Network Working Group
Internet-Draft
Intended status: Experimental
Expires: 21 August 2026

AILEX, Ed.
AILEX Inc. / VeritasChain Standards Organization

17 February 2026

Verifiable AI Provenance (VAP) Framework and Legal AI Profile (LAP)

draft-ailex-vap-legal-ai-provenance-01

Abstract

This document specifies the Verifiable AI Provenance (VAP) Framework, a cross-domain upper framework for cryptographically verifiable decision audit trails in high-risk AI systems, along with the Legal AI Profile (LAP), a domain-specific instantiation for legal AI and LegalTech systems.

VAP defines common infrastructure including hash chain integrity, digital signatures, unified conformance levels (Bronze/Silver/Gold), external anchoring via RFC 3161 Time-Stamp Protocol and compatible transparency services, a Completeness Invariant pattern guaranteeing no selective logging, standardized Evidence Pack format for regulatory submission, and privacy-preserving verification protocols.

LAP extends VAP for the judicial AI domain, addressing unique requirements including attorney oversight verification (Human Override Coverage), three-pipeline completeness invariants for legal consultation, document generation, and fact-checking, tiered content retention with legal hold protocols for judicial discovery compliance, graduated override enforcement mechanisms, and privacy-preserving fields designed to maintain attorney-client privilege while enabling third-party auditability.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 21 August 2026.

Table of Contents

1. Introduction

The deployment of AI systems in high-risk domains -- including finance, healthcare, transportation, and the administration of justice -- creates a structural accountability gap. AI decisions that affect fundamental rights and societal infrastructure lack standardized, cryptographically verifiable audit trails that independent third parties can inspect.

Current approaches rely on trust-based governance: AI providers assert that their systems are safe and well-logged, but no independent party can cryptographically verify these claims. The Verifiable AI Provenance (VAP) Framework addresses this gap by defining a "Verify, Don’t Trust" architecture for AI decision provenance.

This document defines two complementary specifications:

1. VAP Framework (Part I): A cross-domain upper framework defining common infrastructure for verifiable AI provenance applicable to any high-risk AI domain.

2. Legal AI Profile (LAP) (Part II): A domain-specific profile for legal AI systems, addressing requirements arising from professional regulation of attorneys and high-risk AI system governance.

1.1. Scope

VAP targets AI systems where "system failure could cause significant and irreversible harm to human life, societal infrastructure, or democratic institutions." This intentionally strict scope distinguishes VAP from general-purpose logging frameworks.

LAP specifically addresses legal AI systems that provide AI-powered legal consultation, document generation, and fact-checking services to licensed attorneys.

1.2. Design Philosophy

The core principle is "Verify, Don’t Trust." Rather than relying on AI providers’ claims about the safety and integrity of their systems, VAP enables independent, cryptographic verification of every AI decision’s provenance, completeness, and human oversight.

2. Conventions and Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2.1. Terminology

VAP Verifiable AI Provenance Framework - the cross-domain upper framework defined in this document.

Profile A domain-specific instantiation of VAP (e.g., VCP for finance, CAP for content, LAP for legal).

LAP Legal AI Profile - the judicial AI domain profile defined in this document.

Provenance Cryptographically verifiable record of data origin, derivation, and history.

Completeness Invariant A mathematical guarantee that every attempt event has exactly one corresponding outcome event.

Evidence Pack A self-contained, signed package of provenance events suitable for regulatory submission and third-party audit.

External Anchor Registration of a Merkle root hash with an external trusted timestamping service such as [RFC3161] or a compatible transparency log.

Human Override An event recording a human professional's review, approval, modification, or rejection of an AI-generated output.

Override Coverage The ratio of AI outputs reviewed by a human professional to total AI outputs, expressed as a percentage.

Causal Link A reference from an outcome event to its originating attempt event, establishing referential integrity within a pipeline.

Content Retention Tier One of three levels of content preservation (Full Content, Recoverable Hash, Hash-Only) defining the availability of original event content at a given point in the retention lifecycle.

Legal Hold A directive freezing the current retention tier for events within scope, preventing content deletion or tier transition during litigation, investigation, or regulatory inquiry.

Override Enforcement Level One of four graduated control levels (Metric Only, Warn, Gate, Strict) governing how the system responds when AI outputs lack corresponding HUMAN_OVERRIDE events.

3. VAP Framework Architecture

3.1. Layer Model

VAP is organized into four core layers, a common infrastructure layer, and a domain profile layer:

Integrity Layer Hash chain, digital signatures, timestamps (REQUIRED for all levels).

Provenance Layer Actor, input, context, action, and outcome recording (REQUIRED).

Accountability Layer Operator identification, approval chain,

delegation records (REQUIRED for operator_id; RECOMMENDED for approval chain).

Traceability Layer Trace IDs, causal links, cross-profile references (REQUIRED for trace_id; OPTIONAL for cross-references).

Common Infrastructure Conformance levels, external anchoring, completeness invariant, evidence packs, privacy-preserving verification, retention framework (availability depends on conformance level).

Domain Profile Layer Domain-specific event types, data model extensions, regulatory mappings (defined per profile).

3.2. Domain Profiles

VAP supports multiple domain profiles. Each profile MUST define:

1. Event Types: Domain-specific event type taxonomy.

2. Data Model Extensions: Additional fields beyond the VAP common event structure.

3. Conformance Mapping: Mapping to VAP Bronze/Silver/Gold levels.

4. Regulatory Alignment: Mapping to applicable regulations (informative).

5. Completeness Invariant Application: How the completeness invariant applies to domain-specific event flows.

Registered profiles include VCP (Finance), CAP (Content/Creative AI), and LAP (Legal AI, defined in Part II of this document). Additional profiles for automotive (DVP), medical (MAP), and public administration (PAP) domains are under development.

