_Sign1 ; Same TEE-held key as artifact
}
```

The `attestation-root` is the Merkle root hash of the full MAI-1 artifact. This single fixed-size value commits to the entire governance state—platform identity, model health, provenance chain—without requiring the full artifact to be transmitted with every output. The `attestation-uri` provides a retrieval location for the full artifact when forensic analysis is required.

The `inference-id` in the Decision Receipt **MUST** match the `inference-id` in the corresponding MAI-1 attestation artifact. This cross-reference creates a bidirectional link: from any output, one can retrieve its governance proof; from any attestation artifact, one can identify all outputs it governed.

**Requirement 8.1** (Receipt Signing)**.** *The Decision Receipt **MUST** be signed by the same TEE-held attestation key that signed the MAI-1 artifact. This ensures that the binding between output and governance state was created within the trusted execution boundary and cannot be forged by an external party. A receipt signed by a different key than its referenced attestation artifact is invalid.*

## 8.2 Tamper Resistance

The cryptographic binding between outputs and attestation artifacts **MUST** satisfy three tamper-resistance properties:

**Integrity:** Any modification to the model output, the attestation artifact, or the Decision Receipt after signing **MUST** be detectable through signature verification failure. The COSE_Sign1 envelope provides this property.

**Non-repudiation:** The deploying entity cannot deny that a specific output was produced by a specific model in a specific state. The TEE-rooted signing key, certified by the hardware manufacturer's CA, provides non-repudiation—the key provably existed only within the TEE boundary at the time of signing.

**Replay Prevention:** The verifier-supplied nonce, echoed in both the attestation artifact and the Decision Receipt, prevents an attacker from replaying old evidence from a previously compliant state while currently running a compromised model. A replayed artifact will fail nonce verification.

**Requirement 8.2** (Mismatch Invalidation)**.** *Any mismatch between the Decision Receipt and its referenced attestation artifact—including `inference-id` mismatch, `nonce` mismatch, signing key mismatch, or temporal inconsistency (receipt timestamp outside the attestation artifact's validity window)—**MUST** void the compliance determination. A voided determination **MUST** be treated as equivalent to RED (non-compliant) status by any conforming verifier.*

### 8.3 Retention Requirements

MAI-1 attestation artifacts and Decision Receipts are governance evidence with legal significance. Their retention is subject to the most stringent applicable jurisdiction.

Table 4: Minimum retention periods by regulatory framework.

| **Framework** | **Minimum Retention** | **Basis** |
|---|---|---|
| EU AI Act (Art. 11, 12) | 10 years | Technical documentation and automatic logging |
| FDA SaMD (QMS/QMSR) | Device lifetime + 2 years | Quality management system records |
| SR 11-7 | 7 years (typical) | Model risk management documentation |
| SOX (financial AI) | 7 years | Audit trail requirements |
| GDPR (Art. 5(1)(e)) | Purpose-limited | Data minimization; retain governance evidence, purge personal data |
| Federal procurement | Contract period + 3 years | FAR record retention |
| Insurance | Policy period + statute of limitations | Claims evidence preservation |

**Requirement 8.3** (Retention Policy)**.** *MAI-1 compliant systems **MUST** retain attestation artifacts and Decision Receipts for the longest applicable retention period among all governing frameworks. In multi-jurisdictional deployments, the retention period **SHALL** default to the maximum across all applicable jurisdictions. The retention mechanism **MUST** preserve the cryptographic verifiability of the artifacts—archived artifacts must remain signature-verifiable throughout the retention period, requiring key management practices that account for cryptographic algorithm deprecation and certificate chain validity.*

### 8.4 Certificate Transparency Alignment

The Decision Receipt architecture follows the precedent established by Certificate Transparency (CT) in the TLS ecosystem. CT converted certificate issuance transparency from a moral virtue into a technical requirement for connectivity: since April 2018, Google Chrome mandates that all new certificates be logged in publicly auditable, append-only logs, verified via Signed Certificate Timestamps (SCTs) during the TLS handshake.

MAI-1 adopts the same structural logic: the SCITT receipt embedded in Layer 3 is the AI governance analog of the SCT. Just as a TLS certificate without a CT log entry is untrusted by browsers, a model without a SCITT-registered provenance chain is ungoverned by MAI-1. The Decision Receipt extends this logic to individual inference events: just as CT ensures that every certificate is publicly auditable, MAI-1 ensures that every consequential AI output is governance-auditable.

The historical trajectory is instructive. The TLS ecosystem's transition from voluntary certificate management to mandatory CT logging took approximately four years (2014–2018), driven by browser enforcement rather than regulatory mandate. The AI governance ecosystem is following an accelerated version of the same trajectory, driven by concurrent regulatory (EU AI Act), judicial (AI discovery orders), and market (insurance exclusion) pressure. MAI-1 is positioned to serve the same structural role that the Baseline Requirements and CT played for TLS: the technical specification that transforms governance from aspiration into infrastructure.

# 9 The MAI-1 Conformance Test Suite

The MAI-1 Conformance Test Suite is the enforcement mechanism of this specification. Without conformance testing, a normative interface is merely a suggestion. With it, the interface becomes a **gatekeeper artifact**: a system either passes or it does not, and the result is deterministic, reproducible, and machine-verifiable.

The design of the conformance suite draws on three established security certification frameworks, each selected for a specific structural precedent:

- **FIPS 140-3** (Cryptographic Module Validation Program): provides the precedent for strictly binary pass/fail semantics, mandatory field completeness, and multi-level security tiers with escalating physical and logical requirements. FIPS 140-3 validation averages 542 days and costs $50,000–$200,000+ per module configuration—MAI-1 conformance is designed to be significantly faster and less expensive while preserving the binary rigor.
- **Common Criteria** (ISO/IEC 15408): provides the precedent for Evaluation Assurance Levels (EALs) scaling from lightweight functional testing to formal mathematical verification, Protection Profiles defining reusable security requirements for product classes, and the cascade verdict model (Pass/Inconclusive/Fail resolving to a binary outcome).
- **TCG TPM Certification**: provides the precedent for automated functional compliance test suites with normative reference code, combined with mandatory third-party evaluation at higher assurance levels.

MAI-1 synthesizes these precedents into a conformance framework calibrated for the unique requirements of AI governance: attestation artifacts that are generated continuously (not once at certification time), verified by automated systems (not human auditors at every event), and mapped to multiple regulatory frameworks simultaneously (not a single-jurisdiction standard).

## 9.1 Conformance Levels

MAI-1 defines three conformance levels, each representing an escalating assurance tier. The levels are cumulative: MAI-C1 includes all MAI-C0 requirements, and MAI-C2 includes all MAI-C1 requirements.

### 9.1.1 MAI-C0: Research and Development

MAI-C0 is the entry-level conformance tier, designed for pre-deployment environments where organizations are integrating MAI-1 into their development pipelines.

Table 5: MAI-C0 conformance requirements.

