OVERT

OBSERVABLE VERIFICATION EVIDENCE FOR RUNTIME TRUST

The Working Reference

Glossary · profile reference · rationale · assessment guidance · legal annex · claims grammar · supplementary requirements (Annex G is normative)

DATE	June 2026
PUBLISHED BY	GLACIS Technologies, Inc.
REPRODUCES	Annexes A–G of OVERT 1.1
COMPLETE EDITION	overt.is
CONTACT	overt-review@glacis.io

OFFPRINT NOTICE

This fascicle reproduces Annexes A–G of OVERT 1.1 of OVERT Version 1.1 without modification. Section numbering follows the Complete Edition, which is the sole authoritative text for conformance purposes. Conformance claims cite OVERT 1.1, never an individual fascicle. The Complete Edition and all fascicles are published at overt.is.

Contents of this Volume

ANNEX A: GLOSSARY (INFORMATIVE)
ANNEX B: PROTOCOL PROFILE REFERENCE SUMMARY
B.1 Cryptographic Primitives . . . 9
B.2 Domain Separation and Key Architecture . . . 11
B.3 Canonicalization . . . 11
B.4 Commitment Architecture . . . 12
B.5 Key Hierarchy . . . 12
B.6 Attestation Envelope Architecture . . . 13
B.7 S3P Attestation Schema . . . 14
B.8 Clopper-Pearson Confidence Interval Computation . . . 14
B.9 Receipt Service Architecture . . . 15
B.10 Informative Latency Targets . . . 15
B.11 Informative Default Parameters . . . 15
B.12 Implementation Resources . . . 16
ANNEX C: DESIGN RATIONALE AND CASE STUDIES
C.1 Verification Gaps in High-Stakes AI Deployments . . . 17
C.2 The T.J. Hooper Principle and Potential Standard-of-Care Analysis . . . 18
C.3 Adverse Inference Doctrine and the Duty to Create Records . . . 19
C.4 Consent Attestation and Healthcare AI . . . 20
C.5 Multi-Agent Trust Exploitation . . . 21
C.6 Tiered Certification Analogy . . . 21
C.7 PCI-DSS Contractual Adoption Precedent . . . 22
C.8 FedRAMP and NIST SP 800-53 Adoption History . . . 22
C.9 Insurance Market Interpretation . . . 23
C.10 Non-Egress Architecture and Business Associate Agreement Exposure . . . 23
C.11 Emergent Behavior in Authorized Agentic Systems . . . 24
ANNEX D: RISK SIGNAL FRAMEWORK (INFORMATIVE)
D.1 Signal Properties . . . 28
D.2 Signal Categories . . . 28
D.3 Signal Derivation Requirements . . . 29
D.4 Design Rationale . . . 29
ANNEX E: LEGAL ADMISSIBILITY ANALYSIS (INFORMATIVE)
E.1 Federal Rules of Evidence 902(13): Certified Records of Regularly Conducted Activity (Electronic) . . . 31

E.2 Federal Rules of Evidence 902(14): Certified Data Copied from Electronic Device, Storage Medium, or File . . . 33
E.3 Federal Rules of Evidence 803(6): Business Records Exception to Hearsay . . . 34
E.4 Federal Rules of Civil Procedure 37(e): Failure to Preserve ESI . . . 35
E.5 International Admissibility References . . . 36
United Kingdom: Civil Evidence Act 1995 . . . 36
European Union: eIDAS Regulation (Regulation 910/2014 and eIDAS 2.0) . . . 36
ANNEX F: SAMPLE CITATION LANGUAGE (INFORMATIVE)
F.1 Canonical Conformance Citation Format . . . 38
F.2 Guidance for Referencing OVERT in External Documents . . . 39
F.3 Disclaimer . . . 39
ANNEX G: SUPPLEMENTARY REQUIREMENTS (NORMATIVE — ADDED IN V1.1)
G.1 Local CAS Evidence Retrieval and Retention Integrity . . . 40
G.1.1 Purpose and Scope . . . 40
G.1.2 Evidence Retrieval Interface . . . 41
G.1.3 Proof of Possession . . . 42
G.1.4 Retention Integrity Signal . . . 43
G.2 HTTP Transport Binding for Cross-Boundary Attestation . . . 44
G.2.1 Scope . . . 44
G.2.2 OVERT Context Headers . . . 44
G.2.3 Egress Injection (Upstream Path) . . . 45
G.2.4 Ingress Extraction and Validation (Downstream Path) . . . 46
G.2.5 Header Survival and Reporting . . . 46
G.3 Automated Auditor Discovery and Well-Known Endpoint Protocol . . . 47
G.3.1 Purpose and Scope . . . 47
G.3.2 The OVERT Discovery Document . . . 47
G.3.3 Per-Epoch Artifact Retrieval . . . 48
G.3.4 Auditor Verification Flow . . . 49
G.3.5 Conformance . . . 49
G.4 ControlAction Reference Schema (Informative) . . . 50
G.4.1 Closed Schema . . . 50
G.4.2 Five-Gate Validation (Reference) . . . 52
G.4.3 Test Vectors . . . 52

Annex A: Glossary (Informative)

The following terms and acronyms are used throughout this standard. Where a term has a specific OVERT definition that differs from common usage, the OVERT-specific definition is provided.

Term ID	Term	Definition
A.1	AAL	**Attestation Assurance Level.** One of four tiers (AAL-1 through AAL-4) describing the cryptographic verifiability and independence of governance attestation artifacts. AAL-1: Policy Documentation (self-asserted). AAL-2: Process Records (self-attested, auditor must trust operator). AAL-3: Automated Monitoring (machine-generated, operator-controlled). AAL-4: Cryptographic Attestation (third-party verifiable, zero content access required). See Section 4.1.
A.2	Arbiter	The enforcement sidecar that intercepts AI system actions (tool calls, API requests, data access) and evaluates them against policy before permitting execution. Implemented as an enforcement module; the specific runtime technology is specified by the applicable Protocol Profile. The arbiter generates attestation envelopes for every intercepted action.
A.3	OVERT	**Observable Verification Evidence for Runtime Trust.** This standard.
A.4	Attestation Artifact	A cryptographically signed record produced by the OVERT attestation infrastructure demonstrating that a specific governance control executed at a specific time under a specific configuration. Includes envelopes, receipts, S3P attestations, and ControlActions.
A.5	Attestation Pack	A bundled collection of attestation artifacts sufficient to demonstrate conformance for a defined scope and time period. Includes receipts, transparency log proofs, epoch data, S3P attestations, and ControlActions.
A.6	BLS	**Boneh-Lynn-Shacham.** A pairing-based signature scheme permitting efficient aggregation of multiple signers' signatures into a single compact signature. Used for notary threshold signatures in Protocol Profile 1.0. Not post-quantum resistant; the standard requires hybrid classical + post-quantum constructions, or pure post-quantum constructions, after January 1, 2031. Alternative notary signature constructions (e.g., multi-signature with Ed25519 or ML-DSA) may be specified by other Protocol Profiles.
A.7	CAS	**Content-Addressable Storage.** Local storage within the operator's environment where attestation evidence (prompts, responses, evaluations) is stored indexed by cryptographic commitment. Content never leaves the operator's boundary.

Term ID	Term	Definition
A.8	CBOR	**Concise Binary Object Representation.** Binary data serialization format (RFC 8949). Protocol Profile 1.0 uses CBOR deterministic encoding per Section 4.2 of RFC 8949 for canonical byte-level representation. The standard requires deterministic encoding as a property (Section 17.1); the specific format — CBOR, JSON via JCS (RFC 8785), or other deterministic encoding — is specified by the applicable Protocol Profile.
A.9	CI (Confidence Interval)	A statistical interval computed using the Clopper-Pearson exact method providing upper and lower bounds on a violation rate with guaranteed coverage probability. Used in S3P attestations.
A.10	Co-epoch Binding	The cryptographic binding of an attestation receipt to the binary identity, network isolation state, and configuration of the system under attestation at the time the attestation was produced. Ensures that attestations cannot be replayed across different system configurations.
A.11	ControlAction	A cryptographically attested record of a governance response to a detected violation. Includes action type, timestamp, scope, and co-epoch binding.
A.12	DPL	**Digest Publication Ledger.** A per-epoch publication of request commitments (never raw digests) enabling auditor verification of sampling fairness without content access.
A.13	Epoch	A bounded time interval (configurable; recommended default: 300 seconds) during which attestation parameters remain constant. Epoch boundaries trigger nonce publication, key rotation, and S3P computation.
A.14	HKDF	**HMAC-based Key Derivation Function.** Key derivation per RFC 5869. Protocol Profile 1.0 uses HKDF for deriving tenant_pepper, storage_key, sampling_key, and epoch_secret from root secrets within the split-knowledge key hierarchy. The standard requires key derivation as specified by the applicable Protocol Profile.
A.15	HMAC	**Hash-based Message Authentication Code.** Keyed hash function per RFC 2104. Protocol Profile 1.0 uses HMAC for request commitments, evidence commitments, PRF tags, and S3P sampling tags with domain separation prefixes. The standard requires keyed commitment functions as specified by the applicable Protocol Profile.
A.16	IAP	**Independent Attestation Provider.** An entity structurally independent of the AI system operator that operates notary infrastructure, validates attestations, and publishes transparency log entries. An IAP does not access protected content. Multiple IAPs may operate under different governance models.
A.17	Base Envelope	The baseline attestation envelope emitted for every AI request. Contains 9 fields including blinded identifier, request commitment, encoder binary identity, and

Term ID	Term	Definition
		metadata. Closed schema — no additional fields permitted. See Annex B for architecture; full field-level schema is in the Protocol Profile.
A.18	Extended Envelope	The extended attestation envelope emitted for sampled requests. Contains 10 fields including the full PRF tag for auditor recomputation and policy evaluation scores. Closed schema. See Annex B for architecture; full field-level schema is in the Protocol Profile.
A.19	NETATT	**Network Attestation.** A per-epoch attestation of the system's network isolation state, covering at minimum the effective egress policy, the enforcement component identity, and the TLS certificate pin set; operators may include additional deployment-specific inputs such as network policy controller identity, eBPF state, CNI configuration, and environment variables affecting AI behavior (Section 18.3). Bound to all receipts issued during the epoch.
A.20	OSCAL	**Open Security Controls Assessment Language.** NIST-developed machine-readable format for security control documentation. OVERT attestation packs are expressible as OSCAL Assessment Results.
A.21	POST_HOC Receipt	An attestation receipt generated retroactively after a fail-open period. POST_HOC receipts are reconstruction artifacts excluded from contemporaneous attestation coverage for conformance, risk-signal reporting, and litigation reporting purposes — the normative exclusion is stated in RES-5.2 and Section 19. Distinguishable from contemporaneous receipts in all export packages and signal computations.
A.22	PRF	**Pseudorandom Function.** A deterministic function (HMAC-SHA256 in Protocol Profile 1.0) used to determine whether a given request falls within the attestation sample. Operates on request commitments, not raw content.
A.23	Protocol Profile	A registered implementation specification defining cryptographic constructions, envelope schemas, key derivation methods, and receipt formats that implement this standard. Multiple profiles may coexist. Conformance requires exactly one registered profile per deployment. See Annex B for Protocol Profile 1.0 summary.
A.24	RATS	**Remote ATtestation procedureS.** IETF architecture for remote attestation (RFC 9334). OVERT attestation architecture is complementary to RATS, with roles mapping to Attester (arbiter), Verifier (notary), and Relying Party (auditor/insurer).
A.25	Receipt	A cryptographically signed record issued by the notary service proving that a specific attestation envelope was received, validated, and recorded during a specific epoch. The base receipt contains 9 fields including attestation_hash, epoch binding, binary identity, network state, flags (contemporaneous vs. POST_HOC), and transparency log proofs; for cross-boundary workflows the registered Protocol Profile defines an extended receipt type additionally carrying the Section 4.8