4. Cryptographic Foundation

4.1. Algorithm Requirements

All VAP-conformant implementations MUST support the following cryptographic algorithms:

Category	Primary	Alternative	Post-Quantum (Future)
Hash	SHA-256	SHA-384, SHA-512	SHA3-256
Signature	Ed25519 (RFC 8032)	ECDSA P-256	ML-DSA-65
Encryption	AES-256-GCM	ChaCha20-Poly1305	Kyber-1024

Table 1: Required Cryptographic Algorithms

Implementations MUST include algorithm identifiers in all cryptographic fields to support crypto agility and future algorithm migration.

4.2. Hash Chain Specification

Events MUST be linked in a hash chain where each event's hash includes the hash of the preceding event:

```
EventHash[n] = SHA-256(
  Canonicalize(Event[n] without Signature field)
)

where Event[n].PrevHash = EventHash[n-1]
      Event[0].PrevHash = null  (genesis event)
```

Canonicalization MUST follow RFC 8785 (JSON Canonicalization Scheme).

Chain integrity verification MUST confirm:

1. Each event's hash matches its recomputed hash.

2. Each event's PrevHash matches the preceding event's EventHash.

3. The genesis event has a null PrevHash.

4.3. Digital Signature Requirements

Every event MUST be signed using Ed25519 ([RFC8032]). The signature MUST be computed over the event hash bytes:

```
Signature = Ed25519.Sign(PrivateKey, EventHash_bytes)
Encoded as: "ed25519:" + Base64(Signature)
```

5. Common Event Structure

All VAP-conformant events MUST include the following fields:

```
{
  "vap_version": "1.2",
  "profile": {
    "id": "string (VCP|CAP|LAP|DVP|MAP|PAP|EIP)",
    "version": "semver string"
  },
  "header": {
    "event_id": "UUIDv7 (RFC 9562)",
    "chain_id": "UUIDv7",
    "prev_hash": "sha256:... | null (genesis)",
    "timestamp": "ISO 8601 with timezone",
    "event_type": "string (profile-specific)"
  },
  "provenance": {
    "actor": {
      "actor_id": "string",
      "actor_hash": "sha256:... (privacy-preserving)",
      "role": "string"
    },
    "input": { },
    "context": { },
    "action": { },
    "outcome": { }
  },
  "accountability": {
    "operator_id": "string",
    "last_approval_by": "string",
    "approval_timestamp": "ISO 8601"
  },
  "domain_payload": { },
  "security": {
    "event_hash": "sha256:...",
    "hash_algo": "SHA256",
    "signature": "ed25519:...",
    "sign_algo": "ED25519",
    "signer_id": "string"
  }
}
```

Event identifiers MUST use UUIDv7 ([RFC9562]) to ensure time-ordered sortability. JSON canonicalization MUST follow [RFC8785].

5.1. Numeric Value Encoding

Fields representing monetary amounts, cryptographic values, or high-precision measurements SHOULD be encoded as JSON strings rather than JSON numbers. This recommendation is motivated by:

* IEEE 754 double-precision floating-point, the only numeric type in JSON (per RFC 8259, Section 6), cannot exactly represent all decimal values. Financial and legal contexts require exact decimal representation.

* JSON parsers across programming languages exhibit inconsistent behavior for large integers (exceeding 2^53) and high-precision decimals, leading to silent data corruption.

* Canonicalization stability: [RFC8785] defines specific rules for numeric serialization, but string encoding avoids parser-dependent numeric reformatting entirely, ensuring consistent hash computation across implementations.

Fields where exact precision is not critical (e.g., event_count, token_count) MAY use JSON numbers. Implementations MUST document which fields use string encoding. Implementations that use JSON numbers for counters MUST ensure that any numeric-to-string conversion performed during canonicalization is deterministic and documented, to avoid signature verification ambiguity across languages and libraries.

6. Conformance Levels

VAP defines three conformance levels applicable to all domain profiles. Each level inherits all requirements of lower levels (Gold is a superset of Silver, which is a superset of Bronze).

6.1. Bronze Level

Target: SMEs, early adopters. Core capabilities:

* Event logging for all AI decision points (REQUIRED)

* SHA-256 hash chain linking all events (REQUIRED)

* Ed25519 digital signature on every event (REQUIRED)

* ISO 8601 timestamps with timezone (REQUIRED)

* UUIDv7 event identifiers (REQUIRED)

* Minimum 6-month retention (REQUIRED)

* JSON Schema validation (REQUIRED)

6.2. Silver Level

Target: Enterprise, regulated industries. Additional requirements beyond Bronze:

* Daily external anchoring to a trusted timestamping service conforming to [RFC3161] or an equivalent transparency log (REQUIRED)

* Completeness Invariant verification (REQUIRED)

* Evidence Pack generation capability (REQUIRED)

* Sensitive data hashing for privacy preservation (REQUIRED)

* Minimum 2-year retention (REQUIRED)

* Merkle tree construction for efficient proofs (REQUIRED)

* Third-party verification endpoint (REQUIRED)

6.3. Gold Level

Target: Highly regulated industries. Additional requirements beyond Silver:

* Hourly external anchoring (REQUIRED)

* HSM for signing key storage, FIPS 140-2/3 (REQUIRED)

* Integration with a transparency log service such as IETF SCITT or equivalent (REQUIRED)

* Real-time audit API with sub-second latency (REQUIRED)

* Minimum 5-year retention (REQUIRED)

* 24-hour incident response and evidence preservation (REQUIRED)

* Geographic redundancy, minimum 2 regions (REQUIRED)

* Annual third-party audit (REQUIRED)

* Crypto-shredding support (REQUIRED)

7. External Anchoring

External anchoring proves that events existed at a specific point in time, preventing backdating, forward-dating, and log forking.

