Defending the Real-Time Audit Trail in Regulated Financial AI

Banks, fintechs, asset managers, capital-markets infrastructure

White paper — Tamper-evident audit, multi-witness anchoring, FIPS-grade key brokerage

Enclawed LLC **May 5, 2026**

Executive summary

Financial institutions running AI models on customer data, trading data, KYC/AML pipelines, or fraud-detection engines face a regulatory triad that's been tightening every year: **PCI DSS 4.0**, **SOC 2 Trust Services Criteria**, and the personal-data regulations (**GDPR**, US state laws, the post-CCPA wave). On top of that, the prudential regulators (OCC, FCA, ECB, BaFin) increasingly require demonstrable *model risk management* (SR 11-7, the Basel Committee's *Principles for the sound management of model risk*, the ECB's TRIM and IFRS 9 guidance) — which in practice means an audit trail that ties every model decision to immutable input provenance.

enclawed-enclaved provides the audit-trail integrity substrate. Hash-chained append-only logs, M-of-N quorum attestation, and an optional public-blockchain anchor produce a non-repudiable, tamper-evident record of every gated AI operation — without requiring the institution to operate its own blockchain or rely on a single cloud KMS.

Key point. The hardest part of regulated AI is not the model. It's the ledger that proves what the model saw, what it decided, and that neither the input nor the decision has been altered after the fact. enclawed-enclaved provides that ledger as a primitive.

1 The regulator's actual question

When a model makes a credit decision, a fraud flag, or a market order, the regulator's question is not "how did the model work?" — it's:

- Can you prove what data the model saw at decision time?
- Can you prove that data has not been altered since?

- Can you prove which version of the model produced the decision, and that the model was the one approved by Model Risk Management?
- Can you prove the decision itself has not been altered after the fact?

The risk. A standard cloud-native ML stack stores its audit trail in the same cloud that runs the model. A compromised cloud-KMS operator, a malicious insider, or an adversary who reaches the control plane can edit the log. The regulator's tests assume exactly this attack vector exists.

2 What enclawed-enclaved gives you

2.1 Hash-chained append-only audit log

Every gated operation emits a record that includes the `prevHash` of the previous record. The chain is verifiable end-to-end with a single SHA-256 walk; any tampering produces a visible mismatch at the tampered index.

Compliance. PCI DSS 4.0 **10.x** (audit logs); SOC 2 **CC7.2** (monitoring), **CC7.3** (anomaly detection); GDPR Article 30 (records of processing) and Article 32 (security of processing).

2.2 M-of-N independent accreditation

A closed-tree extension lets you configure N independent witness nodes (operated by separate business units, separate cloud accounts, or separate legal entities) to counter-sign the log head. A regulator's verification requires only M of those signatures to be valid; no single party — including the platform operator — can forge a record.

2.3 Permissioned blockchain accreditation

For regulated infrastructures that already run a private chain (R3 Corda, Hyperledger Fabric, etc.), a closed-tree extension wraps the audit log in a K-of-N Byzantine-fault-tolerant chain whose validator quorum is set by the institution. The chain is in-process (no external blockchain), but the consensus rules match a real BFT chain.

2.4 Public-blockchain tail-truncation anchor (optional)

For institutions that want a public, third-party-verifiable anchor without operating any internal blockchain, a closed-tree extension periodically anchors the audit log's head hash to a public chain. The on-chain anchor event is sufficient for any regulator with read-only access to verify the log has not been tail-truncated.

2.5 Regulator-readable evidence package: NIST OSCAL emit

The boundary emits the **full set of four NIST OSCAL 1.2.2 submission models** an examiner's GRC platform expects:

- *Component Definition* — which 800-53 controls the boundary implements, with status, narrative, and evidence pointers per control.
- *Assessment Results* — per-sample observations and per-control findings from each adversarial run.
- *System Security Plan starter* — deployment-specific document with the institution's FIPS-199 categorisation, authorization-boundary description, information types, and users; the framework's contributions are pre-filled, the institution's deployment-specific fields are hard-gated (no placeholder defaults).
- *Plan of Action and Milestones* — one item per partial-status control's deployer-owned residual plus any not-satisfied finding from a paired Assessment Results.

All four are validated against NIST's published JSON Schemas, detached-signed with Ed25519, and written into the same hash-chained audit log as every gate decision. Bundled SVG architecture, audit-chain, and admission-gate diagrams ride along as OSCAL back-matter resources. The format is what an FFIEC examiner, an OCC information-technology examination team, or a SOC 2 auditor's GRC tool consumes directly: machine-readable, tamper-evident, and importable into the institution's regulatory package without manual rekey. Maps to NIST SP 800-53 **CA-2** (Control Assessments), **CA-5** (POA&M), **CA-7** (Continuous Monitoring), **CM-6** (Configuration Settings inventory), **PT-3** (Personally Identifiable Information Processing Purposes), **PT-5** (Privacy Notice).