Term ID	Term	Definition
		parent-reference fields. Each receipt type is a closed schema. See Annex B for architecture; full field-level schemas are in the Protocol Profile.
A.26	S3P	**Statistical Safety Signal Protocol.** The normative auditor-reproducible sampling and measurement method defined in Section 9 (MEASURE). Uses commit-then-reveal epoch nonces, keyed-function-based sampling (HMAC in Protocol Profile 1.0), and Clopper-Pearson exact confidence intervals to produce statistically rigorous safety signals.
A.27	Severity Class	A classification of governance violations by severity, defined in GOV-3.2. Maps to risk-signal computation and response escalation requirements.
A.28	Split-Knowledge Key Hierarchy	The key management architecture in which content-binding keys (operator-managed, e.g., tenant_pepper) and sampling/identity keys (platform-managed, e.g., sampling_key, epoch_secret) are held by different parties. Ensures that routine audit is a zero-content-knowledge operation.
A.29	SPKI	**Subject Public Key Info.** DER-encoded public key information used for TLS certificate pinning in NETATT.
A.30	STH	**Signed Tree Head.** A signed commitment to the current state of the transparency log Merkle tree, enabling split-view detection.
A.31	SUT	**System Under Test.** The AI system being governed. The attestation system treats the SUT as untrusted — self-reports from the SUT are insufficient for AAL-4 conformance. This designation is specific to the OVERT attestation relationship and is distinct from the NIST SP 800-207 Zero Trust Architecture for network security.
A.32	TEVV	**Test, Evaluation, Verification, and Validation.** The systematic process of evaluating AI system performance, safety, and governance compliance. OVERT provides the attestation infrastructure for TEVV activities.
A.33	Transparency Log	An append-only, cryptographically verifiable log (per RFC 6962) in which attestation receipts and S3P attestations are recorded. Supports inclusion proofs (proving a receipt is in the log), consistency proofs (proving the log has not been tampered with), and signed tree heads for split-view detection.
A.34	Notary Network	One or more notary nodes operated under a published governance model (ATT-5.1) that validate attestations on behalf of relying parties. Where multiple nodes are deployed, t-of-n agreement is required before a valid receipt can be issued and no single node can unilaterally issue or suppress a receipt. A single structurally independent notary satisfies the AAL-4 independence requirement but not the AAL-4 resilience requirement; single-IAP deployments disclose this limitation per Section 4.7.1 and the conformance statement grammar (Section 22.4), where the requirement is stated. Multi-entity sets provide both indepen-

Term ID	Term	Definition
		dence and resilience. The signature or verification construction achieving the t-of-n property (where applicable) is specified in the registered Protocol Profile.
A.35	Scanner	A runtime monitoring sidecar or component that inspects inputs, outputs, and intermediate states of an AI system to detect policy violations, security threats, or behavioral drift. It works in conjunction with the arbiter.
A.36	Local Classifier	A local evaluation component that runs classification or inference models to categorize inputs, outputs, or agent behaviors for policy decision-making.
A.37	Proof of Possession (PoP)	A challenge–response operation by which an operator demonstrates, without egressing protected content, that it retains the evidence payload bound to a given `evidence_commitment` at the time of challenge. Defined in Annex G, Section G.1.3.
A.38	OVERT Discovery Document	A machine-readable document published under the `/.well-known/` URI namespace (RFC 8615) that enumerates the endpoints from which an auditor retrieves the cryptographic artifacts required for independent verification. Defined in Annex G, Section G.3.

Annex B: Protocol Profile Reference Summary

> ***Informative —*** *Protocol Profile 1.0 is the initial registered profile authored by GLACIS Technologies. Additional profiles may be submitted by any party meeting the registration requirements defined in Section 22.6. This annex summarizes Protocol Profile 1.0 for reference. The authoritative specification is the Protocol Profile document itself.*

B.1 Cryptographic Primitives

Protocol Profile 1.0 specifies the following cryptographic constructions:

Primitive	Usage	Specification
SHA-256	All digests, binary hashes, commitments	FIPS 180-4
HMAC-SHA256	PRF tags, request/evidence commitments, bearer tokens, S3P sampling	RFC 2104
HKDF-SHA256	Key derivation for both operator and platform key hierarchies	RFC 5869
Ed25519	Arbiter signatures, notary signatures, controller signatures	RFC 8032
Notary signature (BLS threshold in Profile 1.0)	Notary network t-of-n verification	draft-irtf-cfrg-bls-signature (construction selected by Protocol Profile 1.0; the core standard requires only the t-of-n trust property and permits alternative constructions including multi-signature schemes)
Deterministic encoding (CBOR in Profile 1.0, JCS for JSON profiles)	Canonical encoding of attestation structures	RFC 8949, Section 4.2 (CBOR); RFC 8785 (JCS)

Primitive	Usage	Specification
Merkle trees	Transparency log inclusion and consistency proofs	RFC 6962
Clopper-Pearson	Exact binomial confidence intervals for S3P	Standard statistical method

> ***Non-Normative Note on Post-Quantum Cryptography:*** *A post-quantum migration path for the signature layer using ML-DSA-65 (FIPS 204) is under development. This is a forward-looking implementation note only and does not alter the normative requirements or conformance level definitions of this version.*

Post-quantum migration path: Protocol Profile 1.0 uses BLS threshold signatures and Ed25519 single-signer signatures, both of which rely on the computational Diffie-Hellman problem and are vulnerable to quantum attack via Shor's algorithm. Section 18 requires that after January 1, 2031, conformant implementations use hybrid classical + post-quantum constructions or pure post-quantum constructions. For single-signer operations, the recommended migration is ML-DSA (FIPS 204) + Ed25519. For notary network operations, Protocol Profiles using multi-signature constructions can migrate each notary independently to ML-DSA; profiles using threshold signatures require threshold-compatible post-quantum schemes. Pure classical signature schemes become non-conformant after that date; this is OVERT's own profile-binding cutoff, informed by the NIST IR 8547 (draft) transition timeline, which deprecates quantum-vulnerable classical signatures after 2030 and disallows them after 2035.

IETF RATS alignment and EAT forward reference: OVERT attestation envelopes are structurally aligned with the IETF RATS architecture (RFC 9334). The Entity Attestation Token (EAT, RFC 9711) defines a CBOR/JSON token format for attester-generated claims about entity identity and state. Future Protocol Profile revisions should evaluate expressing OVERT attestation envelopes as EAT profiles, enabling interoperability with the broader RATS ecosystem. In particular, an Agentic AI EAT Capability Attestation profile — encoding arbiter binary identity, co-epoch binding, and capability-scoped policy claims as EAT claims — would position OVERT attestation artifacts for direct consumption by EAT-aware verifiers and relying parties within the IETF trust model. This alignment is a design goal for Protocol Profile 2.0 and does not affect Protocol Profile 1.0 conformance.

SCITT alignment (informative). OVERT transparency logs implement the same pattern the IETF SCITT (Supply Chain Integrity, Transparency and Trust) architecture generalizes: an append-only, independently auditable log of signed statements. A registered Protocol Profile MAY operate an OVERT transparency log as a SCITT transparency service, registering receipts as signed statements

and exposing SCITT receipts as inclusion evidence; this is targeted for Protocol Profile 2.0 alongside the EAT alignment above. Relatedly, the build-pipeline-compromise threat (Section 4.5) is addressed in the supply-chain ecosystem by in-toto and SLSA provenance attestations, which a deployment MAY bind to the arbiter binary identity verified under ATT-2.2.

B.2 Domain Separation and Key Architecture

Protocol Profile 1.0 uses versioned domain separation prefixes on all HMAC operations to prevent cross-protocol attacks. Each HMAC operation — request commitment, evidence commitment, sampling PRF, epoch bearer token, and S3P sampling — uses a distinct prefix with a version suffix enabling future protocol evolution.

The split-knowledge key hierarchy ensures that content-binding keys (operator-managed) and sampling/identity keys (platform-managed) are held by different parties. This separation prevents any single party from both reversing content AND verifying sampling fairness. The specific prefix strings, salt values, and key derivation parameters are specified in the Protocol Profile.

B.3 Canonicalization

Protocol Profile 1.0 canonicalizes all OVERT messages per RFC 8949 Section 4.2 (Deterministic Encoding). Two conformant encoders encoding the same logical data must produce identical byte sequences, which is the prerequisite for all hash-based verification. The standard requires deterministic canonicalization as a property (Section 17.1); the specific encoding format is specified by the applicable Protocol Profile. Protocol Profiles using JSON are expected to specify JCS (RFC 8785) for deterministic canonicalization; Protocol Profiles using CBOR are expected to specify RFC 8949 Section 4.2.

Protocol Profile 1.0 prohibits IEEE-754 floating-point numbers in attestation envelopes, requiring scaled integers for deterministic cross-platform hashing. The S3P schema uses decimal strings for rates and bounds. Timestamps use uint64 nanoseconds since Unix epoch. Indefinite-length encodings, NaN, and +/-Inf are rejected. Protocol Profiles using JSON encodings are expected to specify equivalent numeric safety requirements (e.g., string-encoded decimals, integer-only numeric fields, or explicit precision annotations) to satisfy the numeric losslessness property of Section 17.1.1.

B.4 Commitment Architecture

Protocol Profile 1.0 defines a layered commitment architecture:

- **Request commitments** are computed by HMAC over the content digest using an operator-managed key derived from the operator's root secret via HKDF. The content digest stays local; only the HMAC commitment crosses the trust boundary.
- **Evidence commitments** follow the same pattern for policy evaluation evidence.
- **PRF sampling tags** are computed using a platform-managed key, operating on the request commitment (not the raw content digest). This ensures auditors can verify sampling fairness without holding content-reversing keys.

The specific HMAC constructions, HKDF derivation parameters, and key hierarchy are specified in the Protocol Profile.

Critical constraint: Operator-managed content-binding keys never leave the operator's environment. Deep audits requiring content verification are conducted on-premises under operator control and legal authority.

B.5 Key Hierarchy

The split-knowledge key hierarchy has two branches:

Operator-managed keys (content binding): Derived from an HSM-backed root secret. Includes keys for content commitment and local storage encryption. These keys never cross the operator's trust boundary.

Platform-managed keys (identity and sampling): Derived from an HSM-backed root secret managed by the notary network operator. Includes keys for sampling, epoch management, and notary signing. In Protocol Profile 1.0, notary signing keys are BLS threshold shares distributed across notary nodes; other Protocol Profiles may use per-notary signing keys with independent key management.

Forward secrecy: Epoch-scoped keys are deleted after the subsequent epoch begins. Compromise of a current epoch secret does not reveal past epoch secrets.

Recovery: Shamir k-of-n (recommended: 3-of-5) across geographic regions or HSMs for operator root secrets. Platform key recovery is specified in the registered Protocol Profile (Protocol Profile 1.0 uses BLS threshold key shares across notary nodes; multi-signature profiles use standard per-node key backup procedures).

The specific HKDF derivation paths, salt values, and key tree structure are specified in the Protocol Profile.

B.6 Attestation Envelope Architecture

Protocol Profile 1.0 defines three closed-schema structures:

Base Envelope (all requests — 9 fields): Emitted for every in-scope AI action. Contains a blinded identifier, request commitment, encoder binary identity, non-content metadata, monotonic counter, nanosecond timestamp, key identifier, arbiter instance identifier, and signature. No additional fields are permitted (closed schema).