| Requirement | Specification |
|---|---|
| Field completeness | All mandatory fields present in attestation artifact |
| Encoding validity | CBOR encoding conforms to RFC 8949; COSE_Sign1 envelope parseable |
| Signature verification | Signature verifiable against a declared public key (TEE binding not required) |
| Invariant reporting | All four base invariants reported with measured values and thresholds |
| Compliance flag logic | GREEN/YELLOW/RED determination consistent with reported values and thresholds |
| Nonce echo | Verifier-supplied nonce correctly echoed in response |
| Verification model | Self-attestation permitted |
| SRAM thermal invariant | Recommended, not required |

MAI-C0 establishes that an implementation can produce *well-formed* artifacts. It does not verify that the measurements are *trustworthy*—the signing key need not be TEE-held, and self-attestation is permitted. MAI-C0 is **not sufficient** for regulatory compliance, insurance underwriting, or procurement eligibility.

### 9.1.2 MAI-C1: Commercial Deployment

MAI-C1 is the standard conformance tier for commercial AI systems deployed in contexts where governance evidence is expected by customers, partners, or market regulators.

Table 6: MAI-C1 conformance requirements (cumulative with MAI-C0).

| **Requirement** | **Specification** |
|---|---|
| TEE-rooted signing | Signing key **MUST** be generated within and held by a hardware TEE; certificate chain verifiable to manufacturer CA |
| Platform attestation | Layer 1 hardware quote verifiable against manufacturer's attestation service |
| Invariant measurement frequency | Minimum once per hour during active operation |
| Decision Receipt generation | Receipts generated for all inference events; output-to-attestation binding verifiable |
| Provenance binding | SCITT receipt present and verifiable against a registered transparency service |
| Third-party verification | Attestation artifacts **MUST** be verified by an independent verification service (not self-attested) |
| Replay resistance | Nonce freshness verified; replayed artifacts rejected |
| SRAM thermal invariant | Recommended |
| Retention | Minimum 7 years |

MAI-C1 ensures that attestation artifacts are *cryptographically trustworthy*: measurements are computed by verified code on verified hardware, signed by TEE-held keys, and verified by independent services. The third-party verification requirement eliminates self-grading. MAI-C1 is the minimum tier that satisfies commercial insurance underwriting requirements and enterprise vendor risk management (SOC 2+ equivalent for AI governance).

### 9.1.3 MAI-C2: Regulated, Insured, and Federal

MAI-C2 is the highest conformance tier, required for AI systems operating in regulated sectors, covered by specialist AI insurance policies, or deployed under federal procurement authority.

Table 7: MAI-C2 conformance requirements (cumulative with MAI-C1).

| Requirement | Specification |
|---|---|
| Per-inference attestation | Attestation artifact generated for every inference event (not sampled) |
| SRAM thermal invariant | Mandatory; continuous monitoring with FATAL breach semantics |
| Composite TEE verification | GPU attestation evidence independently verified; composite evidence aggregated per CMW Collection format |
| Invariant breach simulation | Conformance testing **MUST** include injection of simulated invariant breaches to verify correct detection, status transition, and remediation response |
| Hardware quote mismatch test | Conformance testing **MUST** include presentation of invalid or expired hardware quotes to verify rejection |
| Clock skew tolerance test | Attestation artifacts with timestamps exceeding configurable skew tolerance **MUST** be rejected |
| Replay attack simulation | Conformance testing **MUST** include presentation of previously valid artifacts with stale nonces to verify rejection |
| Tamper detection test | Conformance testing **MUST** include presentation of artifacts with modified fields (bit-flip in invariant values, altered compliance flags) to verify signature failure detection |
| Decision Receipt cross-reference | Bidirectional linkage between receipts and attestation artifacts verified for completeness |
| Regulatory evidence package | Full attestation artifact **MUST** be exportable in a format satisfying the regulatory mapping defined in §10 |
| Retention | Maximum applicable period per Table 4; cryptographic verifiability preserved throughout |

MAI-C2 adds *adversarial testing* to the conformance suite. It is not sufficient to demonstrate that a system produces correct artifacts under normal conditions; the system **MUST** demonstrate correct behavior under attack. The invariant breach simulation, hardware quote mismatch, replay attack, and tamper detection tests verify that the system fails safely—detecting and rejecting compromised evidence rather than silently accepting it.

This adversarial testing requirement reflects the threat model established by prEN 18282 (EU AI Act cybersecurity specifications): AI systems must be resilient not only against operational failures but against deliberate manipulation of their governance infrastructure.

## 9.2 Pass/Fail Semantics

**Requirement 9.1** (Binary Compliance)**.** *MAI-1 conformance is **strictly binary**. A system either passes all requirements for its claimed conformance level or it does not. There is no partial compliance, conditional certification, or "substantially conformant" status.*

This design decision follows the precedent of FIPS 140-3 (where a module's overall security level equals the minimum level achieved across all requirement areas) and Common Criteria (where the overall verdict is Pass if and only if every constituent verdict is Pass). The rationale is identical: partial compliance creates ambiguity that regulators, insurers, and procurement officers cannot resolve. A system is either governable or it is not.

When a conformance test identifies a deficiency, the implementer may remediate and resubmit. The conformance process is iterative, but the *result* is binary. This mirrors the FIPS 140-3 Coordination phase, where CSTLs work with vendors to remediate deficiencies before submission to CMVP, but the final validation decision is strictly pass or fail.

## 9.3 Conformance Verification Roles

Three roles participate in MAI-1 conformance verification:

**The Implementer** builds the MAI-1 compliant system, integrates the attestation endpoint, and submits for conformance testing. The Implementer is analogous to the vendor in FIPS 140-3 or the developer in Common Criteria.

**The Conformance Testing Entity (CTE)** executes the conformance test suite against the Implementer's system. For MAI-C0, the CTE may be the Implementer itself (self-attestation). For MAI-C1 and MAI-C2, the CTE **MUST** be an independent third party with no financial relationship to the Implementer other than the testing engagement. This independence requirement follows the FIPS 140-3 model (NVLAP-accredited laboratories) and Common Criteria model (accredited evaluation facilities). The accreditation criteria for CTEs, including technical competency requirements, conflict-of-interest rules, and audit procedures, are specified in the private version of this framework.

**The Governance Authority** reviews the CTE's test results and issues the conformance determination. For deployments under regulatory obligation, the Governance Authority may be a Notified Body (EU AI Act), a regulatory agency (FDA, OCC), or an insurance underwriter (Armilla AI, Munich Re, AIUC). The Governance Authority does not re-execute tests; it verifies that the CTE's procedures were correct and that the results support the claimed conformance level.