7.1. Anchoring Service Types

VAP defines an abstract anchoring interface that can be realized by multiple service types. The baseline anchoring service is [RFC3161] Time-Stamp Authority (TSA). Additional service types include transparency logs and public blockchains.

RFC 3161 TSA (Baseline) Traditional enterprise timestamping via X.509 PKI ([RFC3161]). This is the normative baseline. Trust model: CA trust hierarchy.

Transparency Log (e.g., IETF SCITT) Append-only transparency logs providing public verifiability. IETF SCITT ([IETF-SCITT]) is one such service; implementations MAY use any transparency log providing equivalent guarantees. Trust model: public append-only log.

Blockchain Bitcoin or Ethereum anchoring for maximum decentralization. Trust model: PoW/PoS consensus. This option is non-normative and provided for environments requiring decentralized trust.

Gold Level implementations MUST use at least one transparency log service (such as SCITT) or equivalent, in addition to or instead of RFC 3161 TSA. Implementations SHOULD use multiple independent anchoring services for critical deployments.

7.2. Anchor Record Format

```
{
  "anchor_id": "UUIDv7",
  "anchor_type": "RFC3161 | TRANSPARENCY_LOG | BLOCKCHAIN",
  "merkle_root": "sha256:...",
  "event_count": 1000,
  "first_event_id": "UUIDv7",
  "last_event_id": "UUIDv7",
  "first_event_timestamp": "ISO 8601",
  "last_event_timestamp": "ISO 8601",
  "anchor_timestamp": "ISO 8601",
  "anchor_proof": "Base64-encoded proof",
  "service_endpoint": "https://tsa.example.com"
}
```

7.3. Merkle Tree Construction

Events MUST be batched into a binary Merkle hash tree for efficient anchoring and selective disclosure. The tree construction uses SHA-256 as the hash function and follows a standard binary tree structure:

```
Leaf[i]      = SHA-256(EventHash[i])
Interior[j]  = SHA-256(Left_child || Right_child)
MerkleRoot   = Interior[root]
```

If the number of leaves is not a power of two, the last leaf MUST be duplicated to complete the tree. The resulting Merkle root is submitted to the external anchoring service.

Implementations MAY follow the tree construction specified in [RFC6962] (Certificate Transparency) or any equivalent binary Merkle tree construction that produces deterministic, verifiable inclusion proofs.

Merkle inclusion proofs enable selective disclosure: a verifier can confirm that a specific event is included in an anchored batch without accessing other events in the batch.

8. Completeness Invariant

The Completeness Invariant is a mathematical guarantee that every "attempt" event has exactly one corresponding "outcome" event. This prevents selective logging -- the omission of inconvenient records.

General form:

```
For each pipeline P:
  Count(P_ATTEMPT) = Count(P_SUCCESS)
                   + Count(P_DENY)
                   + Count(P_ERROR)
```

The invariant enforces three properties:

Completeness Every ATTEMPT has an outcome. Violation indicates missing events.

Uniqueness Each ATTEMPT has exactly one outcome. Violation indicates duplicate records.

Referential Integrity Every outcome contains a causal link to its originating ATTEMPT. Violation indicates orphan events.

Domain profiles MUST specify which event types constitute attempts and outcomes for the invariant. Each outcome event MUST contain a causal link field referencing the originating attempt event's identifier. Verification SHOULD account for a configurable grace period for in-flight operations.

9. Evidence Pack Specification

An Evidence Pack is a self-contained, signed package of provenance events suitable for regulatory submission and third-party audit.

9.1. Pack Structure

An Evidence Pack MUST contain:

* manifest.json: Pack metadata and integrity information
* events/: Event batches (max 10,000 events per file)
* anchors/: External anchor records
* merkle/: Merkle tree structure and selective disclosure proofs
* keys/: Public keys for signature verification
* signatures/: Pack-level signature

9.2. Pack Manifest

The manifest MUST include the following fields:

pack_id (REQUIRED) UUIDv7 uniquely identifying this Evidence Pack.

vap_version (REQUIRED) VAP framework version (e.g., "1.2").

profile (REQUIRED) Object containing profile id and version.

conformance_level (REQUIRED) "Bronze", "Silver", or "Gold".

generated_at (REQUIRED) ISO 8601 timestamp of pack generation.

time_range (REQUIRED) Object with start and end ISO 8601 timestamps.

statistics (REQUIRED) Object containing total_events and events_by_type breakdown.

completeness_verification (REQUIRED for Silver+) Object containing invariant_type, invariant_valid boolean, and per-pipeline results.

integrity (REQUIRED) Object containing checksums (SHA-256 per file), merkle_root, and pack_hash.

external_anchors (REQUIRED for Silver+) Array of anchor records referencing this pack's time range.

retention_status (REQUIRED for Silver+ in LAP) Object containing events_at_tier1, events_at_tier2, events_at_tier3 counts, and active_legal_holds count and identifiers. See Section 18.

enforcement_metrics (REQUIRED for Silver+ in LAP) Object containing enforcement_level, warnings_issued, gates_blocked, gates_overridden, and rapid_approvals count and percentage. See Section 16.

The manifest MAY include additional profile-specific fields as defined by the domain profile specification.

10. Privacy-Preserving Verification

VAP enables verification of system integrity without disclosure of sensitive data. This is achieved through:

1. Hash-based attestation: Sensitive fields are stored as cryptographic hashes, enabling existence verification without content disclosure.

2. Selective disclosure via Merkle proofs: Individual events can be proven to exist within an Evidence Pack without revealing other events.