Extended Envelope (sampled requests — 10 fields): Emitted for requests selected by the sampling PRF. Contains a reference to the matching Base Envelope, request and evidence commitments, the full PRF tag for auditor recomputation, policy evaluation scores, monotonic counter, timestamp, key and arbiter identifiers, and signature. Closed schema.

Receipt (9 fields, issued by notary service): Contains the attestation hash (cryptographic digest of the submitted envelope), validated epoch, notary-derived binary hash, network state hash, monotonic counter, issuance timestamp, flags (contemporaneous vs. POST_HOC), notary signature (single-signer, multi-signature, or threshold, as specified in the Protocol Profile), and transparency log proofs (inclusion proof, consistency proof, signed tree heads). Closed schema. A receipt's `attestation_id`, as referenced in Section 4.8 and Annex G, is its attestation-hash field under the name reflecting its role as a cross-boundary reference key. For cross-boundary workflows, the Protocol Profile defines an **extended receipt type** that additionally carries the Section 4.8 parent-reference fields (`parent_attestation_id`, `parent_reference_status`); each receipt type is itself a closed schema.

The `flags` field distinguishes contemporaneous receipts (`0x00`) from POST_HOC receipts (`0x01`, generated after a fail-open period per RES-5.2). Auditors and risk-signal computations filter on this field to separate contemporaneous attestation from retroactive reconstruction.

The complete field-by-field schemas with types, constraints, and signature scopes are specified in the Protocol Profile.

B.7 S3P Attestation Schema

The S3P attestation schema is a 14-field closed structure capturing all data needed for auditor-reproducible safety verification. Every field is necessary and sufficient for independent recomputation of statistical bounds.

The schema includes: epoch identifier, violation type, total and sampled request counts, sampling and observed rates (decimal strings to avoid IEEE-754 variance), confidence level, Clopper-Pearson lower and upper bounds (decimal strings), sampling threshold, epoch nonce commitment, status indicator, and notary signature.

The three status values are: `"OK"` (valid computation), `"ERR_INSUFFICIENT_SAMPLE"` (sample size below minimum), and `"ERR_NONCE_NOT_PUBLISHED"` (verification failure — epoch nonce was not published after epoch close).

The complete schema with field types and encoding rules is specified in the Protocol Profile.

B.8 Clopper-Pearson Confidence Interval Computation

The Clopper-Pearson method provides exact (not approximate) binomial confidence intervals with guaranteed coverage probability. It provides exact binomial interval coverage under the S3P sampling model and remains valid for small sample sizes without normal-approximation assumptions. The upper bound is conservative by construction.

Given `k` violations observed in `n` sampled requests at confidence level `1 - alpha`:

```
CI_lower = Beta_inv(alpha/2;     k,     n - k + 1)     for k > 0, else 0
CI_upper = Beta_inv(1 - alpha/2; k + 1, n - k)         for k < n, else 1
```

Where `Beta_inv(p; a, b)` denotes the p-th quantile of the Beta distribution with shape parameters `a` and `b`.

Properties:

- Exact coverage: P(p_true in [CI_lower, CI_upper]) >= 1 - alpha for all p_true
- Conservative: The interval is wider than approximate methods (Wald, Wilson), never narrower
- Valid for all sample sizes, including small samples
- No distributional assumptions required

B.9 Receipt Service Architecture

The receipt service accepts a closed-schema request containing only a hash and an epoch identifier, and returns a signed receipt. The API schema enforces the non-egress architecture at the protocol level: the service is structurally incapable of receiving content because its schema does not contain fields for content. Unknown fields are rejected.

This constraint is architectural, not merely a validation rule. The receipt service API schema is specified in the Protocol Profile.

B.10 Informative Latency Targets

The following latency targets are informative recommendations for Protocol Profile 1.0. Specific latency requirements are deployment-dependent and are not normative requirements of the standard.

Phase	Operation	Informative Target
Phase 1 — Enforcement	Local policy evaluation	< 5 ms P50
Phase 1 — Enforcement	Distributed policy evaluation	< 25 ms P50
Phase 2 — Attestation	Receipt round-trip	< 50 ms P50
Phase 3 — Commitment	Transparency log inclusion	< 100 ms P95

Total overhead (enforcement + attestation): informative target < 50 ms P50, which is negligible relative to typical LLM inference latency (500-5000 ms).

B.11 Informative Default Parameters

The following default parameters are informative recommendations for Protocol Profile 1.0. Operators configure these values according to their deployment requirements.

Parameter	Informative Default	Notes
Epoch duration	300 seconds (5 minutes)	Configurable per deployment policy
Tool-call recursion depth	25	Configurable threshold defined in deployment policy
Clock skew tolerance	<= 2 seconds	Bounded skew tolerance; stale submissions rejected
Override review SLA	Within operator-defined SLA	Recommended: 24 hours

Parameter	Informative Default	Notes
TEVV testing interval	Per operator's risk management policy	Not to exceed 12 months or as required by applicable regulation

B.12 Implementation Resources

Protocol Profile 1.0 includes CBOR diagnostic notation examples, reference test vectors for S3P computation, and auditor verification procedures. These materials enable implementers to validate their implementations against known-good results. Protocol Profiles using other encodings are expected to provide equivalent notation examples and test vectors in their respective formats.

Reference test vectors and implementation examples are available in the Protocol Profile document. Organizations implementing OVERT using Protocol Profile 1.0 should obtain the Protocol Profile from the OVERT Protocol Profile Registry.

Annex C: Design Rationale and Case Studies

> ***Informative —*** *This annex provides design rationale and contextual analysis. It describes legal, operational, and institutional conditions relevant to the standard's development. It does not impose requirements on implementers or assert legal conclusions. The normative requirements of the standard are specified in Parts 1-5. It is structured as Design Decision, Rationale, and Supporting Analysis.*

C.1 Verification Gaps in High-Stakes AI Deployments

Design Decision: OVERT requires independent, third-party verifiable attestation (AAL-4) for governance controls in high-stakes deployments, rather than relying on self-attestation or contractual governance alone.

Rationale: Contractual governance has proven structurally insufficient as the sole enforcement mechanism for AI safety controls. When disputes arise between AI system providers and their customers over safety control execution, neither party can independently verify what controls actually ran if no attestation infrastructure exists.

Supporting Analysis: In early 2026, a series of disputes between major AI laboratories and government agencies demonstrated this proof gap with extraordinary clarity. In one instance, an AI company insisted on contractual red lines regarding prohibited uses, while the government customer demanded unrestricted operational access. Neither party could independently verify whether AI use complied with stated restrictions during operational deployment. The dispute was adjudicated through contract negotiations, leaked internal memoranda, public conference statements, and executive action — rather than through independent verification of actual system behavior.

Simultaneously, competing AI laboratories publicly accused each other of inadequate safety practices, with characterizations ranging from "safety theater" to "mendacious" claims about governance controls. These mutual accusations could not be independently adjudicated because no party had deployed infrastructure capable of producing verifiable records of what safety controls actually

executed on any given interaction. The disputes were resolved — or remain unresolved — through political, commercial, and reputational channels rather than through technical verification.

This pattern illustrates a structural problem: contractual governance produces assertions about intended behavior, not verifiable records of actual behavior. When the only evidence of safety control execution is the operator's own claims, disputes become contests of credibility rather than questions of fact. If verification technology becomes commercially deployable at scale, the continued reliance on unverifiable self-attestation may become relevant to courts, regulators, and insurers evaluating evidentiary and governance posture under applicable legal and policy frameworks.

OVERT addresses this gap by specifying how to produce tamper-evident, independently verifiable, temporally bound proof that AI governance controls executed — without exposing protected content.

C.2 The T.J. Hooper Principle and Potential Standard-of-Care Analysis

Design Decision: OVERT is designed as an open standard that can serve as one reference point in discussions of verifiable AI governance, recognizing that courts — not industries — ultimately determine the standard of care.

Rationale: The T.J. Hooper principle holds that an entire industry can be found negligent for failing to adopt available safety technology, regardless of industry custom.

Supporting Analysis: In The T.J. Hooper, 60 F.2d 737 (2d Cir. 1932), Judge Learned Hand held that tugboat operators were negligent for failing to carry radio receivers that would have warned of an approaching storm — even though no tugboat company used radios at the time. The court stated: "a whole calling may have unduly lagged in the adoption of new and available devices… Courts must in the end say what is required; there are precautions so imperative that even their universal disregard will not excuse their omission."

The principle has been applied consistently for nearly a century. The RAND Corporation's report on AI tort liability explicitly cited T.J. Hooper, noting that "courts can still find AI companies negligent even if they did follow industry custom" and that safety-conscious companies developing standards "could establish benchmarks for the whole industry in future litigation."

For AI governance, the implication is narrower. If cryptographic attestation technology becomes commercially deployable and operationally mature, failure to consider its adoption could become relevant to negligence analysis, depending on jurisdiction, commercial availability, deployment maturity, and the surrounding facts. An open standard may strengthen that analysis by documenting

an interoperable approach, but publication of a standard alone does not establish that the technology is available, required, or legally obligatory.

The English equivalent is Bolitho v. City and Hackney Health Authority [1998] AC 232, where Lord Browne-Wilkinson held that courts may reject professional custom as unreasonable if "the professional opinion is not capable of withstanding logical analysis." The Australian statutory calculus under Section 5B of the Civil Liability Act 2002 (NSW) reaches the same outcome through explicit consideration of the probability of harm, seriousness of harm, burden of precautions, and social utility. The German doctrine of Verkehrssicherungspflichten requires anyone who creates or controls a potential source of danger to take necessary precautions.

C.3 Adverse Inference Doctrine and the Duty to Create Records

Design Decision: OVERT Section 21 (Legal Preservation and Production) requires retention policies, legal hold procedures, immutable export capabilities, and chain-of-custody metadata — addressing the risk that operators could "define away bad evidence" through self-serving retention policies.

Rationale: The adverse inference doctrine permits factfinders to draw unfavorable conclusions when a party fails to preserve records it had the available technology to produce. For AI systems, where the "black box" nature makes governance documentation critical, the absence of attestation technology may be relevant to evidentiary analysis. Whether the failure to deploy attestation creates a substantive claim or an evidentiary disadvantage depends on jurisdiction, applicable duty, commercial availability of attestation technology, and the specific facts. This rationale identifies the doctrinal relevance; it does not assert a specific legal outcome.

Supporting Analysis: Under FRCP 37(e), if electronically stored information that should have been preserved is lost because a party failed to take reasonable steps to preserve it, the court may order measures no greater than necessary to cure the prejudice. Upon finding that a party acted with the intent to deprive another party of the information's use in the litigation, the court may presume that the lost information was unfavorable to the party, instruct the jury that it may or must presume the information was unfavorable, or dismiss the action or enter a default judgment.

In Zubulake v. UBS Warburg, 229 F.R.D. 422 (S.D.N.Y. 2004), failure to preserve digital evidence resulted in a $29.2 million verdict including $21.1 million in punitive damages. The court granted an adverse inference instruction: "if you find that UBS could have produced this evidence… you are permitted, but not required, to infer that the evidence would have been unfavorable to UBS."

Valcin v. Public Health Trust of Dade County, 473 So.2d 1297 (Fla. 3d DCA 1984), provides the closest doctrinal analogue. Where a hospital's file failed to contain an operative note, the court imposed a rebuttable presumption of negligence and shifted the burden of proof for records that should have been created pursuant to a duty. If industry standards (NIST AI RMF, ISO/IEC 42001) and available technology create a de facto duty to record AI safety control execution, the failure to deploy attestation technology triggers a parallel adverse presumption.