### 9.4 Conformance Artifact

Upon successful conformance testing, the CTE produces a **MAI-1 Conformance Report** containing:

1. The claimed conformance level (MAI-C0, MAI-C1, or MAI-C2).
2. A complete enumeration of tests executed, with pass/fail results for each.
3. The system identification (hardware platform, TEE type, model architecture, software stack version).
4. The CTE's identity and accreditation status.
5. A validity period (recommended: 12 months for MAI-C1, 6 months for MAI-C2), after which re-testing is required.
6. A digital signature by the CTE.

The Conformance Report is itself a governance artifact—it can be referenced in procurement responses, insurance applications, regulatory submissions, and vendor risk management questionnaires. Its structure is deliberately aligned with the SOC 2 Type II report model: a time-bounded assurance of operational effectiveness, not a point-in-time snapshot.

## 10 Regulatory and Insurance Evidence Mapping

This section maps MAI-1 attestation artifacts to specific evidence requirements across seven regulatory frameworks, three harmonized standards currently in CEN/CENELEC Enquiry, and the emerging AI insurance ecosystem. The mapping is designed to be directly actionable: a compliance officer, insurance underwriter, or contracting officer can use this section to determine exactly which MAI-1 fields satisfy which requirements.

### 10.1 Regulatory Framework Mapping

Table 8: MAI-1 evidence mapping to regulatory and insurance frameworks.

| **Framework** | **Article / Requirement** | **Evidence Required** | **MAI-1 Artifact** |
|---|---|---|---|
| EU AI Act | Art. 10 (Data Governance) | Training data documentation; contamination bounds | Layer 3: `training-provenance` (multiset hash of corpus); `ai-bom` (data lineage) |
| EU AI Act | Art. 11 (Technical Documentation, Annex IV) | Design specifications; training methodologies; robustness measures; validation results | Full artifact: Layer 2 (invariants as robustness evidence); Layer 3 (provenance as training documentation); Conformance Report as validation record |
| EU AI Act | Art. 12 (Record-Keeping) | Automatic logging of system operation | Decision Receipt chain: continuous, signed, tamper-evident log of all inference events with governance state binding |
| EU AI Act | Art. 15(1)(3)(4) (Accuracy, Robustness, Cybersecurity) | Appropriate accuracy levels; resilience against errors and manipulation; anti-tampering measures | Layer 1 (anti-tampering via TEE); Layer 2 (accuracy/robustness via invariants); Conformance Report (adversarial testing at MAI-C2) |
| prEN 18282 | Cybersecurity specifications (WG5) | Lifecycle security controls; adversarial resilience; data integrity | Layer 1 (platform integrity); Layer 3 (data provenance); MAI-C2 adversarial test results |
| prEN 18229-2 | Accuracy & robustness methodology (WG4) | Metric selection; test data integrity; perturbation testing protocols | Layer 2 (invariant metrics as continuous accuracy evidence); `thresholds` (certified bounds as methodology documentation) |
| prEN 18286 | QMS for AI Act (WG2), Clause 4.4 | Technical File with traceability from requirement to specification to verification evidence | Full artifact provides the verification evidence tier in the traceability chain mandated by Clause 4.4 |

*Continued on next page*

| Framework | Article / Requirement | Evidence Required | MAI-1 Artifact |
|---|---|---|---|
| SR 11-7 | Conceptual Soundness | Theory-based validation; sufficiently detailed documentation for independent review | Layer 2 (Lyapunov-based gradient stability as formal theoretical grounding); Conformance Report (independent third-party verification) |
| SR 11-7 | Ongoing Monitoring | Stability metrics (PSI and equivalents); periodic review | Layer 2 (continuous signed invariant stream: entropy, drift, coherence); Decision Receipts (continuous audit trail) |
| SR 11-7 | Effective Challenge | Independent, objective review of model decisions without proprietary access | Layer 2 + Layer 3 (external validators verify attestation artifacts without accessing model weights—`selective_disclosure` enables weight-private validation) |
| FDA SaMD PCCP | Data Management Evidence | Retraining data matches documented corpus from required multiple clinical sites | Layer 3: `training-provenance.data-hash` (multiset hash proves corpus consistency) |
| FDA SaMD PCCP | Performance Evaluation | Model health within validated parameters after each modification | Layer 2 (invariant measurements after each update); `pccp-baseline` hash (delta from validated baseline) |
| FDA SaMD PCCP | Cumulative Drift | Detection when iterative modifications shift model beyond validated envelope | Layer 2: `drift-kl` (continuous KL divergence monitoring against `pccp-baseline`) |
| FDA SaMD PCCP | Audit Trail (QMSR) | QMS documentation; ISO 13485 transition | Full attestation chain as QMS documentation; Decision Receipt chain as audit trail |

*Continued on next page*

| Framework | Article / Requirement | Evidence Required | MAI-1 Artifact |
|---|---|---|---|
| OMB M-25-21 | High-Impact AI: Pre-deployment testing | Testing even without access to source code or data | Conformance Report (third-party verification of attestation infrastructure); Layer 2 (invariant baseline established at validation) |
| OMB M-25-21 | High-Impact AI: Ongoing monitoring | Performance degradation and adverse impact monitoring | Layer 2 (continuous invariant stream); Decision Receipt chain (longitudinal governance record) |
| OMB M-25-21 | High-Impact AI: Human oversight | Intervention capability | Layer 2: `compliance-status` RED triggers documented `remediation-action` including human escalation pathways |
| OMB M-26-04 | LLM Transparency: Model/system/-data cards | Model cards, system cards, data cards | Layer 2: `model-identity` (architecture, parameters, quantization); Layer 3: `ai-bom` + `training-provenance` |
| OMB M-26-04 | LLM Enhanced Transparency: Bias evaluation, benchmarks, red-teaming | Bias evaluation results; benchmark scores; red-teaming documentation | Conformance Report (MAI-C2 adversarial testing as structured red-teaming evidence); Layer 2 (benchmark-equivalent invariant measurements) |
| FY2026 NDAA §1513 | AI/ML cybersecurity framework: adversarial tampering, model tampering, prompt injection, supply chain risks | Comprehensive security framework leveraging NIST SP 800 series | Layer 1 (anti-tampering); Layer 2 (model integrity monitoring); Layer 3 (supply chain provenance); MAI-C2 adversarial test suite |

*Continued on next page*

| Framework | Article / Requirement | Evidence Required | MAI-1 Artifact |
|---|---|---|---|
| FY2026 NDAA §1533 | Standardized framework for evaluating, procuring, and overseeing AI models | Assessment of all major DoD AI systems by January 2028 | MAI-1 Conformance Report as standardized evaluation artifact; procurement language (§11) as contract integration |

### 10.2 Insurance Evidence Mapping

The AI insurance market's structural shift from silent coverage to explicit exclusion creates an urgent demand for governance evidence that current artifacts cannot supply. MAI-1 attestation artifacts are designed to serve as the **underwriting evidence package** that specialist AI insurers require.

Table 9: MAI-1 evidence mapping to AI insurance underwriting requirements.