3. Per-tenant salting: Hash salts are unique per tenant to prevent cross-tenant correlation attacks.

This mechanism is particularly critical for LAP, where attorney-client privilege prevents disclosure of consultation content while still requiring verifiable audit trails.

11. Retention Framework

Level	Events	Anchor Records	Evidence Packs	Keys
Bronze	6 months	N/A	On-demand	1 year after last use
Silver	2	5 years	2 years	3 years after

	years			last use
Gold	5 years	10 years	5 years	7 years after last use

Table 2: Retention Requirements by Conformance Level

Retention periods MUST be extended upon: regulatory investigation notification, legal hold orders, security or safety incidents, and third-party audit requests.

Domain profiles MAY specify extended retention periods beyond the VAP baseline where domain-specific regulations require longer retention (see Section 18 for LAP extensions).

For privacy regulation compliance (e.g., [GDPR] "right to be forgotten"), implementations at Silver level and above SHOULD support crypto-shredding: encrypting personal data with per-user keys and deleting those keys to render the data cryptographically unrecoverable while preserving hash chain integrity.

11.1. Content Retention Tiers

Implementations face a tension between privacy-preserving verification (which favors early deletion of sensitive content) and legal discovery obligations (which may require content disclosure). To address this, VAP defines three content retention tiers that domain profiles MAY adopt:

Tier 1 - Full Content Retention All event content (inputs, outputs, documents) is retained in encrypted form alongside provenance hashes. Duration: configurable per tenant.

Tier 2 - Recoverable Hash Retention Original content is deleted but a content recovery key is escrowed with a designated custodian. The escrowed key enables re-association of content from encrypted backups if a legal hold or disclosure order is triggered. Duration: from end of Tier 1 to the applicable retention period endpoint.

Tier 3 - Hash-Only Retention Only provenance hashes remain. Content is cryptographically unrecoverable (crypto-shredding complete). Duration: remainder of the retention period.

Transition from Tier 1 to Tier 2 MUST NOT occur while any Legal Hold is active for the affected events. Transition from Tier 2 to Tier 3 MUST NOT occur while any Legal Hold is active.

Implementations MUST log all tier transitions as RETENTION_TIER_CHANGE events in the provenance chain.

11.2. Legal Hold Protocol

A Legal Hold freezes the current retention tier for all events within scope, preventing content deletion or tier transition.

Legal Hold triggers:

1. Court Order: Judicial order for evidence preservation.
2. Regulatory Investigation: Government regulatory action.
3. Litigation Hold: Reasonably anticipated litigation.
4. Professional Body Inquiry: Investigation by a professional regulatory body (e.g., bar association).
5. Disciplinary Proceedings: Professional disciplinary matters.

When a Legal Hold is activated:

1. All events within scope MUST be frozen at their current retention tier (Tier 1 or Tier 2).
2. A LEGAL_HOLD_ACTIVATED event MUST be recorded in the provenance chain, including: hold_id (UUIDv7), hold_trigger (enumerated trigger type), scope (event time range or case identifiers), activated_by (actor identity hash), and activation_timestamp (ISO 8601).
3. Tier transitions for in-scope events MUST be blocked until a corresponding LEGAL_HOLD_RELEASED event is recorded.
4. The Legal Hold itself MUST be included in Evidence Packs generated during the hold period.

Legal Hold activation MUST be available at Bronze level. Automated Legal Hold detection (e.g., triggered by court filing notifications or regulatory inquiry receipt) is RECOMMENDED at Gold level.

11.3. Judicial Disclosure Response

When a court or regulatory body orders full content disclosure for specific events, the response depends on the current retention tier:

Events at Tier 1 Content is available in encrypted form and can be

disclosed subject to appropriate access controls and professional review.

Events at Tier 2 Content recovery key is retrieved from escrow. Content is reconstructed from encrypted backups. A CONTENT_RECOVERY_EXECUTED event MUST be logged in the provenance chain.

Events at Tier 3 Content is cryptographically unrecoverable. The implementation MUST provide:

1. Hash chain integrity proof demonstrating unbroken provenance.

2. Tier transition log showing when and why content was deleted.

3. Certification that no Legal Hold was active at the time of deletion.

4. Statistical metadata (token counts, timestamps, event types) that remains available.

This three-tier approach enables implementations to demonstrate to judicial authorities that content deletion followed a documented, auditable process rather than constituting potential evidence spoliation.

NOTE: The adequacy of hash-only evidence for specific judicial proceedings is a jurisdiction-specific legal determination outside the scope of this specification. This framework provides the maximum available technical evidence in each retention tier.

12. Third-Party Verification Protocol

Verification is available at three access levels:

Public Access to Merkle roots only. Verifies anchor existence.

Auditor Access to Evidence Packs. Full chain and completeness verification.

Regulator Real-time API access (Gold level). Live monitoring capability.

Verification steps:

1. Anchor Verification: Confirm Merkle root in external timestamping service or transparency log.

2. Chain Verification: Validate hash chain integrity from genesis to latest event.

3. Signature Verification: Authenticate all events with public keys.

4. Completeness Verification: Check invariant for the time period.

5. Selective Query (optional): Verify specific events via Merkle proofs.

13. Legal AI Profile (LAP) Overview

The Legal AI Profile (LAP) is a VAP domain profile for judicial AI and LegalTech systems. LAP addresses unique challenges in the legal domain:

Unauthorized Practice of Law Risk Proving that AI does not independently practice law, through HUMAN_OVERRIDE events documenting attorney oversight.

Hallucination Recording fact-check provenance through LEGAL_FACTCHECK events with citation chain verification.

Selective Logging Preventing omission of inconvenient AI outputs through three-pipeline Completeness Invariant.

Attorney-Client Privilege Maintaining confidentiality through privacy-preserving fields (prompt hashes instead of raw content) and tiered content retention with legal hold support.