OVERT Section 21 directly mitigates this risk by requiring operators to define retention policies, implement legal hold procedures, and maintain export capabilities — ensuring that attestation artifacts are available when needed for legal proceedings, regulatory investigations, or insurance claims.

C.4 Consent Attestation and Healthcare AI

Design Decision: OVERT HITL-1 requires cryptographic attestation of patient consent in healthcare AI deployments, with consent receipts that are independently verifiable.

Rationale: AI systems that generate their own compliance records without actual human attestation create a novel and dangerous category of false documentation.

Supporting Analysis: Recent class-action litigation regarding an ambient clinical documentation system deployed without all-party consent illustrates this risk. The complaint alleged violations of the California Invasion of Privacy Act (CIPA) and the Confidentiality of Medical Information Act (CMIA).

The most significant allegation: the AI tool allegedly inserted false statements into patient charts claiming patients "were advised" and "consented" to recording when they had not. This represents AI systems generating their own false compliance documentation — a pattern that conventional audit methods (reviewing the documentation itself) cannot detect, since the documentation asserts the very compliance whose absence it conceals.

Estimates suggest 100,000+ patient encounters may have been affected. The legal theories — wiretapping, unauthorized third-party disclosure, false consent documentation, retention failures — represent patterns emerging across healthcare AI deployments.

OVERT consent attestation addresses this by requiring that consent events be cryptographically attested with independent verification: the consent receipt is signed by the notary network, not generated by the AI system itself. The receipt proves that a consent interaction occurred at a specific time, was recorded through a specified mechanism, and was attested by an independent party.

This approach also responds to the California Invasion of Privacy Act's requirement for "all party" consent to recording, as well as SB 53's incident reporting requirements effective January 1, 2026.

C.5 Multi-Agent Trust Exploitation

Design Decision: OVERT Sections 11-16 (Agentic AI Controls) require per-call attestation, capability-based access control, and multi-agent trust boundary enforcement.

Rationale: Multi-agent AI systems exhibit systematic vulnerability to trust exploitation through prompt injection, tool misuse, and cross-agent privilege escalation.

Supporting Analysis: In a 2025 evaluation of 17 state-of-the-art models, 82.4% executed malicious commands when requested by peer agents — even where they resisted the identical instruction delivered directly — demonstrating inter-agent trust exploitation as the dominant agentic attack surface, alongside prompt injection through shared context, capability escalation via delegated tool access, and information exfiltration through inter-agent communication channels (The Dark Side of LLMs: Agent-based Attacks for Complete Computer Takeover, arXiv:2507.06850, 2025). The CaMeL framework (Google DeepMind, 2025) proposed capability-based prompt injection defense through separation of privileged and quarantined execution contexts — a design pattern that OVERT formalizes through per-call attestation and capability-scoped access control.

The "policy-quality gap" is particularly acute in multi-agent systems: an attestation system that faithfully records and attests to the outputs of a compromised agent produces cryptographically valid records of invalid outputs. OVERT addresses this through capability-based access control (CAP-1, CAP-2) that constrains what each agent can do, combined with per-call attestation (TOOL-1 through TOOL-5) that records what each agent actually did. The combination enables forensic reconstruction of multi-agent interactions and detection of capability violations even when individual agent outputs are compromised.

C.6 Tiered Certification Analogy

Design Decision: OVERT is structured as a tiered standard (AAL-1 through AAL-4) with progressive requirements, permitting organizations to adopt attestation incrementally while establishing a clear ceiling (AAL-4) for the highest assurance tier.

Rationale: Tiered certification avoids all-or-nothing adoption barriers and creates a progressive path toward comprehensive governance. Building-certification systems provide a useful structural analogue.

Supporting Analysis: The relevant lesson from tiered certification systems is structural, not economic: progressive assurance levels can lower adoption friction, and separation between a standard-setter and an independent verifier can improve trust in published claims. For OVERT, the

relevant point is narrower: tiered adoption and separation between standard-setting and independent verification can accelerate uptake without changing the standard's underlying technical claims. External regulatory, contractual, or market incentives may influence adoption, but they do not alter what OVERT itself proves.

Cautionary lessons: Tiered systems can incentivize point gaming or over-interpretation of lower-tier certifications. For AI governance, this informed OVERT's emphasis on distinguishing documentation, operator-controlled telemetry, and independently verifiable attestation rather than treating all conformance levels as equivalent.

C.7 PCI-DSS Contractual Adoption Precedent

Design Decision: OVERT includes a signal architecture and independent-verification model that can be incorporated into contractual and oversight processes, paralleling the way PCI-DSS was operationalized through private agreements.

Rationale: Standards adoption often accelerates when contractual incentives align with verification requirements.

Supporting Analysis: PCI-DSS illustrates that private agreements can embed verification expectations into commercial relationships without waiting for legislation. The relevant point for OVERT is that procurement, platform, insurance, or sector-specific contracts may reference verifiable governance evidence and independent verification artifacts. OVERT is designed to be referenceable in those settings, but it does not prescribe a particular market structure, assessment industry, or contractual model.

C.8 FedRAMP and NIST SP 800-53 Adoption History

Design Decision: OVERT provides crosswalks to NIST SP 800-53 Rev 5 and FedRAMP (in the companion document `OVERT_v1.1_CROSSWALKS.md`) and supports OSCAL-formatted attestation packs.

Rationale: Federal adoption of security standards follows established patterns through NIST framework alignment, FedRAMP authorization, and OSCAL-based automation.

Supporting Analysis: Federal security standards often spread through crosswalks, machine-readable artifacts, and reuse in existing compliance workflows. The relevant point for OVERT is interoperability: by mapping to NIST/FedRAMP concepts and supporting OSCAL-compatible out-

puts, implementers can present attestation evidence in familiar oversight formats. These references explain integration paths, not adoption forecasts or official endorsement.

C.9 Insurance Market Interpretation

Design Decision: OVERT Section 4.6 (Risk Signal Architecture) and Annex D (Risk Signal Framework) are primary design pillars, not afterthoughts.

Rationale: Insurance market reactions illustrate that external risk bearers may seek more verifiable runtime evidence for AI systems. Those reactions are informative context; they do not determine the scope or legal effect of this standard.

Supporting Analysis: Insurance markets have begun issuing both exclusions and affirmative products addressing AI risk. Those developments are relevant here only as evidence that some external risk bearers are beginning to differentiate among AI governance postures; they do not imply insurer endorsement of OVERT, any required coverage position, or any specific underwriting outcome.

These developments show why independently verifiable runtime evidence may become relevant to external risk assessment. OVERT provides a signal and evidence architecture that can be evaluated in that context, but it does not determine coverage availability, pricing, or legal entitlement.

C.10 Non-Egress Architecture and Business Associate Agreement Exposure

> ***Informative —*** *This section describes architectural properties relevant to regulatory analysis. It does not constitute legal advice. Organizations should obtain qualified legal counsel regarding data processing agreements and BAA requirements for their specific deployments.*

Design Decision: Section 17.5 states that the non-egress architecture "SHOULD be designed to prevent the transmission of Protected Health Information (PHI) or other regulated content" while explicitly hedging: "The applicability of data processing agreements or Business Associate Agreements remains a question of applicable law and regulatory interpretation."

Rationale: The architectural argument for reduced BAA exposure is strong but the legal conclusion is not yet settled. OVERT preserves the argument without overclaiming.

Supporting Analysis: Under HIPAA, a Business Associate is any person or entity that "creates, receives, maintains, or transmits" PHI on behalf of a covered entity (45 CFR §160.103). The OVERT non-egress architecture is specifically designed so that the attestation layer — including the receipt service, notary network, and transparency log — never receives PHI. Only cryptographic commitments (HMAC-SHA256 digests with tenant-scoped keys) cross the operator's trust boundary. The raw content remains in the operator's content-addressable storage, never leaving the covered entity's environment.

The architectural claim is: if the attestation provider never receives, creates, maintains, or transmits PHI — receiving only irreversible cryptographic commitments from which PHI cannot be reconstructed — the attestation provider may not meet the statutory definition of a Business Associate. This may reduce the factual basis for treating the attestation provider as a recipient of PHI, but BAA obligations remain a matter of applicable law and deployment-specific facts.

However, OCR has not issued guidance specifically addressing whether receipt of cryptographic commitments derived from PHI constitutes "receiving" PHI. The closest analogue is the de-identification safe harbor (45 CFR §164.514(b)), which permits disclosure of health information from which specified identifiers have been removed. HMAC commitments are arguably stronger than de-identification: they are computationally irreversible without the tenant-scoped key, which the attestation provider never possesses.

The hedge in Section 17.5 reflects the current state: the architectural argument is sound, the legal conclusion requires either OCR guidance or judicial interpretation, and the standard should not assert a legal conclusion that applicable law has not yet confirmed. Healthcare operators should consult qualified HIPAA counsel regarding their specific deployment architecture.

C.11 Emergent Behavior in Authorized Agentic Systems

Design Decision: OVERT Section 16 (Behavioral Drift Governance) introduces five control families (DRIFT-1 through DRIFT-5) addressing emergent behavioral changes in agentic AI systems that occur entirely within authorized operational bounds.

Rationale: Existing governance frameworks — including earlier per-action runtime control models — are designed to detect and prevent policy violations: individual actions that breach a defined rule. Agentic AI systems introduce a qualitatively different governance challenge: emergent behavior

where every individual control passes but the system's aggregate behavior drifts, cascades, or produces ungovernable complexity. This gap cannot be closed by tightening existing controls; it requires a new category of governance capability.

Supporting Analysis: Per-action attestation — the foundational model used by earlier runtime-governance designs and comparable frameworks — operates on a premise inherited from conventional access control: that governance is the sum of individual authorization decisions. This premise holds for request-response systems where each invocation is independent. It does not hold for agentic AI systems, where persistent agents accumulate state, spawn subordinate agents, and operate across extended time horizons. Six illustrative scenarios demonstrate the structural inadequacy of per-action governance for agentic deployments.

Spawn chain complexity. An orchestrator agent, operating within its declared capability set, spawns sub-agents to decompose a complex task. Each sub-agent, also operating within its declared capability set, spawns further sub-agents. Every individual spawn decision is authorized under the system's capability policy. The resulting execution graph, however, may comprise dozens or hundreds of leaf agents operating in parallel, each issuing tool calls, consuming resources, and producing outputs that feed into sibling and parent agents. The aggregate topology — the total number of active agents, the depth of the spawn hierarchy, the fan-out at each level — may far exceed what any human operator anticipated or any governance process was designed to oversee. Existing controls such as MULTI-2 attest the topology of agent hierarchies but do not evaluate whether the observed topology complexity exceeds the deployment's declared operational baseline.

Within-bounds behavioral drift. An agent produces outputs that individually conform to all applicable policy constraints across successive operational epochs. No single output is flagged, rejected, or escalated. Over time, however, the statistical distribution of those outputs shifts: risk scores trend higher or lower, topic coverage narrows, tool selection patterns change. The shift may be gradual enough that no individual epoch-to-epoch comparison triggers concern, yet the cumulative drift from the system's initial behavioral baseline is substantial. Measurement and evaluation controls such as MEA-2 assess whether individual outputs violate policy thresholds; they do not detect distributional shifts in the population of authorized outputs. Such drift may indicate model degradation, subtle prompt manipulation that biases rather than violates, or environmental changes that alter the agent's effective decision-making.