| Insurer / Standard | Underwriting Requirement | MAI-1 Evidence |
|---|---|---|
| Armilla AI | Model performance and robustness testing (50+ automated tests across 6 dimensions) | Layer 2 invariants provide continuous robustness evidence; MAI-C1/C2 Conformance Report provides structured test documentation equivalent to Armilla's FingerPrint™ validation |
| Armilla AI | AI system inventories; controls documentation; human oversight protocols | Full MAI-1 artifact as system-level governance evidence; `compliance-status` with `remediation-action` as oversight protocol documentation |
| Armilla AI / A-LIGN | ISO/IEC 42001 certification with preferential insurance terms | MAI-1 attestation artifacts provide the cryptographic backing that strengthens ISO 42001 from self-reported documentation to verifiable evidence |
| Munich Re aiSure | Technical due diligence: model architecture, data pipeline, performance monitoring, quality management | Layer 2: `model-identity` (architecture); Layer 3: `training-provenance` (pipeline); Layer 2: continuous invariants (monitoring); Conformance Report (quality management evidence) |
| Munich Re aiSure | Predictive robustness via conformal prediction; historical performance data | Layer 2: `drift-kl` calibrated via conformal prediction methodology; Decision Receipt chain provides longitudinal performance history |

*Continued on next page*

| Insurer / Standard | Underwriting Requirement | MAI-1 Evidence |
|---|---|---|
| Testudo | External AI risk profile assessment; litigation-data-driven pricing | MAI-1 Conformance Report as standardized risk profile artifact; `compliance-status` history as claims-relevant governance evidence |
| AIUC-1 | Six-pillar certification (security, safety, reliability, data/privacy, accountability, societal risks) | Layer 1 maps to security pillar; Layer 2 maps to safety and reliability pillars; Layer 3 maps to data/privacy and accountability pillars; Conformance Report maps to certification audit |
| Lloyd's LMA | Acceptable use policies; AI tool inventories; human oversight; training records; performance monitoring baselines | Full MAI-1 deployment documentation satisfies LMA September 2025 E&O underwriting guidance; Layer 2 invariant baselines serve as monitoring benchmarks |
| General (all carriers) | Governance maturity signal for premium calibration | MAI-1 conformance level (C0/C1/C2) serves as a graduated governance maturity signal—analogous to safe-driver discounts—enabling better terms as conformance level increases |

The insurance mapping reflects a critical insight from the specialist AI insurance market: underwriters need **continuous, cryptographically signed evidence**, not point-in-time documentation. Munich Re's conformal prediction methodology, Armilla's continuous monitoring via Trustible integration, and AIUC-1's certification-to-premium pipeline all converge on the same requirement—the requirement that MAI-1's Layer 2 invariant stream and Decision Receipt chain are designed to satisfy.

**Proposition 10.1** (Epistemic to Actuarial)**.** *MAI-1 compliance transforms AI insurance underwriting from epistemic uncertainty ("we believe this system is governed") to actuarial risk quantification ("this system's attestation history demonstrates a quantifiable failure probability"). The transition from epistemic to actuarial is the structural prerequisite for the AI insurance market to scale from its current approximately $80 million in annual premiums to the $4.7 billion projected by 2032.*

# 11 Procurement Language

This section provides copy-pasteable contract language enabling procurement officers, insurance underwriters, and vendor risk management teams to mandate MAI-1 compliance without requiring bespoke technical evaluation. The clauses are designed to be inserted directly into solicitations, master service agreements, insurance questionnaires, and vendor assessment frameworks. Each clause is self-contained and jurisdiction-neutral.

The historical precedent is precise. PCI-DSS became mandatory not through legislation but through modification of the Merchant Agreement—the foundational contract between a merchant and their acquiring bank. SOC 2 became mandatory not through regulation but through Vendor Risk Management policies at enterprise buyers. TLS certificate standards

became mandatory not through statute but through Browser Root Store Policies. In each case, procurement language was the vector through which a technical standard became a business requirement. MAI-1 follows the same structural logic.

### 11.1 Federal and Government Procurement

The following clause is designed for insertion into federal Requests for Proposal (RFPs), Statements of Work (SOWs), and contract modifications under OMB M-25-22 (AI acquisition) and OMB M-25-21 (high-impact AI governance). It is compatible with the anticipated DFARS amendments under FY2026 NDAA §1513 and the standardized evaluation framework mandated by §1533.

> ***AI Governance Attestation Requirement.*** *All artificial intelligence systems delivered under this contract that meet the applicability criteria defined in the Model Attestation Interface specification (MAI-1, Clause AI-5, Auburn Patent Family)* ***SHALL*** *expose an MAI-1 compliant attestation endpoint. The Contractor* ***SHALL*** *provide, at minimum, MAI-C1 conformance for commercial deployments and MAI-C2 conformance for systems classified as high-impact AI under OMB M-25-21 or subject to the cybersecurity framework requirements of FY2026 NDAA §1513. The Contractor* ***SHALL*** *deliver a current MAI-1 Conformance Report from an independent Conformance Testing Entity as part of the deliverable package. Systems unable to provide valid MAI-1 attestation receipts are ineligible for acceptance. The Government reserves the right to verify attestation artifacts independently at any time during the contract period and for the duration of the applicable retention period.*

### 11.2 EU Conformity Assessment

The following clause is designed for insertion into conformity assessment documentation, Notified Body engagement letters, and technical file submissions under the EU AI Act. It maps to the presumption of conformity mechanism (Article 40) and the QMS documentation requirements of prEN 18286 (Clause 4.4).

> ***Attestation Evidence for Conformity Assessment.*** *The Provider* ***SHALL*** *demonstrate compliance with Articles 11, 12, and 15 of Regulation (EU) 2024/1689 by maintaining continuous MAI-1 compliant attestation at conformance level MAI-C2 for all high-risk AI systems within scope. The Technical File required under Article 11 and Annex IV* ***SHALL*** *include the MAI-1 attestation artifact schema, the current Conformance Report, and a representative sample of Decision Receipts covering the most recent reporting period. The Provider* ***SHALL*** *make the MAI-1 attestation endpoint available to the Notified Body for independent verification during the conformity assessment procedure. Attestation artifacts* ***SHALL*** *be retained for a minimum of ten years in accordance with Article 11(1) and* ***SHALL*** *remain cryptographically verifiable throughout the retention period.*

### 11.3 Insurance Underwriting

The following clause is designed for insertion into AI insurance applications, underwriting questionnaires, and policy terms. It is compatible with the governance-to-premium pipeline established by the A-LIGN/Armilla AI partnership, the AIUC-1 certification framework, and the Munich Re aiSure technical due diligence process.