Accountability Ambiguity Recording "who, when, and on what basis" through the Accountability Layer and graduated override enforcement.

13.1. Profile Registration

Field	Value
Profile ID	LAP
Full Name	Legal AI Profile
Domain	Legal AI / LegalTech
Regulatory Scope	Attorney regulation, AI governance (informative)

Time Precision	Second
Profile Version	0.3.0

Table 3: LAP Profile Registration

14. LAP Event Type Taxonomy

LAP defines three functional pipelines, one cross-cutting control event type, and administrative event types for retention and enforcement management:

14.1. Pipeline 1: Legal Query

AI-powered legal consultation:

* LEGAL_QUERY_ATTEMPT: Question submission to AI
* LEGAL_QUERY_RESPONSE: AI response generated successfully
* LEGAL_QUERY_DENY: Response refused (content filter, unauthorized role)
* LEGAL_QUERY_ERROR: System error (API failure, timeout)

14.2. Pipeline 2: Document Generation

AI-assisted legal document drafting:

* LEGAL_DOC_ATTEMPT: Document generation request
* LEGAL_DOC_RESPONSE: Document generated successfully
* LEGAL_DOC_DENY: Generation refused (insufficient consent, unauthorized)
* LEGAL_DOC_ERROR: System error (API failure, parse error)

14.3. Pipeline 3: Fact Check

AI-powered legal fact verification:

* LEGAL_FACTCHECK_ATTEMPT: Fact-check request
* LEGAL_FACTCHECK_RESPONSE: Fact-check completed
* LEGAL_FACTCHECK_DENY: Fact-check refused (OPTIONAL)

* LEGAL_FACTCHECK_ERROR: System error

Implementations MAY define LEGAL_FACTCHECK_DENY for cases where a fact-check request is refused due to rate limiting, insufficient permissions, or consent constraints. The deny_reason field SHOULD distinguish these from system errors.

If an implementation does not support LEGAL_FACTCHECK_DENY, refusal conditions MUST be recorded as LEGAL_FACTCHECK_ERROR with a deny_equivalent indicator set to true in the error detail, ensuring the Completeness Invariant is maintained.

14.4. Cross-Cutting: Human Override

HUMAN_OVERRIDE events record an attorney's review of any AI output:

* APPROVE: Attorney confirms AI output without modification

* MODIFY: Attorney edits AI output (modification hash recorded)

* REJECT: Attorney rejects AI output entirely

HUMAN_OVERRIDE events reference the target outcome event via target_event_id (establishing a causal link) and include the attorney's identity (bar number hash), override type, and optional modification details.

14.5. Retention and Enforcement Events

LAP defines additional event types for retention management and override enforcement:

RETENTION_TIER_CHANGE Records transitions between content retention tiers. Fields: previous_tier, new_tier, affected_event_range, reason, authorized_by.

LEGAL_HOLD_ACTIVATED Records activation of a legal hold. Fields: hold_id, hold_trigger, scope, activated_by.

LEGAL_HOLD_RELEASED Records release of a legal hold. Fields: hold_id, released_by, release_reason.

CONTENT_RECOVERY_EXECUTED Records recovery of content from Tier 2 escrow. Fields: hold_id, recovered_event_range, recovered_by, court_order_reference_hash.

REVIEW_WARNING_ACKNOWLEDGED Records that an attorney acknowledged an

unreviewed-output warning before proceeding with export. Fields: target_event_id, warning_type, acknowledged_by.

REVIEW_GATE_BLOCKED Records that an export attempt was blocked due to missing HUMAN_OVERRIDE. Fields: target_event_id, document_type, blocked_action.

REVIEW_GATE_OVERRIDE Records that an attorney bypassed a review gate with explicit justification. Fields: target_event_id, override_reason, overridden_by.

These events are NOT part of the three-pipeline Completeness Invariant (they are control and administrative events, similar to HUMAN_OVERRIDE). However, they MUST be included in the hash chain and signed.

15. LAP Completeness Invariant

LAP applies the Completeness Invariant independently to all three pipelines:

For each pipeline P in {QUERY, DOC, FACTCHECK}:

```
Count(LEGAL_{P}_ATTEMPT)
= Count(LEGAL_{P}_RESPONSE)
+ Count(LEGAL_{P}_DENY)    [if supported]
+ Count(LEGAL_{P}_ERROR)
```

Expanded:

```
LEGAL_QUERY_ATTEMPT = LEGAL_QUERY_RESPONSE
                    + LEGAL_QUERY_DENY
                    + LEGAL_QUERY_ERROR

LEGAL_DOC_ATTEMPT   = LEGAL_DOC_RESPONSE
                    + LEGAL_DOC_DENY
                    + LEGAL_DOC_ERROR

LEGAL_FACTCHECK_ATTEMPT = LEGAL_FACTCHECK_RESPONSE
                        + LEGAL_FACTCHECK_DENY  [if supported]
                        + LEGAL_FACTCHECK_ERROR
```

For implementations that do not support LEGAL_FACTCHECK_DENY, the invariant simplifies to ATTEMPT = RESPONSE + ERROR for Pipeline 3. Refusal conditions recorded as ERROR with deny_equivalent MUST be counted toward the invariant.

Each outcome event MUST contain a causal link field referencing the originating attempt event's identifier, ensuring referential integrity can be verified independently of event ordering.

16. Override Coverage and Enforcement

16.1. Override Coverage Metric

HUMAN_OVERRIDE events are outside the Completeness Invariant but LAP defines Override Coverage as a critical operational metric:

```
Override Coverage =
  Count(HUMAN_OVERRIDE) /
  (Count(LEGAL_*_RESPONSE) + Count(LEGAL_*_DENY))
```

This metric quantifies the degree to which human professionals review AI outputs. In jurisdictions where regulations require that a licensed professional personally scrutinize AI-generated work products, this metric provides measurable evidence of compliance.