Cascading depth exploitation. Consider a three-level agent hierarchy where each level spawns three sub-agents. The resulting execution graph contains twenty-seven leaf agents. Each individual agent operates within its authorized bounds — its tool calls are permitted, its outputs conform to policy, its resource consumption falls within declared limits. But the combinatorial complexity of the full execution graph — the total volume of tool invocations, the interaction patterns between agents at different levels, the aggregate resource consumption, the effective attack surface — may be orders

of magnitude beyond the deployment's design assumptions. Existing recursion depth limits operate per-trace and do not evaluate the aggregate complexity of concurrent execution graphs sharing a common orchestrator.

Tool selection drift. An agent authorized to invoke multiple tools shifts its selection distribution over time. Where the agent previously selected one tool for approximately sixty percent of invocations and another for approximately forty percent, the ratio gradually inverts. Neither tool is prohibited; every individual invocation is authorized. The change in selection distribution, however, may indicate that the agent's underlying decision-making behavior has materially changed. Existing controls log individual tool invocations but do not track selection distributions across tools over time, and therefore cannot detect distributional shifts that leave every individual action compliant.

Propagated drift across agent hierarchies. When a parent agent drifts in the manner described above, its outputs — which serve as inputs to downstream agents — change in distribution. Child agents, whose models, policies, and configurations remain unchanged, alter their behavior in response to the changed input distribution. The behavioral drift propagates through the attestation DAG without any agent individually violating its policy. Existing controls evaluate each agent's behavior independently and do not correlate behavioral changes across parent-child attestation linkages, rendering propagated drift invisible to per-agent governance.

Human oversight quality degradation. Human reviewers responsible for overseeing AI outputs initially conduct substantive reviews: they spend adequate time, apply corrections at rates consistent with the system's risk signals, and demonstrate decision patterns that correlate with output characteristics. Over time — through automation bias, workload pressure, or miscalibrated trust — review duration decreases, modification rates decline, and the statistical correlation between risk signals and review decisions weakens. The review process continues to occur, and existing controls attest that it occurred, but the review ceases to be substantively meaningful. Approval velocity controls may cap the rate of approvals but do not assess whether the cognitive engagement underlying each approval is sufficient for the decision's risk level.

These six scenarios share a common structural feature: the governance gap lies not in any individual action but in the relationship between per-action compliance and system-level behavior. An attestation system that evaluates each action independently and finds no violation may nonetheless fail to detect that the system's aggregate behavior has materially changed — potentially in ways that alter its risk profile, undermine its fitness for purpose, or erode the effectiveness of human oversight. DRIFT-1 through DRIFT-5 close this gap by requiring that conformant systems declare their intended behavioral baseline (DRIFT-1), detect deviations from that baseline using sequential statistical methods (DRIFT-2), evaluate execution topology complexity against declared bounds (DRIFT-3), trace behavioral drift propagation across agent hierarchies (DRIFT-4), and assess the substantive quality of human oversight processes (DRIFT-5). The standard specifies what conformant systems

must detect and attest. The specific statistical methods, evaluation instruments, and enforcement mechanisms are specified in the registered Protocol Profile.

Annex D: Risk Signal Framework (Informative)

This annex describes the framework for OVERT risk signals. Signal definitions, mathematical formulas, derivation procedures, and minimum credibility thresholds are specified in the registered Protocol Profile or companion signal specification.

D.1 Signal Properties

All OVERT risk signals satisfy the properties specified in Section 4.6 (where the normative requirements are stated):

1. Content-free derivation
2. Verifiability classification
3. Temporal granularity
4. Statistical rigor
5. Scope binding

D.2 Signal Categories

OVERT risk signals are organized into three categories:

Category	Scope	Examples
Operational Signals	Attestation infrastructure health	Coverage ratios, exposure windows, response latency, retention integrity
Governance Risk Signals	Policy compliance indicators	Violation rate bounds, override frequency, consent coverage, review completion
Agentic Risk Signals	Agentic system behavioral indicators	Behavioral drift rate, graph complexity, spawn authorization, review quality

Agentic Risk Signals apply only to systems classified as "Automation" or "Agentic" under IDE-1.2 and are required for OVERT Agentic conformance.

D.3 Signal Derivation Requirements

Signal specifications in the registered Protocol Profile are required to include, for each signal (see Section 4.6):

- **Signal identifier** — unique, namespaced (e.g., OVERT-INS-NNN for insurance signals, or other prefixes as defined by companion specifications)
- **Definition** — precise natural-language description
- **Formula** — mathematical formula with defined numerator, denominator, and unit
- **Data type and unit** — including encoding requirements to avoid floating-point variance
- **Source artifacts** — which attestation artifacts are required for computation
- **Derivation procedure** — step-by-step auditor-reproducible computation method
- **Aggregation window** — temporal scope (epoch, daily, policy-period)
- **Missing-data handling** — behavior when source artifacts are unavailable
- **Minimum credibility threshold** — minimum sample size for statistically credible interpretation
- **Severity classification** — threshold-based severity levels

D.4 Design Rationale

Risk signals are a primary design goal of OVERT because the verification gap described in the Foreword affects defenders, auditors, regulators, procurement reviewers, and external risk assessors alike. Quantitative risk signals — independently verifiable where the denominator source supports it, operator-dependent where it does not (see Section 4.6) — enable:

- **Security operations:** Monitoring of coverage, overrides, exposure windows, and other runtime indicators within the attested scope
- **Audit and investigation:** Recomputable evidence for control-execution and anomaly analysis
- **Regulatory and oversight reporting:** Quantified posture reporting without content exposure
- **External risk analysis:** Structured inputs for insurance, procurement, or other third-party evaluations, subject to the verifiability classification of the underlying signals

Signal specifications are maintained in the Protocol Profile rather than this standard so that signal definitions can evolve with operational experience and measurement practice, without requiring standard revisions.

Annex E: Legal Admissibility Analysis (Informative)

> ***Informative —*** *This annex does not constitute legal advice. Admissibility determinations are made by courts applying jurisdiction-specific rules. Organizations should consult qualified legal counsel regarding the admissibility of attestation artifacts in their jurisdictions.*

This annex analyzes how AAL-4 attestation artifacts produced by OVERT-conformant systems may relate to evidentiary rules governing the admissibility of electronic records. The discussion identifies how OVERT design features address foundational admissibility concepts — authenticity, integrity, chain of custody, and hearsay exceptions — without asserting that any specific attestation artifact will be admitted in any specific proceeding.

E.1 Federal Rules of Evidence 902(13): Certified Records of Regularly Conducted Activity (Electronic)

Rule: FRE 902(13), effective December 1, 2017, provides for self-authentication of "a record of a regularly conducted activity" in electronic form, when accompanied by a certification from a qualified person that the record: (A) was made at or near the time of the occurrence of the matters set forth by a person with knowledge, or from information transmitted by such a person; (B) was kept in the course of the regularly conducted activity; and (C) was made as a regular practice of that activity.

OVERT Mapping:

FRE 902(13) Requirement	OVERT Feature	Relevant Controls
"made at or near the time of the occurrence"	Attestation receipts include nanosecond-precision timestamps (wall_time_ns), co-epoch binding, and transparency log inclusion with signed tree heads providing independent temporal verification.	ATT-1, ATT-2, Section 18

FRE 902(13) Requirement	OVERT Feature	Relevant Controls
"by a person with knowledge, or from information transmitted by such a person"	OVERT records are machine-generated by the arbiter and notary network. Courts increasingly accept automated systems as sources of business records when the system's reliability is established. The notary network's independent derivation of binary identity and network state provides an additional reliability indicator.	ATT-2.2, ATT-3.3, ATT-5
"kept in the course of regularly conducted activity"	OVERT attestation is continuous and automatic — operating on every in-scope AI action as a regular practice, not created in anticipation of litigation. The transparency log provides a tamper-evident, append-only record.	ATT-4, Section 21.1
"made as a regular practice"	AAL-4 conformance requires continuous attestation for all in-scope actions. The DPL publishes request commitments per epoch as a regular operational practice.	Section 22 (Conformance)
Certification by qualified person	Section 21.3(e) (Custodian Certification) requires the export package to include custodian identity, timestamp, scope declaration, and hash of the export package. This certification can be prepared by the operator's designated custodian of records.	Section 21.3, 21.5

Analysis: OVERT attestation artifacts are designed to address the structural elements of FRE 902(13). Whether a specific court accepts these artifacts under FRE 902(13) will depend on the proponent's compliance with notice requirements (FRE 902(13) requires written notice and opportunity to inspect), the court's assessment of the underlying system's reliability, the proponent's ability to establish the "regular practice" and "person with knowledge" elements for machine-generated records, and the specific facts of the deployment. This analysis identifies design alignment; it does not predict admissibility outcomes.

E.2 Federal Rules of Evidence 902(14): Certified Data Copied from Electronic Device, Storage Medium, or File

Rule: FRE 902(14), also effective December 1, 2017, provides for self-authentication of data "copied from an electronic device, storage medium, or file" when accompanied by a certification from a qualified person that the process of digital identification used to verify the data is trustworthy, typically through cryptographic hash verification.

OVERT Mapping:

FRE 902(14) Requirement	OVERT Feature	Relevant Controls
Data "copied from" electronic device	OVERT immutable export packages (Section 21.3) are copied from the operator's content-addressable storage and the transparency log.	Section 21.3
Process of digital identification	SHA-256 hashing, HMAC commitments, Ed25519/BLS signatures, and Merkle tree inclusion proofs provide multiple layers of cryptographic verification.	ATT-1, Section 18, Annex B
Process "used to verify" is trustworthy	The entire OVERT verification chain is publicly specified, uses NIST-approved cryptographic primitives (SHA-256, HMAC-SHA256, HKDF), and is independently reproducible by any party.	Protocol Profile, Annex B
Certification by qualified person	Section 21.3(e) requires custodian certification with identity, timestamp, scope, and hash of the export package.	Section 21.3(e)

Analysis: FRE 902(14) was specifically designed to accommodate cryptographic hash verification of electronic data. OVERT attestation artifacts employ multiple layers of cryptographic integrity verification (content hashes, commitment chains, Merkle tree proofs, notary signatures) designed to address the requirements of 902(14) authentication. The publicly documented verification procedures enable opposing parties and courts to assess the trustworthiness of the digital identification process.

E.3 Federal Rules of Evidence 803(6): Business Records Exception to Hearsay

Rule: FRE 803(6) excludes from the hearsay rule a record of a regularly conducted activity if: (A) made at or near the time by someone with knowledge; (B) kept in the course of a regularly conducted business activity; (C) making the record was a regular practice; and (D) these conditions are shown by testimony of the custodian or another qualified witness, or by a certification under FRE 902(11), (12), or (13). The record may be excluded if "the source of information or the method or circumstances of preparation indicate a lack of trustworthiness."

OVERT Mapping:

Elements (A), (B), (C), and (D) map identically to the FRE 902(13) analysis above. The additional trustworthiness inquiry under FRE 803(6)(E) is addressed by OVERT's design properties:

Trustworthiness Factor	OVERT Feature
Source reliability	Attestation records are generated by cryptographically verified arbiters (binary identity derived by independent notaries, not self-reported) operating within verified network isolation (NETATT).
Preparation circumstances	Attestation is continuous, automatic, and not created in anticipation of litigation. The system operates identically regardless of whether litigation is pending.
Tamper evidence	Transparency log provides append-only storage with Merkle tree consistency proofs. Any modification is detectable through signed tree head comparison.
Independent verification	Any party can independently verify attestation integrity using publicly available verification procedures without operator cooperation.