3 Composable cryptographic primitives

enclawed-enclaved's *FIPS Mode of Operation* routes every cryptographic operation through a FIPS-validated provider. For financial institutions this means the audit-trail integrity — and the model-decision signing — inherits the strongest available cryptographic posture without bespoke plumbing.

Risk surface	enclawed-enclaved primitive
PII / PCI account-data leakage from model prompts	*DLP scanner* (regex-based, severity-graded, ships with US classification banners + AWS / GCP / Azure secret patterns + IBAN / BIC / SSN / credit-card patterns)
Prompt-injection attacks on conversational AI	*prompt-shield* (structural sanitization + 75-language imperative-override detection)
Cross-model / cross-tenant key separation	Bell–LaPadula classification lattice with the `financial-services` scheme (none, internal, sensitive, mnpi)
Outbound data exfiltration through tool calls	*two-layer egress allowlist* covering both high-level HTTP and low-level socket egress; VPN-only posture available for classified deployments

Backdoored or compromised plugins reaching attacker hosts	*biconditional admission* per extension: signed manifest plus an explicit per-extension destination allowlist; runtime deviations are audited
Post-init configuration drift, hostile module load	Mandatory zero-trust accreditor at boot that gates every extension load and audits + blocks any post-init tamper attempt
Cloud-KMS single-point-of-trust failure	Zero-trust quorum key broker requiring multiple independent custodians to agree before any key material is released
Model-version provenance	Module-signing with a customer-controlled trust root

Table 1: Financial-services risk surface coverage.

4 Standards mapping (concise)

Standard	**enclawed-enclaved coverage**
PCI DSS 4.0 **3.x / 4.x / 10.x / 11.x**	FIPS-grade crypto, encrypted transmission, hash-chained log, cloud-security monitor
SOC 2 **CC6.x / CC7.x / CC8.x**	Logical access, monitoring, anomaly handling, change management
GDPR Art. 5(1)(f), 25, 30, 32	Integrity + confidentiality, by-design, RoPA, security of processing
CCPA / CPRA	DLP / classification / audit
ISO/IEC 27001:2022	Org / people / physical (out of scope) / technological controls
NIST CSF 2.0	Identify / Protect / Detect / Respond / Recover
SR 11-7, Basel *Principles MRM*	Tamper-evident decision provenance, model-version binding

5 Integration patterns

5.1 Pattern A: shadow audit only

Run enclawed-enclaved alongside the existing model. Every production decision is also signed and logged through the closed-tree's audit primitives; the existing ML stack is unchanged. Lowest-risk integration; useful as a stepping stone before moving the actual gating into enclawed-enclaved.

5.2 Pattern B: gated production

Wrap the model's input + output in the closed-tree's policy + DLP + prompt-shield primitives. Decisions cannot reach the customer without passing through the chain. Provides the strongest regulator-facing posture.

5.3 Pattern C: cross-cloud key brokerage

For multi-cloud institutions, the runtime is configured to require a quorum of independent key custodians, each in a separate cloud account (for example AWS KMS, Azure Key Vault, GCP KMS). Every key fetch is gated on quorum approval and produces an evidence record that the audit log absorbs; no single cloud breach exposes the keys.

6 Direct empirical evidence

This is not a marketing claim. We back it with a published statistical in-vivo harness (`enclawed/test/security/in-vivo/llm-narrative.mjs` in the public repository) that mediates 1600 chat-message samples through three subjects against real Discord and Telegram bot endpoints. Across the full run, upstream OpenClaw achieves **recall = 0.000** on every failure mode (F1 gate-bypass, F2 audit-forgery, F3 silent-host-failure, F4 wrong-target); both enclawed-oss and enclawed-enclaved achieve **precision = recall = F1 = 1.000** on all four. Total wall-clock for the full pass: 42 seconds. Per-sample ground-truth labels and per-subject decisions are written to a CSV; every gate decision lands in a tamper-evident audit log; every witness record is independently re-verifiable from a journal file. The companion paper, *Architectural Obsolescence of Unhardened Agentic-AI Runtimes*, formalizes the methodology.

Talk to us

Enclawed LLC
Alfredo Metere <alfredo.metere@enclawed.com>
Reference deployments and a per-pattern technical design package available on request.