Zero-Trust Module Attestation for Operational Technology AI

Energy, water, transport, utilities, OT/ICS owner-operators

White paper — Multi-witness accreditation against the supply-chain-compromise threat

Enclawed LLC **May 5, 2026**

Executive summary

Operational technology (OT) operators — the utilities, the transport networks, the water authorities, the grid balancing authorities — are deploying AI models on monitoring data streams (SCADA, historian databases, smart-meter telemetry, substation phasor data). The threat surface is not the same as the IT enterprise's. The dominant adversary is a *supply-chain-compromise* actor: a state-grade adversary who reaches the operator's environment through a compromised software update, library, or model artifact.

enclawed-enclaved addresses the supply-chain question head-on with three composable mechanisms: a customer-rooted module-signing trust chain; a zero-trust K-of-N key broker that treats every external custodian (cloud KMS, HSM-as-a- service) as untrusted; and three independent accreditor extensions that produce non-repudiable, multi-party-attested records of every gated operation.

Key point. For OT, the question is not "can the model be wrong?" — it's "can a state-grade adversary substitute a malicious model without leaving a forensic trace?" enclawed-enclaved makes the substitution detectable at every gated entry, and a detection cryptographically traceable to the moment of substitution.

1 The OT threat model

1.1 What OT actually looks like

- Asset lifetimes measured in decades, not quarters.
- Patch windows measured in years, not weeks.
- Operating systems often pinned to a specific vendor-validated build.

- Network segmentation against the IT plane (Purdue Level 3.5 / 3 / 2 boundaries; IEC 62443 zones + conduits).
- A regulator that does not accept "move fast and break things" as an operational philosophy.

1.2 The dominant adversary

The risk. The OT adversary is not a fraudster trying to steal a credit- card number. It is a state-grade actor whose objective is to position malicious code in a critical-infrastructure operator's environment for years before activation. The vector is software supply chain: a compromised library update, a substituted model artefact, a poisoned training-data feed.

The 2025 **NIS2** directive in the EU, the **NERC CIP** framework in North America, and the **CRA (Cyber Resilience Act)** all converge on the same operator obligation: cryptographically demonstrable software provenance for every component running in the OT environment.

2 enclawed-enclaved's three layers of attestation

2.1 Layer 1 — Module signing and the customer trust root

Every component that loads in the closed tree carries a signed `enclawed.module.json` manifest. The signature is verifiable against a trust root that the customer *controls* — typically pinned in an air-gapped HSM or a hardware Yubikey held by the operator's CISO office. A substituted module fails verification at load time; the substitution is recorded in the audit log.

2.2 Layer 2 — Zero-trust K-of-N key broker

Cryptographic keys for the operator's AI models do not come from a single cloud KMS. They come from K-of-N independent custodians (cloud KMSs, in-house HSMs, federated key authorities). The broker treats every custodian as untrusted: each response must carry a provider-signed attestation, and a quorum of attestations must AGREE before any key material is exposed to the host.

Compliance. NIST SP 800-53 **SC-12** (Cryptographic Key Establishment + Management); IEC 62443-3-3 **SR 7.3** (Control system backup); NIST CSF 2.0 *Identify* (Asset Management and Risk Management functions).

2.3 Layer 3 — Multi-witness audit-chain accreditation

The closed tree ships a family of audit-chain accreditation extensions that produce multi-party-attested records of every gated operation:

Mode	Trust model

M-of-N local quorum	Independent local witnesses, each with its own signing key, agree on a hash-chained journal of every gated operation. *Fully FIPS-mode compatible.*
K-of-N permissioned blockchain	Permissioned-chain validators with two-layer trust (receipt-level + block-level quorum). *Fully FIPS-mode compatible.*
Public-blockchain anchor (optional)	Optional public-chain anchor for tail-truncation defense without operating an internal chain. Deployed outside the FIPS boundary because public-chain primitives are not part of the validated module.

Table 1: Three accreditation modes, three trust models. Mix and match per-deployment.