Analysis: OVERT attestation artifacts are designed with properties relevant to the FRE 803(6) trustworthiness inquiry. Whether a specific court finds a particular deployment's attestation artifacts trustworthy under 803(6)(E) will depend on the deployment-specific facts, including system reliability, operational consistency, and the circumstances of record creation. This analysis identifies design alignment; it does not predict admissibility or trustworthiness determinations.

E.4 Federal Rules of Civil Procedure 37(e): Failure to Preserve ESI

Rule: FRCP 37(e) addresses the consequences of failing to preserve electronically stored information (ESI) that should have been preserved in anticipation or conduct of litigation. If ESI is lost because a party failed to take reasonable steps to preserve it, and it cannot be restored or replaced through additional discovery, the court may: (1) upon finding prejudice, order measures no greater than necessary to cure the prejudice; or (2) upon finding that the party acted with intent to deprive, presume the information was unfavorable, instruct the jury accordingly, or dismiss the action.

OVERT Mitigation:

OVERT Section 21 (Legal Preservation and Production) directly addresses FRCP 37(e) exposure:

FRCP 37(e) Element	OVERT Mitigation	Relevant Section
"reasonable steps to preserve"	Section 21.1 requires operators to define and publish retention policies meeting or exceeding the longer of regulatory requirements or applicable statutes of limitation. Section 21.2 requires legal hold procedures upon receipt of litigation hold notice or preservation demand.	Section 21.1, 21.2
"lost because a party failed"	The transparency log provides an independent, tamper-evident record of what attestation artifacts existed. Even if the operator's local copy is lost, the transparency log entries (hashes, inclusion proofs) remain, proving that the artifacts existed and establishing their content hashes.	ATT-4, ATT-4 (Section 8)
"cannot be restored or replaced"	Section 21.3 requires immutable export capabilities. The transparency log + notary signatures provide partial reconstruction capability even if local evidence is lost. Co-epoch binding and receipt hashes enable a court to determine the scope of loss.	Section 21.3, 21.4
"intent to deprive"	OVERT audit trails make intentional destruction detectable. If an operator deletes local evidence, the transparency log still contains the receipts — showing what was attested and when. The gap between transparency log entries and available local evidence is itself evidence of deletion.	ATT-4, ATT-4 (Section 8)

Analysis: OVERT does not eliminate FRCP 37(e) exposure — no technical system can prevent a party from destroying evidence if they are willing to accept the consequences. However, OVERT creates a structural environment where: (a) preservation obligations are documented in the operator's published retention policy; (b) legal hold procedures are defined and attestable; (c) the transparency log provides an independent record of what artifacts existed, making destruction detectable; and (d) the gap between what the log shows existed and what the operator can produce is itself a measurable, verifiable retention-integrity signal.

E.5 International Admissibility References

United Kingdom: Civil Evidence Act 1995

The Civil Evidence Act 1995 abolished the common law rule against hearsay in civil proceedings, making all relevant evidence admissible subject to weight. Section 9 provides that a document shown to form part of the records of a business may be received in evidence without further proof, with a certificate signed by an officer of the business sufficing to establish that the document forms part of those records; Section 8 separately permits a statement in a document to be proved by production of the document or an authenticated copy. OVERT attestation artifacts, accompanied by custodian certification (Section 21.3(e)), are designed to support business-records certification under Section 9 and proof by authenticated copy under Section 8. The weight given to such evidence remains at the court's discretion, informed by factors including the reliability of the computer system and the manner in which the data was processed.

The UK has not enacted standalone AI legislation as of March 2026, maintaining a "pro-innovation" regulatory approach with cross-sector principles applied through existing regulators. The Online Safety Act 2023, with a February 2026 amendment bringing standalone AI chatbots within scope, creates an expanding statutory backdrop. In the absence of AI-specific evidentiary rules, OVERT attestation artifacts would be assessed under general principles of electronic evidence admissibility.

European Union: eIDAS Regulation (Regulation 910/2014 and eIDAS 2.0)

The eIDAS Regulation provides a legal framework for electronic identification and trust services across EU member states. Under eIDAS:

- **Electronic signatures** (Article 25): An electronic signature shall not be denied legal effect solely on the grounds that it is in electronic form. Qualified electronic signatures have the equivalent legal effect of handwritten signatures.

- **Electronic seals** (Article 35): An electronic seal shall not be denied legal effect solely on the grounds that it is in electronic form. Qualified electronic seals enjoy a presumption of integrity of the data and correctness of the origin.
- **Electronic time stamps** (Article 41): An electronic time stamp shall not be denied legal effect solely on the grounds that it is in electronic form. Qualified electronic time stamps enjoy a presumption of accuracy.
- **Electronic documents** (Article 46): An electronic document shall not be denied legal effect solely on the grounds that it is in electronic form.

OVERT attestation artifacts — cryptographically signed, timestamped, and integrity-verified — are designed to satisfy the structural requirements for electronic evidence under eIDAS. Organizations operating under eIDAS may additionally seek qualified trust service provider (QTSP) status for their notary network operations, which would provide the legal presumptions associated with qualified electronic signatures, seals, and time stamps. The eIDAS 2.0 update extends the framework to include electronic ledgers, which may be relevant to OVERT transparency log operations.

For EU AI Act operationalization specifically, the Article 12 logging obligations are expected to be implemented through harmonized standards developed by CEN-CENELEC JTC 21; referencing OVERT within, or contributing it to, those deliverables is the highest-leverage path to regulatory recognition. Separately, W3C Verifiable Credentials is a natural encoding for both the conformance claim (Section 22.4) and the IDENT-1 delegation chain, and is worth tracking as agent-identity standardization (OAuth token exchange, GNAP, and emerging agent-identity efforts) matures. These are informative positioning notes, not conformance requirements.

The revised EU Product Liability Directive (Directive 2024/2853), with transposition deadline December 9, 2026, explicitly treats software and AI systems as products. Article 9 creates rebuttable presumptions of defectiveness where a defendant fails to comply with disclosure obligations. OVERT attestation artifacts are designed to support the operator's ability to respond to disclosure obligations with verifiable, tamper-evident records.

Annex F: Sample Citation Language (Informative)

This annex provides canonical citation forms for referencing OVERT conformance in legal, procurement, insurance, and regulatory contexts.

F.1 Canonical Conformance Citation Format

The standard citation form for an OVERT conformance claim follows the grammar defined in Section 22.4. All claims include a human-readable scope summary and exclusions summary. Level 3 and Level 4 claims additionally include coverage percentage, denominator source and verifiability classification, scope hash, and exposure-window duration.

Example (Level 3):

> ***Example —*** *OVERT Level 3 Agentic — v1.0.0, Protocol Profile 1.0, Scope Summary: sys-agent-010 patient-facing agentic workflows (API gateway gw-prod-01, FHIR interface, voice endpoint), Exclusions: None (full coverage verified), Scope: 85% of inbound API traffic, Denominator: Independent, Scope Statement: sha256:[scope-hash], Exposure Window: 0h (0%), IAP Topology: Multi-IAP, as of 2026-06-15*

Example (Level 2):

> ***Example —*** *OVERT Level 2 Core — v1.0.0, Profile v1.0, Scope Summary: sys-cda-001 clinical documentation API (FHIR R4 interface, HL7v2 ADT feed), Exclusions: Not assessed: batch-analytics-002 (scheduled for Q3 assessment), as of 2026-03-15*

F.2 Guidance for Referencing OVERT in External Documents

Organizations referencing OVERT conformance in procurement, insurance, regulatory, or legal contexts SHOULD:

1. Use the canonical citation format from Section F.1, which includes scope summary, exclusions, denominator source, and exposure-window fields.
2. Not imply that OVERT conformance establishes legal compliance, regulatory approval, insurance coverage, or a judicially recognized standard of care.
3. Not imply that OVERT conformance covers systems, interfaces, or traffic classes outside the declared scope.
4. Consult qualified legal counsel when drafting contract, insurance, or regulatory language that references OVERT.
5. Clearly distinguish between independently verifiable signals and operator-dependent signals when making claims about evidence quality.

> ***NOTE*** — *Previous versions of this annex included sample legal, procurement, insurance, and regulatory paragraphs. Those samples were removed because copy-paste-ready advocacy language in a standard creates misrepresentation risk. Organizations should draft context-specific language with qualified counsel using the canonical citation format and the scope/exclusions/denominator disclosures required by Section 22.4.*

F.3 Disclaimer

> ***NOTE*** — *This annex is informative only and does not constitute legal advice. OVERT has not been judicially recognized as defining a standard of care. No insurer, regulator, or court has adopted OVERT as dispositive evidence. Citation forms should be adapted to the specific legal jurisdiction, regulatory framework, and contractual context in which they are used. Organizations should consult qualified legal counsel when referencing OVERT in any external document.*

Annex G: Supplementary Requirements (Normative — added in v1.1)

> ***Status.*** *This annex is* ***normative****. It defines supplementary requirements added in version 1.1 that operationalize, and do not modify, the Part 1–22 normative core. Each subsection states the conformance level and scope at which its requirements apply. Where a subsection states that its requirements apply at AAL-4 for Level 4 Agentic-Extended claims, conformance to that subsection is required for the corresponding claim in addition to the conformance matrix of Section 22.5. Section G.4 is an informative reference that documents an artifact already mandated normatively by Section 10; it introduces no new obligation.*

G.1 Local CAS Evidence Retrieval and Retention Integrity

> ***Status.*** *This section defines normative requirements (SHALL/SHALL NOT/MAY per RFC 2119) added in version 1.1. It operationalizes, and does not replace, the content-verification provisions of Section 20.3(c) and Section 20.4, the immutable-export provisions of Section 21.3, and the retention-integrity Operational Signal defined in Annex D, Section D.2. All endpoint paths, wire encodings, and message schemas referenced below are specified in the registered Protocol Profile; this section specifies required behavior, not transport.*

G.1.1 Purpose and Scope

OVERT's non-egress architecture (Section 17) requires that only cryptographic commitments and profile-defined metadata cross the operator's trust boundary during routine attestation. Section 20.3(c) and Section 20.4 establish that content verification — auditor access to the protected payloads underlying attested interactions — is the exception, permitted only under legal authority or contractual agreement and accompanied by cryptographic proof that the accessed artifacts

are genuine and contemporaneous. This subsection defines the normative interface and integrity-assurance requirements by which a relying party exercises that exception against the operator's Local Content-Addressable Storage (CAS, Annex A.7), and by which an external risk bearer obtains continuous assurance that required evidence is being retained.

The requirements in this subsection are normative at AAL-4 for Level 4 Agentic-Extended conformance claims. They are OPTIONAL for Level 1 through Level 3 claims and for Core and Agentic scope. Implementations that do not expose a content-verification interface are not required to implement this subsection; an operator that omits it SHALL declare the omission in the conformance statement Exclusions field with architectural justification (Section 22.4).

G.1.2 Evidence Retrieval Interface

An operator that exposes a content-verification interface SHALL provide a deterministic evidence-retrieval operation keyed on the `evidence_commitment` field of the Extended Envelope (Annex B.6). The operation accepts one or more `evidence_commitment` values identifying targeted receipts, a declared governance reason for the retrieval, and the verifiable identity of the requesting principal; it returns, for each located commitment, the original canonicalized payload (the prompt, response, and policy-evaluation scores as committed), together with an enumeration of commitments that could not be located.

The interface SHALL satisfy the following requirements:

(a) **Authorization.** The interface SHALL require out-of-band, mutually authenticated authorization. Retrieval requests SHALL NOT be served on the routine receipt-service egress path (Section 17.4); content verification is the exception path of Section 20.4, not part of routine attestation.