> ***Governance Evidence for AI Insurance.*** *The Applicant* ***SHALL*** *provide evidence of MAI-1 conformance at the level appropriate to the risk classification of the insured AI system: MAI-C1 for standard commercial deployments, MAI-C2 for regulated, safety-critical, or high-exposure systems. Required evidence includes: (a) a current MAI-1 Conformance Report from an independent Conformance Testing Entity, (b) a summary of the attestation history for the policy period, including the distribution of GREEN/YELLOW/RED compliance status determinations, and (c) documentation of any invariant breaches, their severity,*

*and the remediation actions taken. The Insurer may adjust premium terms based on the conformance level achieved and the attestation history provided. Higher conformance levels and cleaner attestation histories **SHALL** be treated as positive governance maturity signals for underwriting purposes.*

### 11.4 Enterprise Vendor Risk Management

The following clause is designed for insertion into Master Service Agreements (MSAs), vendor security questionnaires, and third-party risk management frameworks. It follows the structural model established by SOC 2 Type II reports in the enterprise procurement cycle.

***AI Vendor Governance Requirement.*** *Vendor **SHALL** maintain MAI-1 conformance at level MAI-C1 or higher for all AI systems processing, generating, or influencing decisions related to Buyer's data, operations, or customers. Vendor **SHALL** provide Buyer with a current MAI-1 Conformance Report annually, or within 30 days of any material change to the AI system's architecture, model version, or deployment configuration. Vendor **SHALL** notify Buyer within 48 hours of any MAI-1 invariant breach resulting in RED compliance status. Buyer reserves the right to request attestation artifacts and Decision Receipts for independent verification. Failure to maintain the required MAI-1 conformance level constitutes a material breach of this Agreement.*

### 11.5 The Procurement Cascade

The four clauses above are designed to trigger the same **supply chain pressure cascade** that converted PCI-DSS and SOC 2 from voluntary standards into business requirements:

1. A government agency or enterprise buyer inserts the procurement clause into a solicitation or MSA.
2. The AI vendor, to satisfy the requirement, obtains MAI-1 conformance certification—which requires implementing the attestation endpoint, integrating TEE-rooted signing, and passing the conformance test suite.
3. The AI vendor's own suppliers (cloud providers, hardware vendors, model providers) must support the attestation infrastructure to enable their customer's compliance.
4. Cloud providers, to serve the growing base of MAI-1 compliant customers, build native attestation endpoint support into their AI platforms.
5. The standard propagates upstream and downstream simultaneously, reaching critical mass when a sufficient fraction of the supply chain treats MAI-1 as table stakes.

This cascade is not speculative. It is the documented mechanism by which PCI-DSS reached global adoption (via acquirer fines and merchant agreements), TLS certificates became universal (via browser root store policies and the "Not Secure" UI campaign), and SOC 2 became mandatory for B2B SaaS (via enterprise VRM policies at Google, Microsoft, and AWS). The procurement clauses above are the insertion points. The cascade is the adoption mechanism.

## 12 What MAI-1 Does Not Guarantee

The credibility of a governance framework is proportional to the honesty with which it confronts its own limitations. A framework that overpromises invites justified skepticism from precisely the technical and regulatory audiences it must convince. MAI-1 faces fundamental constraints at every layer—from mathematical impossibility results to hardware vulnerabilities—that bound what attestation can and cannot deliver. This section catalogs those constraints transparently, following the epistemic honesty established in MSAF §8 (Fields, 2026).

### 12.1 No Behavioral Safety Guarantees

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—and no attestation architecture—can guarantee behavioral safety for arbitrary future inputs.

**Proposition 12.1** (The Process-Behavior Gap)**.** *Let $M$ be a model produced by training process $P$ on dataset $D$. Let $\varphi$ be any non-trivial behavioral property (e.g., "$M$ never produces harmful output"). Then:*

$$\textsc{Attest}(P, D) \not\Rightarrow \varphi(M)$$

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

MAI-1 attests that a model was *healthy* and *provenance-verified* at inference time. It does not and cannot attest that the model's output is *correct*, *safe*, or *unbiased*. This is analogous to financial auditing, which certifies process compliance ("the books were kept according to GAAP") without guaranteeing future solvency ("the company will not go bankrupt").

### 12.2 No Bias Elimination

MAI-1 invariants detect distributional anomalies (drift, entropy collapse, coherence degradation) but do not measure or mitigate bias in the sociological sense. A model that produces biased outputs can be perfectly healthy by every MAI-1 invariant—stable gradients, bounded drift, coherent representations—while systematically discriminating against protected classes. Bias detection and mitigation require domain-specific fairness metrics, demographic impact analysis, and human oversight mechanisms that are complementary to, but outside the scope of, attestation infrastructure.

### 12.3 No Correctness Guarantees

The "Impossibility Sandwich" formalized by Grigore (arXiv 2507.03031, 2025) establishes that the minimum complexity required for an AI system to be useful exceeds the maximum complexity for which safety can be formally verified. MAI-1 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 invariant checks is *more likely* to be operating correctly than one that does not, but the invariants cannot *guarantee* correctness.

### 12.4 Hardware Trust Has Physical Limits

The TEE trust model assumes that silicon is trustworthy even when software is not. A cluster of 2024–2025 research has demonstrated that this assumption has physical limits. Side-channel attacks (power analysis, electromagnetic emanation, cache timing), fault injection attacks, and supply chain attacks against manufacturing processes can compromise hardware trust primitives. MAI-1's Layer 1 provides the strongest commercially available trust anchor, but it is not invulnerable. The honest framing: TEE attestation makes hardware compromise *expensive and detectable*, not *impossible*.

### 12.5 The Measurement Gap

A persistent technical challenge is the non-deterministic loading of large models in GPU memory. Two different loading sessions of the same model may result in bit-level differences due to padding or dynamic memory allocation, complicating hash-based verification. Current practice

measures the model file signature during loading rather than the runtime memory image. Future work on deterministic loaders or semantic hashing of neural networks may close this gap, but as of early 2026, it represents a known limitation in the fidelity of Layer 2 model identity verification.

### 12.6 What MAI-1 Does Provide

Within these constraints, MAI-1 provides:

(i) **Accountability infrastructure**: an immutable record of who deployed what model, in what state, on what hardware, at what time.

(ii) **Probabilistic risk reduction**: continuous health monitoring that detects the known precursors of model failure (entropy collapse, gradient instability, distribution drift, representational fragmentation, thermal degradation).

(iii) **Forensic capability**: the ability to reconstruct, after the fact, the complete governance state of a model at the moment a specific output was produced.

(iv) **Regulatory compliance evidence**: machine-verifiable artifacts satisfying the documentation, monitoring, and auditability requirements of seven regulatory frameworks simultaneously.

(v) **Insurance underwriting data**: the continuous, cryptographically signed evidence stream that enables the transition from epistemic uncertainty to actuarial risk quantification.

This is not everything. It is not a safety guarantee. But it is *infrastructure*—the necessary foundation upon which behavioral safety evaluation, bias auditing, and fairness certification can be built. Without it, those higher-order governance activities rest on unverified assumptions about model identity, health, and provenance. With it, they rest on cryptographic evidence.

