CAP-SRP v0.1.0 Released: Cryptographic Proof of AI Safe Refusals

Open-source reference implementation for verifiable AI refusal logs with signatures, Merkle proofs, RFC 3161 anchoring, and completeness checks

TOKYO, JAPAN, January 28, 2026 /EINPresswire.com/ -- VeritasChain Standards Organization (VSO), a vendor-neutral non-profit developing cryptographic auditability standards, today announced the public release of CAP-SRP v0.1.0, an open-source reference implementation that produces independently verifiable evidence that an AI system refused to generate harmful content.

Logo of the VeritasChain Standards Organization (VSO), a neutral standards body developing cryptographic audit and provenance frameworks for AI systems.

As regulators and auditors increase scrutiny of generative AI safety controls, many providers still rely on internal logs to demonstrate that safeguards worked. Internal logs may be accurate, but they are self-reported and difficult for third parties to verify. CAP-SRP addresses this accountability gap by generating tamper-evident records that can be validated outside the provider environment.

> CAP-SRP turns safe refusal claims into independently verifiable evidence. When regulators ask if your AI refused harmful content, trust our logs is not an acceptable answer."
>
> *Tokachi Kamimura, VeritasChain Standards Organization*

CAP-SRP records the full lifecycle of AI generation requests as signed events, including the initial request commitment and the final outcome. The reference implementation supports four event types: GEN_ATTEMPT for logging the request before evaluation, GEN for successful generation, GEN_DENY for a policy refusal, and GEN_ERROR for technical failures.

CAP-SRP combines multiple cryptographic mechanisms to make refusal provenance independently verifiable. Each event is signed using Ed25519 and linked through a hash chain to detect modification. Events are

also organized into a Merkle tree to enable efficient inclusion proofs, and deployments can optionally anchor Merkle roots via RFC 3161 timestamp authorities for stronger third-party evidence.

In addition, CAP-SRP introduces a completeness invariant that detects missing or manipulated records within any time window:

COUNT(GEN_ATTEMPT) = COUNT(GEN) + COUNT(GEN_DENY) + COUNT(GEN_ERROR)

If the equation fails, the system flags a completeness violation, indicating that events may have been deleted, hidden, or fabricated. This provides a deterministic integrity check designed to support audit workflows at scale.

CAP-SRP v0.1.0 includes a schema-first event model with a discriminator field, event_type, and strict validation rules to reduce ambiguity and improve interoperability. The release also includes automated tests for schema validity, example validation, and rejection of malformed or non-compliant events, along with a Streamlit-based dashboard and CLI utilities for verification and reporting.

CAP-SRP is published as an open-source reference implementation to support auditors, regulators, AI providers, and standards practitioners exploring verifiable evidence for AI safety controls. It is intended as a foundation for further conformance profiles, evidence packaging, and interoperability work.

GitHub repository: https://github.com/veritaschain/cap-srp
Project page: https://veritaschain.org/vap/cap/srp/
IETF draft context (SCITT profile): https://datatracker.ietf.org/doc/draft-kamimura-scitt-vcp/

About VeritasChain Standards Organization (VSO)
VeritasChain Standards Organization is a vendor-neutral non-profit focused on cryptographic auditability standards for AI and automated systems. VSO develops open specifications and reference implementations designed to enable independent verification, transparency, and accountability.

Media Contact
VeritasChain Standards Organization (VSO)
Email: info@veritaschain.org

Website: https://veritaschain.org

TOKACHI KAMIMURA
VeritasChain Co., Ltd.
kamimura@veritaschain.org

Visit us on social media:
LinkedIn
Bluesky
Facebook
YouTube
X
Other