(b) **Governance-reason declaration.** Each request SHALL declare a governance reason drawn from a closed enumeration defined in the registered Protocol Profile (for example: incident response, regulatory audit, insurance claim, legal discovery). The declared reason SHALL be attested and SHALL be retained as part of the operator's audit record.

(c) **Verifiability of retrieved payloads.** The requesting party SHALL be able to independently recompute the `evidence_commitment` over each returned canonicalized payload, using the canonicalization method (Section 17.1) and the commitment construction specified in the registered Protocol Profile, and to confirm that the recomputed value equals the commitment recorded in the corresponding receipt. Because the `evidence_commitment` construction is keyed (Protocol Profile 1.0 derives it via HMAC under operator-held keys within the split-knowledge key hierarchy; see Annex A.28 and Annex B), the operator SHALL provide the requesting party a profile-defined **commitment-opening mechanism** sufficient to recompute and verify the commitment under the request's authorization — for example, disclosure of a one-time, commitment-scoped opening key; a

key-management-service (KMS) verification oracle; or verification within an attested verifier enclave — without disclosing long-term content-binding keys. A payload whose recomputed commitment does not match its recorded `evidence_commitment` SHALL be treated as a content-integrity failure and reported as such.

(d) **Non-egress preservation.** Retrieval under this subsection occurs within the operator's environment under the requesting party's lawful authority; it does not relax the non-egress property of Section 17. Protected content disclosed under this interface is disclosed to an authorized verifier under Section 20.4, not egressed to the attestation layer.

(e) **Closed schema.** The request and response message schemas SHALL be closed (unknown fields rejected; `additionalProperties: false`), consistent with the closed-schema requirement of Section 17.4. The concrete field set, encodings, the endpoint path, and the commitment construction are specified in the registered Protocol Profile.

> ***Note — relationship to immutable export.*** *The evidence-retrieval interface is a targeted, commitment-keyed lookup. It does not replace the immutable export package of Section 21.3, which remains the mechanism for producing a verifiable, signed, transparency-log-consistent corpus for litigation or regulatory examination. Production of operator-local artifacts under either mechanism may require operator cooperation or lawful process (Section 21.3).*

G.1.3 Proof of Possession

An external risk bearer or auditor may require continuous assurance that the operator is in fact retaining the evidence payloads required for future investigation, subrogation, or audit, without compelling the operator to egress those payloads. An operator that makes such an assurance SHALL implement a Proof of Possession (PoP) challenge–response operation with the following properties:

(a) The requesting party generates a fresh challenge value of at least 256 bits from a cryptographically secure random source and submits it together with a subset of `evidence_commitment` values to be challenged.

(b) For each challenged commitment, the operator's CAS retrieves the corresponding canonicalized payload and computes a possession value that cryptographically binds the challenge value to the canonical payload, using the construction specified in the registered Protocol Profile (a keyed or domain-separated hash over the challenge value and the canonical payload). The operator returns the possession values and a count of challenged commitments whose payloads could not be retrieved.

(c) The requesting party SHALL retain the challenge value. If the audit is later escalated under Section 20.4, the requesting party may obtain the full payload under that authority, recompute the

possession value with the retained challenge value, and thereby verify that the operator held the unmodified payload at the time of the original challenge.

(d) The PoP operation SHALL be subject to the same out-of-band, mutually authenticated authorization required by Section G.1.2(a). The challenge construction, possession construction, encodings, and endpoint path are specified in the registered Protocol Profile.

> ***Note.*** *Proof of Possession provides retention assurance, not contemporaneity. A successful PoP demonstrates that the operator can produce a payload matching the recorded* `evidence_commitment` *at the time of challenge; contemporaneity of the original attestation continues to be established by the receipt's co-epoch binding (ATT-2, Section 8) and transparency-log inclusion (Section 20.1), not by PoP.*

> ***Note — operator protection.*** *A Proof of Possession challenge requires the operator to read and hash full payloads; a bulk challenge over very large or numerous payloads (for example, multimodal contexts spanning millions of tokens or large media objects) can impose significant disk-I/O and CPU load amounting to an authorized denial of service. PoP endpoints SHOULD support asynchronous processing, rate-limiting, batch-size limits, and/or a profile-defined sampling discipline (for example, Merkle-tree-based spot-checks over a payload's chunks) to bound operator cost while preserving the integrity guarantee.*

G.1.4 Retention Integrity Signal

The count of challenged commitments whose payloads could not be retrieved during a Proof of Possession operation (Section G.1.3(b)), together with the count of commitments reported as not located during evidence retrieval (Section G.1.2), SHALL feed the **retention integrity** Operational Signal enumerated in Annex D, Section D.2. A non-zero retention-failure count indicates that evidence required to be retained under the operator's retention schedule (Section 21.1) is not retrievable, which constitutes a gap in attestation continuity.

The retention integrity signal SHALL satisfy the risk-signal properties of Section 4.6 and the derivation requirements of Annex D, Section D.3; its identifier, formula (numerator, denominator, unit), source artifacts, aggregation window, missing-data handling, minimum credibility threshold, and severity classification are specified in the registered Protocol Profile or companion signal specification. Retention-integrity failures SHALL be reportable to relying parties consistent with the operator's anomaly-triage obligations (Section 4.7) and SHALL NOT be silently suppressed.

> ***Informative.*** *Downstream relying parties — including parametric risk bearers — may treat a non-zero retention-integrity signal as a governance condition with contractual consequences. Any such consequence is a matter of the relevant contract or policy and is outside the scope of this standard; this standard defines only the signal and its verifiable derivation.*

G.2 HTTP Transport Binding for Cross-Boundary Attestation

> ***Status.*** *This section is normative. It specifies the HTTP wire encoding for the cross-boundary attestation protocol defined normatively in Section 4.8, for HTTP/1.1 and HTTP/2 transports. The* `OVERT-Parent-Attestation-Id` *header is the wire encoding of the existing Section 4.8.2* `parent_attestation_id` *reference and adds no obligation beyond Section 4.8. The* `OVERT-Trace-Id` *correlation header (Section G.2.2) is a new requirement of this binding. Transports other than HTTP are bound by their own Protocol Profile specifications.*

G.2.1 Scope

This binding applies when a conformant arbiter (Section 3.2) participates in a cross-boundary workflow (Section 4.8) and the boundary is crossed over an HTTP/1.1 or HTTP/2 transport. It defines canonical request headers that carry the Section 4.8.2 `parent_attestation_id` reference and a correlation identifier across heterogeneous runtimes, gateways, and service meshes without modification of application code. Transports other than HTTP are specified by their own Protocol Profile bindings. This binding does not alter the Section 4.8.8 conformance condition: cross-boundary controls are normative at AAL-4 for Level 3 and Level 4 claims involving cross-boundary workflows, and are not required for workflows that do not cross trust boundaries.

G.2.2 OVERT Context Headers

An arbiter operating on the upstream (egress) path of a cross-boundary HTTP request SHALL inject the following headers. An arbiter operating on the downstream (ingress) path SHALL extract and validate them per Section G.2.4.

Header	Value	Definition
`OVERT-Trace-Id`	32- or 64-character lowercase hexadecimal string (a 128-bit W3C	Correlation identifier for the multi-boundary execution graph. It is a

Header	Value	Definition
	Trace Context identifier, or a 256-bit identifier)	transport-layer correlation aid only; it is NOT an attestation artifact and SHALL NOT be relied upon as evidence of enforcement.
`OVERT-Parent-Attestation-Id`	64-character lowercase hexadecimal string	The SHA-256 hash of the upstream receipt's `attestation_id` as published in the upstream operator's transparency log, exactly as defined in Section 4.8.2. This is the wire encoding of the `parent_attestation_id` field.

`OVERT-Parent-Attestation-Id` SHALL match `^[a-f0-9]{64}$` (a SHA-256 digest). `OVERT-Trace-Id` SHALL match `^[a-f0-9]{32}([a-f0-9]{32})?$`, accepting either a 32-character (128-bit) W3C Trace Context `trace-id` for interoperability with OpenTelemetry, or a 64-character (256-bit) identifier. Header names are case-insensitive per the HTTP specification; the forms above are the canonical spellings and follow RFC 6648 (no `X-` prefix for newly defined fields).

> ***Note.*** *Earlier drafts of this binding proposed a third header declaring an "Identity Assurance Level (IAL)". OVERT does not define IAL; the standard's assurance taxonomy is the Attestation Assurance Level (AAL-1 through AAL-4, Section 4.1), which classifies attestation artifacts, not transport-asserted identity. No identity-assurance header is defined by this binding. Where an arbiter needs to convey the assurance level of an upstream receipt, that level is established by verifying the referenced receipt itself (Section 4.8.6), not by trusting a transport header.*

G.2.3 Egress Injection (Upstream Path)

For each outbound tool invocation, model API call, or RPC that crosses a trust boundary through a conformant arbiter:

(a) **Trace correlation.** If the incoming request already carries an `OVERT-Trace-Id`, the arbiter SHALL preserve and propagate that value unchanged. If no `OVERT-Trace-Id` is present, the arbiter SHALL generate a new identifier using a cryptographically secure random source.

(b) **Parent binding.** The arbiter SHALL set `OVERT-Parent-Attestation-Id` to the SHA-256 hash of the `attestation_id` of its own provisional receipt for the originating action, computed as specified in Section 4.8.2. The provisional receipt is the Phase 2 artifact defined in Section 8 (control

ATT-3.2). Because Phase 3 notary co-signature is asynchronous (Section 8), the injected reference SHALL be derived from the Phase 2 provisional `attestation_id`; if and when the upstream receipt's transparency-log status changes, the downstream `parent_reference_status` is resolved per Section 4.8.6 and Section 4.8.7, not by re-emitting the header.

G.2.4 Ingress Extraction and Validation (Downstream Path)

When a conformant arbiter intercepts an incoming HTTP request at a trust boundary:

(a) **Extraction.** The arbiter SHALL extract `OVERT-Trace-Id` and `OVERT-Parent-Attestation-Id` where present.

(b) **Validation and mapping.** The arbiter SHALL validate each present value against its pattern (`OVERT-Trace-Id` against `^[a-f0-9]{32}([a-f0-9]{32})?$`; `OVERT-Parent-Attestation-Id` against `^[a-f0-9]{64}$`). If `OVERT-Parent-Attestation-Id` is present and structurally valid, the arbiter SHALL record it as the `parent_attestation_id` of the receipt it generates for the downstream action, per Section 4.8.2.

(c) **Status assignment.** The arbiter SHALL set the downstream receipt's `parent_reference_status` (Section 4.8.7) according to the outcome of header extraction and the subsequent upstream-receipt verification defined in Section 4.8.6:

- If `OVERT-Parent-Attestation-Id` is absent, `parent_reference_status` SHALL be `UNAVAILABLE`.
- If the header is present but fails structural validation, or the referenced upstream receipt fails the Section 4.8.6 verification checks, `parent_reference_status` SHALL be `INVALID`.
- If the upstream transparency log does not respond within the profile-defined timeout, `parent_reference_status` SHALL be `TIMEOUT`.
- Otherwise `parent_reference_status` SHALL be `VALID`.

(d) **Continuation.** When `parent_reference_status` is any value other than `VALID`, the downstream receipt SHALL still be generated; attestation of the downstream boundary's own controls proceeds regardless of upstream availability (Section 4.8.7). The transaction MAY proceed where permitted by local policy. Relying parties SHALL treat a non-`VALID` link as a gap in cross-boundary verification, not as a failure of the downstream boundary's own attestation (Section 4.8.7).