## 13 Versioning, Governance, and Future Clauses

### 13.1 The Modular Principle

MAI-1 enforces a strict separation between **protocol** and **content**:

**Standardize the protocol:** How evidence is generated, signed, transmitted, verified, and stored. The cryptographic primitives (hash functions, signature schemes, commitment schemes), the attestation token format (COSE/EAT), the Merkle tree construction algorithm, and the inter-layer binding mechanism. These are infrastructure—they change slowly and benefit from interoperability.

**Let the content evolve:** What is being proven. The specific health invariants (§7), the threshold values, the monitoring frequencies, the measurement algorithms, and the regulatory mappings. 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 MAI-1 profile without modifying the attestation infrastructure. The protocol remains stable; the content is versioned and updatable. This principle follows the design of the IETF RATS architecture itself, which separates the token format (EAT) from the claims it carries.

### 13.2 Version Identification

MAI-1 attestation artifacts carry a `mai-profile` field identifying the specification version (e.g., `"MAI-1-v1.0"`). Verifiers **MUST** check the profile identifier before processing and **MUST** reject artifacts claiming conformance to unrecognized profile versions. This versioning mechanism enables graceful migration: when MAI-1-v1.1 adds a new invariant or modifies a threshold calibration methodology, verifiers can support both versions during a transition period while requiring the updated version by a specified sunset date.

### 13.3 Future Clause Dependencies

MAI-1 (Clause AI-5) is positioned within the Auburn Patent Family clause architecture as the **composition and interface layer**. Future clauses in the Auburn family will build upon MAI-1 conformant attestation as a prerequisite:

> *Future clauses in the Auburn Patent Family (AI-6 and beyond)* ***SHALL*** *assume MAI-1 conformant attestation as a prerequisite for their governance guarantees. Systems that do not expose an MAI-1 compliant endpoint are outside the scope of all downstream Auburn governance clauses.*

This dependency structure means that MAI-1 is not a standalone document. It is the **foundation layer** upon which all future Auburn governance primitives are constructed. Entropy constraints (AI-8), gradient stability envelopes (AI-2), thermal integrity bounds (AI-4), speculative decoding stability (AI-3), and stateful isolation requirements—all depend on the attestation infrastructure defined here.

### 13.4 Standardization Trajectory

The MAI-1 profile is a natural candidate for standardization through multiple venues:

**IETF RATS:** MAI-1 is structured as a RATS profile (§6.4). Submission as an individual Internet-Draft defining a Foundation Model attestation profile would align with the RATS working group's existing roadmap and the `measured-component` extension's trajectory toward Proposed Standard status.

**CEN/CENELEC JTC 21:** The MAI-1 attestation artifact could be referenced as an acceptable evidence methodology under prEN 18229-2 (accuracy and robustness) and prEN 18282 (cybersecurity), either through normative referencing of external specifications or through Liaison A status enabling direct contribution to the drafting process.

**ISO/IEC JTC 1/SC 42:** MAI-1 conformance evidence strengthens ISO/IEC 42001 (AI Management System) certifications from self-reported governance documentation to verifiable evidence, without requiring changes to the standard's structure.

**Confidential Computing Consortium:** The composite attestation patterns in MAI-1 align with the CCC Attestation SIG's work on interoperable RA-TLS and Veraison plugin architecture.

These standardization pathways are complementary, not competing. MAI-1 can be simultaneously referenced by European harmonized standards, adopted as an IETF profile, and integrated into ISO management system certifications—just as PCI-DSS is simultaneously a private industry standard, a state law reference (Nevada NRS 603A, Minnesota Plastic Card Security Act), and a contractual requirement in merchant agreements worldwide.

# References

## IETF Standards and Drafts


[R1] H. Birkholz, D. Thaler, M. Richardson, N. Smith, W. Pan. *Remote ATtestation procedureS (RATS) Architecture*. RFC 9334, Internet Engineering Task Force, January 2023. https://www.rfc-editor.org/rfc/rfc9334

[R2] L. Lundblade, G. Mandyam, J. O'Donoghue, C. Wallace. *Entity Attestation Token (EAT)*. RFC 9711, Internet Engineering Task Force, April 2025. https://www.rfc-editor.org/rfc/rfc9711

[R3] L. Lundblade, G. Mandyam. *Entity Attestation Token (EAT) Media Types*. RFC 9782, Internet Engineering Task Force, May 2025. https://www.rfc-editor.org/rfc/rfc9782

[R4] J. Schaad. *CBOR Object Signing and Encryption (COSE): Structures and Process*. RFC 9052, Internet Engineering Task Force, August 2022. https://www.rfc-editor.org/rfc/rfc9052

[R5] C. Bormann, P. Hoffman. *Concise Binary Object Representation (CBOR)*. RFC 8949, Internet Engineering Task Force, December 2020. https://www.rfc-editor.org/rfc/rfc8949

[R6] S. Frost, T. Fossati. *EAT Measured Component*. draft-ietf-rats-eat-measured-component, IESG Evaluation, Internet Engineering Task Force, 2025–2026.

[R7] H. Birkholz, T. Fossati, Y. Deshpande, N. Smith, W. Pan. *Concise Reference Integrity Manifests (CoRIM)*. draft-ietf-rats-corim, WGLC, Internet Engineering Task Force, 2025–2026.

[R8] T. Fossati, H. Birkholz, N. Smith. *RATS Conceptual Message Wrappers (CMW)*. IESG-approved as Proposed Standard (v23), Internet Engineering Task Force, December 2025.

[R9] T. Fossati, S. Frost. *EAT Attestation Results (EAR)*. draft-ietf-rats-ear, v01, Internet Engineering Task Force, July 2025.

[R10] T. Fossati, E. Voit. *Attestation Results for Secure Interactions (AR4SI)*. draft-ietf-rats-ar4si, v09, Internet Engineering Task Force, 2025.

[R11] H. Birkholz, A. Delignat-Lavaud, C. Fournet, Y. Supply. *An Architecture for Trustworthy and Transparent Digital Supply Chains (SCITT)*. draft-ietf-scitt-architecture, WGLC, Internet Engineering Task Force, 2025–2026.

[R12] Y. Deshpande, et al. *Multi-Verifier Composition in RATS*. draft-deshpande-rats-multi-verifier, v03, Internet Engineering Task Force, October 2025.

[R13] M. Aylward, et al. *AI Governance and Accountability Protocol (AIGA)*. draft-aylward-aiga-2-00, Active, Internet Engineering Task Force, 2025–2026.

[R14] J. Novak, et al. *Trustworthy Workload Identity (TWI) Attestation*. draft-novak-rats-twi-attestation, Active, Internet Engineering Task Force, 2025–2026.

[R15] I. Mihalcea, et al. *Secure Enrollment and Attestation for TLS (SEAT) Use Cases*. draft-mihalcea-seat-use-cases, Active, Internet Engineering Task Force, 2025–2026.