Coverage	Assessment	Operational Implication
100%	Ideal	Full professional oversight of all AI outputs
70-99%	Good	Majority reviewed; low-risk outputs may be excluded
30-69%	Warning	Insufficient review; operational improvement recommended
<30%	Critical	Professional oversight requirements likely unmet

Table 4: Override Coverage Assessment

ERROR events are excluded from the denominator because they do not produce an output suitable for professional approval or rejection. Completeness of error handling is evaluated separately via the per-pipeline invariant, where ERROR is a first-class outcome type.

16.2. Enforcement Levels

Override Coverage tracking alone provides post-hoc accountability but does not prevent the use of unreviewed AI outputs in professional practice. LAP defines four enforcement levels to address this structural limitation:

Level 0 - Metric Only Override Coverage is computed and reported. No enforcement action is taken. The system relies on the professional's ethical obligations.

Level 1 - Warn When a user attempts to export, copy, or transmit an AI- generated output that has no corresponding HUMAN_OVERRIDE event, the system MUST display a prominent warning indicating that the output has not been professionally reviewed.

The warning MUST: (a) be displayed in-context at the point of export or copy action; (b) require explicit acknowledgment before proceeding; and (c) record a REVIEW_WARNING_ACKNOWLEDGED event in the provenance chain, including the target_event_id, warning_type, acknowledged_by actor identity, and timestamp.

Level 2 - Gate For designated high-risk document types, the system MUST require a HUMAN_OVERRIDE event before permitting export or transmission.

The gate MUST: (a) block export, copy, and transmit actions until a HUMAN_OVERRIDE (APPROVE or MODIFY) event exists for the target output; (b) record a REVIEW_GATE_BLOCKED event when an export attempt is blocked; and (c) allow the gate to be bypassed ONLY by a user with "attorney" role AND explicit override, recorded as a REVIEW_GATE_OVERRIDE event with a mandatory reason field.

Document types subject to gating SHOULD be configurable per tenant. The default gated types are: court filings (briefs, motions, petitions), settlement and mediation documents, client-facing legal opinions, and contracts and agreements.

Level 3 - Strict ALL AI-generated outputs require HUMAN_OVERRIDE before any export action. No bypass mechanism. This level is intended for maximum-compliance environments but is not mandated by this specification due to potential impact on workflow efficiency.

16.3. Enforcement Level Mapping

Enforcement	Bronze	Silver	Gold
Level 0 (Metric Only)	Default	Minimum	N/A
Level 1 (Warn)	OPTIONAL	REQUIRED	REQUIRED
Level 2 (Gate)	N/A	RECOMMENDED	RECOMMENDED
Level 3 (Strict)	N/A	OPTIONAL	OPTIONAL

Table 5: Enforcement Level by Conformance Level

16.4. Override Latency Threshold

To distinguish genuine professional review from perfunctory approval, LAP defines an Override Latency metric:

```
Override Latency = HUMAN_OVERRIDE.timestamp
                 - target_output_event.timestamp
```

Implementations at Silver level and above SHOULD flag HUMAN_OVERRIDE events with Override Latency below a configurable threshold (RECOMMENDED default: 10 seconds) as "rapid approval" in the provenance chain.

This does NOT block the override but records a RAPID_APPROVAL_FLAG in the event metadata, which: (a) is visible in Evidence Packs and audit reports; (b) may indicate insufficient review depth; and (c) can trigger alerts at Gold level when rapid approvals exceed a configurable percentage (RECOMMENDED: 20%).

16.5. Structural Limitation Acknowledgment

This specification acknowledges that no technical mechanism can fully guarantee the quality or depth of human professional review. A licensed attorney may approve an AI output after genuine review or after cursory inspection; the system cannot distinguish between these without introducing unacceptable surveillance of professional judgment.

The enforcement framework therefore aims to: (a) create friction against entirely unreviewed AI output usage; (b) provide auditable evidence of review or lack thereof; (c) enable risk-proportionate controls (stronger for high-risk document types); and (d) preserve professional autonomy while recording accountability metadata.

The combination of enforcement levels, latency monitoring, and comprehensive provenance logging shifts the accountability model from "trust alone" to "trust with verification infrastructure," acknowledging that while absolute prevention is impossible, the cost and detectability of non-compliance can be substantially increased.

17. LAP Privacy-Preserving Fields

Legal AI handles extremely sensitive data protected by professional privilege. LAP extends VAP's privacy-preserving verification with the following hashed fields:

Original Data	Hash Field	Sensitive Content
User query text	PromptHash	Legal consultation content (privileged)
AI response text	ResponseHash	AI-generated legal advice
Document output	OutputHash	Generated legal documents
Case number	CaseNumberHash	Case identifier (high specificity)
Bar number	BarNumberHash	Professional registration number
Party names	PartyHash	Personal information of parties
Modification detail	ModificationHash	Professional's corrections
Factcheck content	TargetContentHash	Content under verification

Table 6: LAP Privacy-Preserving Fields

Hash computation uses per-tenant salts to prevent cross-tenant correlation. Third-party verifiers can confirm event existence and chain integrity without accessing privileged content.

The availability of original content corresponding to these hashes depends on the current Content Retention Tier (Section 11.1). At Tier 1, original content is available in encrypted form. At Tier 2, it is recoverable via escrowed keys. At Tier 3, only hashes remain.