G.2.5 Header Survival and Reporting

A gateway, sidecar, or proxy claiming conformance to this binding SHALL propagate the OVERT context headers without alteration to downstream hops it forwards. Where a downstream service or intermediary strips or fails to propagate the headers, the resulting incomplete link SHALL be recorded with `parent_reference_status = UNAVAILABLE` at the next conformant ingress point (Section

G.2.4(c)) and SHALL be reported through the operator's gap-accounting reporting (ATT-3.4) and as an exposure window in risk-signal reporting where the missing link corresponds to an unattested interval (Annex D; signal definitions in the registered Protocol Profile). This binding adds no new conformance level and no new control identifier; conformance with it is conformance with Section 4.8 over an HTTP transport.

G.3 Automated Auditor Discovery and Well-Known Endpoint Protocol

> ***Status.*** *This section is normative. The endpoint paths, JSON wire schemas, and conformance test vectors referenced below are specified in the registered Protocol Profile; this annex defines only the discovery obligation and its binding to existing auditability requirements (Section 20) and measurement requirements (Section 9). The requirements of this section apply at Level 4.*

G.3.1 Purpose and Scope

Section 20 (Third-Party Auditability) requires that the attestation system enable third-party verification of governance claims without requiring trust in the operator. Section 20.5 delegates the auditor verification procedures to the registered Protocol Profile. This annex specifies a discovery mechanism that allows a relying party to locate those procedures and the associated verification artifacts automatically, without bespoke per-operator integration.

Independent Attestation Providers (IAPs) qualified under Section 22.7, and operators hosting their own notary infrastructure, that claim Level 4 conformance SHALL publish a machine-readable discovery document under the `/.well-known/` URI namespace defined by RFC 8615. The discovery document SHALL enumerate the endpoints from which an auditor can retrieve the cryptographic artifacts required by Sections 20.1, 20.3(a), and 20.3(b).

G.3.2 The OVERT Discovery Document

A conformant publisher SHALL expose an OVERT discovery document at the well-known location specified in the registered Protocol Profile (Protocol Profile 1.0 registers the path `/.well-known/overt-configuration`). The document SHALL be retrievable by an unauthenticated HTTP GET and SHALL be served over TLS.

The discovery document SHALL be a closed-schema JSON object. The Protocol Profile specifies the complete field-by-field schema, types, and constraints; at minimum it SHALL include:

- the authoritative issuer URI of the IAP or operator;
- the set of registered OVERT Protocol Profile identifiers the publisher supports;
- the base URI of the RFC 6962 transparency log required by Section 20.1;
- the URI template for per-epoch artifact retrieval (see Section G.3.3);
- the URI of the key set containing the notary network's public verification keys.

The discovery document MAY additionally reference a content-retrieval endpoint for verifiable records held in the operator's local storage (Section 20.3(c) content verification). Because content verification is the exception and operates only under legal authority or contractual agreement (Section 20.4), any such endpoint SHALL require out-of-band, mutually authenticated authorization and SHALL NOT be exposed for unauthenticated routine access.

> ***Note.*** *The optional content-retrieval endpoint referenced above corresponds to the Local CAS evidence-retrieval interface of Section G.1. Where an operator does not expose that interface, the corresponding discovery field is omitted or marked reserved.*

G.3.3 Per-Epoch Artifact Retrieval

Section 9 requires post-epoch publication of the Digest Publication Ledger (DPL) (MEA-1.4) and the S3P epoch nonce (MEA-2.5) so that auditors can reconstruct all sampling decisions for a closed epoch. The per-epoch retrieval endpoint advertised in the discovery document SHALL return, for any closed epoch, the artifacts required for that reconstruction.

The per-epoch response SHALL be a closed-schema JSON object specified in the registered Protocol Profile. For each epoch it SHALL convey:

- the epoch identifier;
- the epoch status, drawn from a closed enumeration that distinguishes active epochs (for which the nonce has not yet been revealed) from closed epochs;
- the S3P epoch nonce, which SHALL be present for a closed epoch and SHALL be omitted while the epoch is active (consistent with the withhold-during-epoch requirement of MEA-2.1);
- the Digest Publication Ledger for the epoch, including its Merkle root and the complete, deterministically ordered set of request commitments processed during the epoch (per MEA-1.4);
- the notary signature over the epoch identifier, status, revealed nonce, and DPL root.

For an active (not yet closed) epoch, the publisher SHALL withhold the epoch nonce and the DPL contents that would permit premature reconstruction, consistent with MEA-2.1 and MEA-2.5. Reveal of these artifacts before epoch close is a conformance failure.

G.3.4 Auditor Verification Flow

The discovery document enables, but does not replace, the auditor verification procedures specified in Section 20 and the registered Protocol Profile. A relying party performing independent verification proceeds as follows:

1. Retrieve the OVERT discovery document from the publisher's declared attestation domain.
2. Retrieve the notary network's public verification keys from the key-set URI advertised in the discovery document.
3. Read receipts from the transparency log (Section 20.1) to obtain the epoch identifiers in scope.
4. Retrieve the revealed nonces and DPLs for the closed epochs from the per-epoch retrieval endpoint (Section G.3.3).
5. Recompute the sampling tags and S3P sampling boundaries using the construction specified in the registered Protocol Profile (MEA-1.2, MEA-2.2) and independently verify the coverage ratio (Section 4.6, item 1) and the statistical safety signals (Section 4.6, item 2; MEA-2.4).

Where a publisher advertises endpoints but a relying party finds a closed epoch for which the required artifacts (nonce, DPL, notary signature) are unavailable or fail verification, the relying party SHALL treat the affected epoch as an attestation gap event for the purposes of gap accounting (Section 4.6, item 3; ATT-3.4).

G.3.5 Conformance

A Level 4 publisher (IAP or self-hosting operator) SHALL expose the discovery document and the per-epoch retrieval endpoint as defined in this annex and the registered Protocol Profile. The discovery document SHALL accurately reflect the publisher's live endpoints; a discovery document that advertises endpoints that do not serve the required artifacts is non-conformant. Discovery is a Level 4 obligation; Levels 1 through 3 MAY publish a discovery document but are not required to.

G.4 ControlAction Reference Schema (Informative)

> ***Status: Informative.*** *This section documents the wire-level structure of the* `ControlAction` *artifact and its validation procedure. It introduces no new conformance obligations. The normative requirement to emit, gate, and bound* `ControlAction` *artifacts is established by Section 10, controls RES-1.2, RES-1.3, and RES-1.4; the cryptographic primitives are those already registered in Annex B.1 (Ed25519 for controller signatures, per RFC 8032). The authoritative, field-level specification is the Protocol Profile document itself.*

`ControlAction` is the attestation artifact emitted by the bounded control loop when verified violation metrics exceed a policy threshold (RES-1.2). Its glossary definition is A.11. The arbiter evaluates every `ControlAction` through the five cryptographic gates of RES-1.3 before applying the requested parameters, and enforces the parameter bounds of RES-1.4 (both obligations are normative in Section 10); this section does not restate those obligations, it only specifies the structure over which they operate. In this structure the requested governance response — the "action type" and "scope" of glossary A.11 — is expressed as the delta between `params_before` and `params_after`, and the temporal binding (the A.11 "timestamp" and co-epoch binding) is carried by `epoch` and the co-epoch receipt referenced by `proof_ref`.

G.4.1 Closed Schema

Protocol Profile 1.0 represents the `ControlAction` artifact as a closed-schema structure. No additional fields are permitted. The parameter set (`sampling_prob` , `queue_max` , `rate_limit`) is the set enumerated in RES-1.2; profiles defining additional bounded parameters extend the parameter objects and the corresponding RES-1.4 bounds in the registered Protocol Profile, not in the core standard.

```
{
  "type": "object",
  "properties": {
    "epoch": {
      "type": "integer",
      "description": "Epoch identifier in which the action is applied. Used by the
Gate 2 (epoch currency) check of RES-1.3."
    },
    "binary_hash": {
      "type": "string",
      "pattern": "^[a-f0-9]{64,128}$",
```

```
      "description": "Digest of the controller binary issuing the action (SHA-256 by
default; a larger digest, e.g. SHA-384/512, where the registered Protocol Profile
specifies one)."
    },
    "params_before": {
      "type": "object",
      "properties": {
        "sampling_prob": { "type": "string", "description": "Decimal string or scaled
integer per Annex B.3; IEEE-754 float not permitted in attested structures." },
        "queue_max": { "type": "integer" },
        "rate_limit": { "type": "integer" }
      },
      "description": "System parameters in effect prior to the action.",
      "additionalProperties": false
    },
    "params_after": {
      "type": "object",
      "properties": {
        "sampling_prob": { "type": "string", "description": "Decimal string or scaled
integer per Annex B.3; IEEE-754 float not permitted in attested structures." },
        "queue_max": { "type": "integer" },
        "rate_limit": { "type": "integer" }
      },
      "description": "Requested new parameters. Subject to the RES-1.4 bounds
check.",
      "additionalProperties": false
    },
    "proof_ref": {
      "type": "string",
      "pattern": "^[a-f0-9]{64,128}$",
      "description": "Digest of the epoch metrics bundle (RES-1.1) that triggered the
action (SHA-256 by default; a larger digest where the registered Protocol Profile
specifies one). Resolved by the Gate 4 co-epoch receipt check."
    },
    "signature": {
      "type": "string",
      "description": "Ed25519 signature (Annex B.1, RFC 8032) by the authorized
controller key over the canonical encoding of epoch, binary_hash, params_before,
params_after, and proof_ref."
    }
  },
  "required": ["epoch", "binary_hash", "params_before", "params_after", "proof_ref",
"signature"],
  "additionalProperties": false
}
```

Numeric and digest fields follow the registered Protocol Profile. Per the deterministic-encoding constraints of Annex B.3, IEEE-754 floats are not used in attested structures; `sampling_prob` is therefore represented as a scaled integer or decimal string, as reflected in the schema above. Digest fields (`binary_hash`, `proof_ref`) are SHA-256 (64 hexadecimal characters) by default; a profile

specifying a larger digest (e.g., SHA-384/512 for alignment with a higher-security signature suite) widens them within the `^[a-f0-9]{64,128}$` constraint. The signature scope and canonical byte layout are specified in the Protocol Profile.

G.4.2 Five-Gate Validation (Reference)

The five gates below restate, for reference, the validation sequence already required by RES-1.3 (gates 1–5), whose gate 3 enforces the parameter bounds defined in RES-1.4. They are reproduced here only to map each gate onto the schema fields above; the normative obligation is in Section 10. If any gate fails, the arbiter rejects the action and retains `params_before`.

Gate	Check	RES-1 basis	Field(s)
1	Verify `signature` against the authorized controller public key.	RES-1.3(1)	`signature`, `binary_hash`
2	`epoch` matches the current active epoch; stale actions are rejected.	RES-1.3(2)	`epoch`
3	All `params_after` values fall within the statically defined bounds of RES-1.4 (e.g., `p_min ⩽ sampling_prob ⩽ p_max`), independent of signature validity.	RES-1.3(3), RES-1.4	`params_after`
4	A valid co-epoch receipt exists for the metrics bundle named by `proof_ref`.	RES-1.3(4)	`proof_ref`
5	A valid co-epoch Network Attestation (NETATT, A.19) exists for `epoch`.	RES-1.3(5)	`epoch`

The arbiter is the enforcement component that performs gates 1–5 (Section 3 / A.2).

G.4.3 Test Vectors

Section 22.6.2 requires that a registered Protocol Profile include published test vectors for every cryptographic operation. Protocol Profile 1.0 provides reference `ControlAction` instances together with their canonical encodings and Ed25519 signatures, and a worked five-gate evaluation (one passing and one per-gate failing case), in the Protocol Profile document (cf. Annex B.12).