3 NERC CIP / IEC 62443 / NIS2 mapping

Standard	**Coverage**
NERC CIP-002 BES Cyber System Categorization	Bell–LaPadula classification lattice with operator-defined scheme
NERC CIP-005 Electronic Security Perimeters	Two-layer egress allowlist (HTTP + socket) with VPN-only posture + per-extension biconditional admission
NERC CIP-007-7 R3 (malicious code prevention)	Prompt-shield + DLP + integrity-anchored module loader
NERC CIP-007-7 R4 (security event monitoring)	Hash-chained audit log + accreditor witness chain
NERC CIP-010 Configuration Change Management	Sealed integrity manifest, re-sealing requires customer trust-root key
NERC CIP-013 Supply-Chain Risk Management	Module-signing trust root + accreditor attestation
IEC 62443-3-3 SR 1.x (Identification + AuthN)	Module-signed clearance attestation
IEC 62443-3-3 SR 2.x (Use Control)	Policy + classification + role-service catalog
IEC 62443-3-3 SR 3.x (System Integrity)	FIPS §7.5 integrity test, hash-chained audit
IEC 62443-3-3 SR 4.x (Data Confidentiality)	FIPS-grade crypto, DLP
IEC 62443-3-3 SR 6.x (Timely Response)	Audit log + cloud-security monitor (anomaly detection)
EU NIS2 Article 21 (cybersecurity risk-management)	full closed-tree feature set
EU CRA (Cyber Resilience Act) Annex I	SBOM, module-signing, integrity, vulnerability handling

NIST OSCAL 1.2.2 (machine-readable evidence)	Full FedRAMP submission model set — Component Definition, Assessment Results, System Security Plan starter, Plan of Action and Milestones — emitted per assessment run, schema-validated, Ed25519-signed, with bundled SVG architecture / audit-chain / admission-gate diagrams as back-matter resources

Table 2: OT-applicable standards mapping.

3.1 Machine-readable assessment artefacts

For a North American Electric Reliability Corporation (NERC) CIP audit, an EU NIS2 conformity assessment, or an IEC 62443-2-1 maturity review, the boundary emits its assessment evidence as the full four-model **NIST OSCAL 1.2.2** submission set: a Component Definition (which 800-53 controls the boundary implements), an Assessment Results document (the latest in-vivo adversarial run's per-sample observations and per-control findings), a System Security Plan starter (deployment-specific fields hard-gated for the operator to fill in — FIPS-199 categorisation, authorization boundary, information types, users), and a Plan of Action and Milestones (deployer-owned residuals and any not-satisfied finding). Bundled SVG architecture, audit-chain, and admission-gate diagrams ride along as OSCAL back-matter resources. All four pass the official NIST OSCAL JSON Schema, are detached Ed25519-signed, and arrive directly in the GRC tool the auditor's team is already using. For OT operators with $\geq$30 substations or assets in scope, this is the difference between hand-curating an evidence binder per audit cycle and shipping the same machine-readable record their compliance partner can diff against the prior cycle.

4 Cloud-security monitoring in the OT plane

A closed-tree extension implements the architecture described in Zhang, Bai, Luo (AEECA 2025) — a multi-method cloud-security monitor with five threat categories (unauthorized access 37%, data leakage 26%, abnormal API calls 18%, configuration errors 12%, malware activity 7%). It ships with adapters for AWS CloudTrail, Azure Activity Log, and GCP Audit Log; for OT operators with hybrid cloud presence, it gives a single audit-ingest pipeline covering both their IT and their OT-cloud-bridge planes.

The implementation contains *no AI/ML/neural-network components*. Every paper detector is replaced with a classical-statistical or rule-based equivalent: Welford z-score, EWMA, IQR, regex rules, standardized-distance k-NN. The substitution table is documented in the user guide.

Key point. For OT environments, the absence of opaque ML in the monitoring path is an asset. Every detector's decision is deterministic and auditable. There are no model-update windows in which the monitor's behaviour silently changes.

5 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.

Speak with us

Enclawed LLC
Alfredo Metere <alfredo.metere@enclawed.com>
OT-segmented reference architectures and a NERC CIP / IEC 62443 mapping workbook are available on request.