18. LAP Conformance Level Mapping

Requirement	Bronze	Silver	Gold
Hash Chain	Yes	Yes	Yes
Digital Signature	Yes	Yes	Yes
External Anchoring	No	Daily	Hourly
Completeness Invariant	No	3 Pipelines	3 Pipelines
Override Coverage Tracking	No	Yes	Yes (with alerts)
Override Enforcement Level	Level 0	Level 1 (REQUIRED)	Level 1 (REQUIRED)
Evidence Pack	No	Yes	Yes
Privacy Hashing	No	Yes	Yes
Content Retention Tiers	Tier 3 only	Tier 2 + Tier 3	All 3 Tiers
Legal Hold Protocol	Yes (manual)	Yes (manual)	Yes (automated detection)
Content Recovery Escrow	No	RECOMMENDED	REQUIRED
HSM	No	No	Yes
Retention	6 months	3 years	10 years
Real-time Audit API	No	No	Yes

Table 7: LAP Conformance Matrix

LAP extends the standard VAP retention periods. Silver level requires 3 years (vs. VAP baseline 2 years) and Gold requires 10 years (vs. VAP baseline 5 years). This extension is driven by the longer statutory limitation periods typical in legal proceedings across multiple jurisdictions.

19. LAP Regulatory Alignment (Informative)

This section is entirely informative and non-normative. It illustrates how LAP audit trail capabilities can support compliance with various regulatory frameworks. Legal compliance determinations are jurisdiction-specific and require independent legal analysis.

19.1. Attorney Professional Regulation

Many jurisdictions restrict the practice of law to licensed attorneys. Where AI systems assist attorneys in legal work, regulations may require that the attorney personally review and take responsibility for AI-generated outputs. LAP's HUMAN_OVERRIDE events, Override Coverage metric, and graduated enforcement levels can support demonstrating such oversight.

As an example, the Japanese Ministry of Justice guideline ([MOJ-GUIDELINE]) establishes that AI-based legal services provided to attorneys are permissible when the attorney personally scrutinizes and modifies the output as necessary. LAP audit trails can help meet these expectations through:

* Actor.role and BarNumberHash: supports verification that the user is a licensed attorney.

* HUMAN_OVERRIDE (APPROVE/MODIFY): supports demonstrating attorney scrutiny.

* ModificationHash: supports evidence of attorney modifications.

* Enforcement Level 1/2: provides system-level friction against use of unreviewed outputs.

* Override Latency monitoring: flags potentially insufficient review.

19.2. High-Risk AI System Governance

Legal AI systems may be classified as high-risk under AI governance frameworks such as the EU AI Act ([EU-AI-ACT]), particularly under Annex III "Administration of justice" category. LAP Silver level and above provides audit trail capabilities that can help satisfy record-keeping requirements, including:

* Automatic event logging (supports Article 12 logging requirements)
* Hash chain continuity (supports lifetime recording)
* HUMAN_OVERRIDE events (supports human oversight documentation)
* Causal links between events (supports traceability)
* Tiered content retention with legal hold (supports discovery obligations under Article 12 and national procedural law)

The degree to which these capabilities satisfy specific regulatory requirements should be evaluated on a per-jurisdiction basis.

20. Security Considerations

VAP-LAP implementations face several security considerations:

Key Compromise Compromise of signing keys allows event forgery. Bronze implementations SHOULD rotate keys annually. Silver MUST rotate semi-annually. Gold MUST use HSM storage and quarterly rotation.

Hash Collision Resistance SHA-256 provides 128-bit collision resistance, considered sufficient for the foreseeable future. Implementations MUST support algorithm migration (crypto agility) for post-quantum transition.

Privacy Leakage Per-tenant salting prevents cross-tenant hash correlation. Implementations MUST NOT share salts across tenants. Event metadata (timestamps, event types, counts) may leak statistical information even when content is hashed.

Availability Attacks Denial-of-service attacks against the logging infrastructure could prevent event recording, violating completeness. Gold level implementations MUST have geographic redundancy.

External Anchor Trust The security of external anchoring depends on

the trusted timestamping service. Implementations SHOULD use multiple independent anchoring services for critical deployments.

Completeness Invariant Circumvention An adversary with write access to the event store could insert fabricated ERROR events to satisfy the invariant while hiding actual outcomes. External anchoring at Silver level and above mitigates this by making post-hoc insertion detectable.

Clock and Time Source Integrity Timestamp rollback or clock skew can cause false completeness verification failures and undermine event ordering guarantees. Implementations SHOULD use monotonic time sources and SHOULD cross-validate local timestamps against external anchoring timestamps. External anchoring at Silver level and above provides an independent time reference.

Content Retention and Discovery Risk The tension between privacy-preserving hash-only retention and judicial discovery obligations creates a risk that hash-only evidence may be deemed insufficient by courts. The Tiered Retention model (Section 11.1) mitigates this by maintaining recoverable content during periods when discovery is most likely, while the Legal Hold Protocol (Section 11.2) prevents premature content destruction when litigation is anticipated.

Override Enforcement Circumvention Enforcement Levels 1 and 2 (Section 16.2) can be circumvented by users who copy AI output through means outside the system’s control (e.g., screen capture, manual transcription). Technical enforcement addresses the common case of system-mediated export but cannot prevent all forms of circumvention. The provenance chain nonetheless records the absence of HUMAN_OVERRIDE events, providing post-hoc accountability.

Rapid Approval Gaming Users aware of the Override Latency threshold (Section 16.4) might delay approval clicks to avoid the RAPID_APPROVAL_FLAG. This is a known limitation. Statistical analysis of approval latency distributions across users and document types can detect such gaming patterns in audit reviews.

Legal Hold Integrity Legal Hold activation and release events must be protected from unauthorized modification. Gold level implementations SHOULD require multi-party authorization for Legal Hold release.

21. IANA Considerations

This document has no IANA actions at this time.