[R16] D. Deeglaze. *CoRIM Profile for AMD SEV-SNP*. draft-deeglaze-amd-sev-snp-corim-profile, Individual, Internet Engineering Task Force, 2025.

[R17] K. Kamimura, et al. *SCITT Refusal Events*. draft-kamimura-scitt-refusal-events, Individual, Internet Engineering Task Force, 2026.

[R18] J. Condrey, et al. *Witnessed Revocation for RATS*. draft-condrey-rats-witnessd-revocation, Active, Internet Engineering Task Force, 2025–2026.

### European Standards and Regulation

[E1] European Parliament and Council. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)*. Official Journal of the European Union, August 2024.

[E2] CEN/CENELEC JTC 21, WG5. *prEN 18282: Cybersecurity specifications for AI systems*. Enquiry Phase, February 2026.

[E3] CEN/CENELEC JTC 21, WG4. *prEN 18229-2: AI trustworthiness framework — Part 2: Accuracy and robustness*. Enquiry Launch February 11, 2026.

[E4] CEN/CENELEC JTC 21, WG4. *prEN 18229-1: AI trustworthiness framework — Part 1: Transparency, logging and human oversight*. Enquiry Launch January 23, 2026.

[E5] CEN/CENELEC JTC 21, WG2. *prEN 18286: Artificial intelligence — Quality management system for EU AI Act regulatory purposes*. Enquiry Closed December 2025.

[E6] CEN/CENELEC JTC 21, WG2. *prEN 18285: Conformity Assessment Framework for AI Systems*. Active Drafting, 2026.

[E7] European Commission. *Standardisation Request M/593 to CEN and CENELEC (as amended by M/613)*. 2024–2025.

[E8] ISO/IEC. *ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system*. International Organization for Standardization, 2023.

[E9] CCRA. *Common Criteria for Information Technology Security Evaluation (CC:2022), ISO/IEC 15408:2022*. Parts 1–5 and CEM (ISO/IEC 18045).

[E10] European Commission. *EUCC: European Cybersecurity Certification Scheme under the Cybersecurity Act (EU) 2019/881*. 2024.

[E11] European Parliament and Council. *Cyber Resilience Act (CRA)*. Regulation on horizontal cybersecurity requirements for products with digital elements, 2024.

### US Federal Regulation and Guidance

[F1] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency. *SR 11-7 / OCC 2011-12: Supervisory Guidance on Model Risk Management*. April 2011.

[F2] Office of Management and Budget. *OMB M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust*. April 3, 2025.

[F3] Office of Management and Budget. *OMB M-25-22: Driving Efficient Acquisition of AI in Government*. April 3, 2025.

[F4] Office of Management and Budget. *OMB M-26-04: Increasing Public Trust in Artificial Intelligence Through Unbiased AI Principles*. December 11, 2025.

[F5] United States Congress. *FY2026 National Defense Authorization Act (P.L. 119-60).* Signed December 18, 2025. Sections 1512, 1513, 1532, 1533, 6604.

[F6] US Food and Drug Administration. *Predetermined Change Control Plans for Machine Learning-Enabled Device Software Functions: Guidance for Industry.* Finalized December 4, 2024; updated August 18, 2025.

[F7] National Institute of Standards and Technology. *AI Risk Management Framework (AI RMF 1.0).* NIST AI 100-1, January 2023.

[F8] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative AI Profile.* NIST AI 600-1, July 2024.

[F9] National Institute of Standards and Technology. *FIPS 140-3: Security Requirements for Cryptographic Modules.* March 2019 (effective September 2019). Derived from ISO/IEC 19790:2012.

[F10] Executive Office of the President. *Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence.* January 2025.

### Industry Standards and Specifications

[I1] PCI Security Standards Council. *Payment Card Industry Data Security Standard (PCI DSS) v4.0.* March 2022.

[I2] Trusted Computing Group. *TPM 2.0 Library Specification, Version 184.* March 2025.

[I3] Trusted Computing Group. *PC Client Platform TPM Profile, Version 1.06.* April 2025.

[I4] Trusted Computing Group. *DICE Attestation Architecture, v1.1.* 2024.

[I5] OWASP Foundation. *CycloneDX SBOM Standard, v1.6.* Including AI/ML BOM extensions. 2024–2025.

[I6] OWASP Foundation. *OWASP AI Exchange and Top 10 for LLMs.* 2024–2025.

[I7] JEDEC Solid State Technology Association. *JESD79-5B: DDR5 SDRAM Standard.* Thermal specification appendices.

[I8] DMTF. *Security Protocol and Data Model (SPDM) Specification, v1.1+.* 2023–2025.

[I9] Coalition for Content Provenance and Authenticity. *C2PA Technical Specification, v2.0+.* 2024–2025.

[I10] Artificial Intelligence Underwriting Company. *AIUC-1: Certification Standard for AI Agents.* July 2025.

[I11] National Association of Insurance Commissioners. *Model Bulletin on the Use of Artificial Intelligence Systems by Insurers.* December 2023. Adopted by 24+ states and DC.

### Auburn Patent Family — Internal References

[A1] R. Fields. *The Model State Attestation Framework: Evidence-Based Governance for Foundation Models.* Auburn Patent Family, February 2026.

[A2] R. Fields. *Clause AI-2: The Gradient Starvation Envelope — A Formal Compliance Primitive for Sparse Mixture-of-Experts Training Dynamics*. Auburn Patent Family, February 2026.

[A3] R. Fields. *Clause AI-3: Lyapunov Stability for Speculative Decoding*. Auburn Patent Family, 2026.

[A4] R. Fields. *Clause AI-4: SRAM Thermal Integrity Bound*. Auburn Patent Family, 2026.

[A5] R. Fields. *Clause AI-5: The Model Attestation Interface (MAI-1) — A Normative Profile and Conformance Protocol for Foundation Model Governance*. Auburn Patent Family, February 2026. **This document.**

[A6] R. Fields. *Clause AI-8: Attention Thermodynamics — Entropy Collapse Constraint*. Auburn Patent Family, 2026.

[A7] R. Fields. *The Stateful Isolation Law*. Auburn Patent Family, 2026.

### Academic and Research References

[S1] H.G. Rice. "Classes of recursively enumerable sets and their decision problems." *Transactions of the American Mathematical Society*, 74(2):358–366, 1953.

[S2] R. Grigore. "On the Mathematical Impossibility of Safe Universal Approximators." arXiv:2507.03031, 2025.

[S3] K. Zhang, et al. "Compositional Neural Certificates for Networked Dynamical Systems." *Proceedings of L4DC*, MIT REALM, 2023.

[S4] Various. "Certificates in AI: Learn but Verify." *Communications of the ACM*, 2025.

[S5] Various. "Constant-Size Cryptographic Evidence Structures." arXiv:2511.17118, 2025.

[S6] Intel Labs. *Atlas: TEE-backed ML Pipeline Provenance*. Apache 2.0, v0.1.0, February 2025.

[S7] Mithril Security. *AICert: Virtual TPM-based Model Certification*, v1.0, September 2024.

[S8] Ghent University–imec. *AIBoMGen: AI Bill of Materials Generation via in-toto Attestations*. Accepted at CAIN 2026.

[S9] Various. "ACAI: Accelerator Attestation for Confidential Computing." *Proceedings of USENIX Security*, 2024.

[S10] Various. "Careful Whisper: Scalable Peer-to-Peer Attestation." arXiv:2507.14796, 2025.

[S11] Various. "Min-K% Prob: Detecting Pretraining Data from Large Language Models." *Proceedings of ICLR*, 2024.

[S12] Various. "Contamination Detection via Output Distribution Analysis (CDD)." *Proceedings of ACL*, 2024.

[S13] Stanford University HAI. *AI Index Report 2025*. Stanford, CA, 2025.

# A Acronyms and Abbreviations

| Acronym | Definition |
|---|---|
| AIGA | AI Governance and Accountability Protocol |
| AI BOM | AI Bill of Materials |
| AR4SI | Attestation Results for Secure Interactions |
| ASV | Approved Scanning Vendor (PCI-DSS) |
| CA | Certificate Authority |
| CBOR | Concise Binary Object Representation |
| CC | Common Criteria (ISO/IEC 15408) |
| CCC | Confidential Computing Consortium |
| CDDL | Concise Data Definition Language |
| CDAO | Chief Digital and Artificial Intelligence Officer |
| CMW | Conceptual Message Wrappers |
| CoRIM | Concise Reference Integrity Manifests |
| COSE | CBOR Object Signing and Encryption |
| CRA | Cyber Resilience Act (EU) |
| CSR | Certificate Signing Request |
| CT | Certificate Transparency |
| CTE | Conformance Testing Entity |
| CVM | Confidential Virtual Machine |
| CWT | CBOR Web Token |
| DFARS | Defense Federal Acquisition Regulation Supplement |
| DICE | Device Identifier Composition Engine |
| DKL | Kullback–Leibler Divergence |
| DMTF | Distributed Management Task Force |
| EAL | Evaluation Assurance Level (Common Criteria) |
| EAR | EAT Attestation Results |
| EAT | Entity Attestation Token |
| FAR | Federal Acquisition Regulation |
| FDA | US Food and Drug Administration |
| FIPS | Federal Information Processing Standards |
| FSP | Foundation Security Processor (NVIDIA) |
| GPU | Graphics Processing Unit |
| GSP | GPU System Processor (NVIDIA) |
| HBM | High Bandwidth Memory |
| hEN | Harmonised European Standard |
| IETF | Internet Engineering Task Force |
| JTC | Joint Technical Committee |
| JWT | JSON Web Token |
| MAI | Model Attestation Interface |
| MGA | Managing General Agent |
| MoE | Mixture of Experts |
| MSAF | Model State Attestation Framework |
| NDAA | National Defense Authorization Act |
| NRAS | NVIDIA Remote Attestation Service |
| NSB | National Standardization Body |
| NVLAP | National Voluntary Laboratory Accreditation Program |
| OCC | Office of the Comptroller of the Currency |
| OJEU | Official Journal of the European Union |

| Acronym | Definition |
|---|---|
| OMB | Office of Management and Budget |
| PCCP | Predetermined Change Control Plan |
| PCI-DSS | Payment Card Industry Data Security Standard |
| PKI | Public Key Infrastructure |
| PP | Protection Profile (Common Criteria) |
| QMSR | Quality Management System Regulation |
| RATS | Remote ATtestation procedureS |
| RFP | Request for Proposal |
| RIM | Reference Integrity Manifest |
| SaMD | Software as a Medical Device |
| SCITT | Supply Chain Integrity, Transparency, and Trust |
| SCT | Signed Certificate Timestamp |
| SEV-SNP | Secure Encrypted Virtualization–Secure Nested Paging |
| SOC | Service Organization Control |
| SOW | Statement of Work |
| SPDM | Security Protocol and Data Model |
| SRAM | Static Random-Access Memory |
| TCB | Trusted Computing Base |
| TCG | Trusted Computing Group |
| TDX | Trust Domain Extensions (Intel) |
| TDISP | TEE Device Interface Security Protocol |
| TEE | Trusted Execution Environment |
| TLS | Transport Layer Security |
| TOCTOU | Time-of-Check-to-Time-of-Use |
| TPM | Trusted Platform Module |
| TWI | Trustworthy Workload Identity |
| VRM | Vendor Risk Management |

# B Auburn Patent Family Clause Cross-Reference

Table 11 maps the Auburn Patent Family clause designations to their titles, the MAI-1 components they support, and their role in the attestation architecture.

Table 11: Auburn Patent Family clause dependencies for MAI-1.

| Clause | Title | MAI-1 Role | Layer |
|---|---|---|---|
| AI-2 | Gradient Starvation Envelope | Invariant 2: gradient stability | Layer 2 |
| AI-3 | Lyapunov Stability for Speculative Decoding | Extended stability analysis | Layer 2 |
| AI-4 | SRAM Thermal Integrity Bound | Invariant 5: thermal integrity | Layer 1 / 2 |
| **AI-5** | **Model Attestation Interface (MAI-1)** | **This document** | **All** |
| AI-8 | Attention Thermodynamics | Invariant 1: entropy floor | Layer 2 |
| — | Stateful Isolation Law | Multi-tenant security model | Layer 1 |
| — | Model State Attestation Framework | Three-tier architecture foundation | All |

The clause numbering reflects the order of formal specification, not a dependency hierarchy. However, MAI-1 (AI-5) serves as the **composition layer**: it defines the interface through which all other clauses' governance guarantees are delivered as verifiable evidence. Future clauses (AI-6 and beyond) will assume MAI-1 conformant attestation as a prerequisite, as specified in §13.3.

# Intellectual Property (IP) Declaration

## Auburn Patent Family

The methods, logic structures, interface specifications, invariant definitions, conformance test architectures, and composition protocols contained in this work are the sole property of Ryan Fields.

## Public License (Non-Commercial)



- **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 interface specifications, invariant definitions, conformance test architectures, or composition protocols 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 AI governance platforms or compliance automation software.
- Integration into commercial attestation services, verification infrastructure, or conformance testing products.
- Implementation of the MAI-1 interface specification in commercial AI deployment platforms or cloud services.
- Use by insurance underwriters, financial institutions, or government contractors for operational compliance without commercial license.
- Use within proprietary high-frequency trading (HFT) risk models.
- Integration into commercial high-assurance AI governance software.

## Contact for Commercial Licensing

Entities seeking to license this framework for commercial applications, or to integrate the MAI-1 interface specification, conformance test suite, or attestation architecture into institutional governance infrastructure, must contact the author directly at:

**Email:** UncleBroFields@proton.me
fieldsryanchristopher@gmail.com