Future versions of this document might request registration of a media type for VAP Evidence Pack manifests (e.g., "application/vnd.vap.evidence-pack+json") and an IANA registry for VAP Domain Profile identifiers. Until then, profile identifiers are managed by the VeritasChain Standards Organization (VSO). The initial registered profiles are VCP (Finance), CAP (Content/Creative AI), and LAP (Legal AI).

22. References

22.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8032] Josefsson, S. and I. Liusvaara, "Edwards-Curve Digital Signature Algorithm (EdDSA)", RFC 8032, DOI 10.17487/RFC8032, January 2017, <https://www.rfc-editor.org/info/rfc8032>.

[RFC8785] Rundgren, A., Jordan, B., and S. Erdtman, "JSON Canonicalization Scheme (JCS)", RFC 8785, DOI 10.17487/RFC8785, June 2020, <https://www.rfc-editor.org/info/rfc8785>.

[RFC9562] Davis, K., Peabody, B., and P. Leach, "Universally Unique IDentifiers (UUIDs)", RFC 9562, DOI 10.17487/RFC9562, May 2024, <https://www.rfc-editor.org/info/rfc9562>.

[RFC3161] Adams, C., Cain, P., Pinkas, D., and R. Zuccherato, "Internet X.509 Public Key Infrastructure Time-Stamp Protocol (TSP)", RFC 3161, DOI 10.17487/RFC3161, August 2001, <https://www.rfc-editor.org/info/rfc3161>.

22.2. Informative References

[RFC6962] Laurie, B., Langley, A., and E. Kasper, "Certificate Transparency", RFC 6962, DOI 10.17487/RFC6962, June 2013, <https://www.rfc-editor.org/info/rfc6962>.

[EU-AI-ACT] European Parliament and Council, "Regulation (EU) 2024/1689 - Artificial Intelligence Act", 2024.

[JAPAN-ATTORNEY-ACT] Government of Japan, "Attorney Act (Bengoshi-ho), Act No. 205 of 1949", 1949.

[MOJ-GUIDELINE] Ministry of Justice, Japan, "Regarding the Relationship between AI-based Contract Document Support Services and Attorney Act Article 72", August 2023.

[JFBA-AI-GUIDANCE] Japan Federation of Bar Associations, "Precautions Regarding the Use of Generative AI in Attorney Practice", September 2025.

[IETF-SCITT] IETF SCITT Working Group, "An Architecture for Trustworthy and Transparent Digital Supply Chains", Work in Progress, Internet-Draft, draft-ietf-scitt-architecture, 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-scitt-architecture>.

[GDPR] European Parliament and Council, "Regulation (EU) 2016/679 - General Data Protection Regulation", 2016.

Appendix A. Profile Comparison

Aspect	VCP (Finance)	CAP (Content)	LAP (Legal)
Time Precision	Nanosecond	Second	Second
Key Invariant	SIG to ORD	GEN_ATTEMPT to GEN/DENY/ERROR	3 Pipeline Invariants
Unique Feature	Signal integrity	Safe Refusal (SRP)	Human Override Coverage + Enforcement

Regulatory Focus	Financial regulation	Content regulation	Attorney regulation + AI governance
Privacy Model	Trade data	Creative content	Professional privilege + Tiered Retention
Retention (Gold)	5 years	5 years	10 years

Table 8: Comparison of VAP Domain Profiles

Appendix B. Change Log

This section tracks changes between Internet-Draft revisions and will be removed before publication.

draft-ailex-vap-legal-ai-provenance-01

* Added Tiered Content Retention model (Section 11.1) addressing the tension between privacy-preserving verification and judicial discovery obligations.
* Added Legal Hold Protocol (Section 11.2) with five trigger types and provenance chain integration.
* Added Judicial Disclosure Response procedures (Section 11.3) for court-ordered content disclosure across all retention tiers.
* Expanded Override Coverage (Section 16) with four-level Enforcement Framework (Metric Only/Warn/Gate/Strict) providing graduated controls against unreviewed AI output usage.
* Added Override Latency Threshold (Section 16.4) with Rapid Approval Flag to detect perfunctory review patterns.
* Added Section 16.5 acknowledging structural limitations of technical enforcement of professional review quality.
* Added seven new event types for retention management and enforcement tracking (Section 14.5).
* Extended Evidence Pack manifest with retention_status and enforcement_metrics fields (Section 9.2).

* Extended LAP Conformance Matrix with Content Retention Tiers, Legal Hold Protocol, Content Recovery Escrow, and Override Enforcement Level rows (Section 18).

* Expanded Security Considerations with content retention risk, override enforcement circumvention, rapid approval gaming, and legal hold integrity analysis (Section 20).

* Updated LAP Profile Version from 0.2.0 to 0.3.0.

* Updated Abstract to reflect new capabilities.

draft-ailex-vap-legal-ai-provenance-00 Initial submission.

Acknowledgments

The VAP Framework and LAP Profile were developed with input from: the CAP v1.0 Safe Refusal Provenance (SRP) design experience, the VCP v1.1 operational feedback, regulatory engagement from legal practitioners, and open-source community contributions.

LAP v0.3 design draws from the AILEX SaaS reference implementation, the Ministry of Justice guideline on AI services and [JAPAN-ATTORNEY-ACT] Article 72 (August 2023), the [JFBA-AI-GUIDANCE] on generative AI in attorney practice (September 2025), and operational feedback from pilot law firms regarding judicial discovery and attorney oversight workflows.

Author's Address

AILEX (editor)
AILEX Inc. / VeritasChain Standards Organization
1-10-8 Dogenzaka, Shibuya-ku, Tokyo,
150-0043
Japan
Email: info@ailex.co.jp
URI: https://ailex.co.jp