TELOS: Forming the Record of Trust
Cryptographic governance receipts for AI agent actions
TELOS AI Labs Inc.
August 2026
TELOS: Forming the Record of Trust
Cryptographic governance receipts for AI agent actions
An agent’s purpose is not a constraint. It is its constitution.
Disclosures
Generative AI disclosure. This document was developed through collaborative analysis between the author and LLM based research agents operating under the TELOS research team methodology. The core research, architectural decisions, and validation framework originated from the author. Research agents validated, extended, and formalized specific sections.
Conflict of interest disclosure. This research was conducted and funded by TELOS AI Labs Inc., which has a commercial interest in the TELOS governance framework.
Evidence status. The quantitative studies reported here were conducted by TELOS AI Labs. They are self-scored detection studies, not independent certifications or independently judged demonstrations of harm prevention. The operating thresholds and composite weights were tuned on development suites without a held-out calibration split. Results are preliminary and subject to peer review.
Abstract
When an organization deploys an AI agent that performs consequential work, the organization carries the consequences. The agent sends correspondence, files reports, changes records, calls tools, and may initiate transactions. Once agents perform enough work to create material value, no human can inspect every action. The deployer therefore needs evidence: a reviewable account of what an agent was authorized to do, what it did, and whether the two remained aligned.
TELOS implements the Record of Trust Framework, a method for producing cryptographic governance receipts for AI agent actions. Its premise is that accountability cannot be added only after an agent acts. The agent must be formed from a declared behavioral specification, bounded by it, and observed against it action by action on every orchestration path integrated with the governance interface.
A human-approved specification compiles to a fixed external reference in embedding space called the Purpose Anchor. Each submitted action is measured against that reference by an independent embedding model. The reference does not drift with conversational context. A configured policy records the result and determines whether a departure is surfaced or enforced. In TELOS’s current internal observe/open configuration, the layer is passive and non-interfering.
Each scored result passed through the opened governance session signing path can produce an Ed25519 signed receipt carrying covered measurement and decision fields. The public prototype currently demonstrates unsigned integrity hashes and chain linkage, not authorship or trusted time. The opened signed receipt schema does not bind a Manifest version, approval root, or predecessor hash. Public cold verification against a published key, approval-root signing, chain-level signing, and verified retention enforcement remain COMING.
The evidence is bounded. The July 2026 run observed 0 undetected prompts across 2,550 prompts under TELOS’s own scoring at an aggressive operating point with a matched benign false positive rate of 74 percent. In the cited 139-event AgentDojo report, 54 of 139 event-level classifications matched the expected labels (38.8%): the 54 detected attack-event rows represent 27 unique attack texts evaluated at two surfaces, while all 85 benign texts were overflagged in one event each under the deliberately generic safety anchor. These are TELOS self-scored classification outcomes, not evidence that execution was blocked. A governed rerun reproduced seven of nine published rows, recorded one scorer-profile non-reproduction at 82.39 percent, and could not execute one row. These results demonstrate a provisional detection mechanism on tested corpora. They do not establish general harm prevention, cross-model generalization, or production effectiveness.
Two boundaries govern the entire paper. TELOS measures alignment to declared purpose, not the correctness of outputs. The evidence described here does not itself establish legal compliance.
1. Why this is needed
An agent is no longer only a demonstration. It sends the email, files the report, changes the record, and moves information or money. The deployer carries the consequences of each action. Past the point where agents create operational value, a bare output is insufficient evidence because nothing in the output alone proves that the work remained inside the authority the human granted.
Research on long-context and multi-turn performance shows that model behavior can degrade as context accumulates. Information in the middle of a long context can lose influence, and recent information can displace earlier instructions. See Liu et al. on lost-in-the-middle behavior, Laban et al. on multi-turn degradation, and Wu et al. on position bias. These findings support a practical concern: a boundary declared at the start of a session may become less behaviorally salient later.
The deeper problem is structural. A language model computes against representations inside the same context that is changing. A governance reference stored only in that context competes with every later token for influence. A reference intended to remain fixed should therefore live outside the governed model.
Three common approaches address adjacent problems but do not produce the same account:

- **Guardrails** screen for general categories of harm. They measure against the guardrail’s rules, not necessarily against the specific authority this principal gave this agent.
- **Training-time alignment and provider safeguards** provide essential model-level safety floors. They do not by themselves express a session-specific or assignment-specific mandate.
- **Compiled rulebooks** translate policy or regulation into enforcement points. They can show whether a rule was encoded, but they do not necessarily show whether every observed action remained aligned with a human-approved purpose.

Regulatory and quality frameworks increasingly call for ongoing documentation, risk monitoring, traceability, and human oversight. The EU AI Act includes post-market monitoring obligations for high-risk systems. The NIST AI Risk Management Framework calls for risk tracking over time. The NAIC Model Bulletin on the Use of AI Systems by Insurers describes governance and documentation regulators may request. These sources corroborate the direction of travel. They do not certify TELOS or make its records legally sufficient.
The practical standard is simpler: show the work.
2. Constitutive governance
2.1 Human authority and agent authority

A governed agent is a constituent of its principal coordinator's purpose. It is formed from a declared behavioral specification, bounded by that specification, and observed against it action by action.

The human principal holds constituent power. The agent holds constituted power. The human defines the assignment, tools, limits, and escalation policy. The agent acts only as a technical component of that declared body of work. The term constituent does not make an agent a legal person, stakeholder, or bearer of liability. Accountability continues to run to the human and organizational actors responsible for deployment.

This ordering follows the distinction between constituent and constituted power associated with Sieyes and the accountability relationship described by Bovens. It also fits the information asymmetry described in principal-agent theory by Jensen and Meckling. TELOS applies those ideas to technical architecture: the declared specification is the norm, telemetry is the account, the human principal is the forum, and the deployer remains the accountable actor.

Bound in the whether, free in the how. A constituent of a purpose is not a marionette. The specification fixes what the work is for and where it may not go, and leaves the method open. Constitutive governance binds the destination and observes the path. It does not script the path.

Why this differs mechanically from a wall. An agent has no relationship to a barrier placed around it, because the barrier belongs to the environment rather than to the agent's mandate. Where such a barrier does not track the mandate, two failure modes become available: the agent routes around it, or it stops work the mandate actually permitted. Neither registers as a departure from the mandate, because the mandate was never what the barrier was expressed in.

A bound the agent is formed from occupies a different position, because it is the reference the work is measured against. Every observed action has a defined distance from it, including the actions that depart. An exterior blocker can log whatever its operator configures, including the actions it allowed, but it cannot report distance from the mandate, because the mandate is not what it was configured with. An interior bound yields a continuous account of how closely the work held to what was authorized, action by action.

This account extends only to actions submitted through an instrumented path. An action taken outside that path is not measured and does not appear in the record.

2.2 The human is the loop

Traditional designs often place a human checkpoint inside the model's operating loop. TELOS inverts that topology. Human intent is the fixed reference. The human sits on the exception path as the authority to whom departures are surfaced.

The proportional measurement relationship is:

Here, is the measured departure at action or turn , is a configured proportional factor, and is an escalation magnitude. In the current observation layer, is a measurement signal, not an actuating force.

The behavioral specification is not authored by the governed model. Human intent is encoded mathematically, and every action submitted to the scoring path is measured against it. An optional automated micro-judge may be enabled for an uncertain band, but it is off by default and its use should itself be recorded. The model remains the governed subsystem.

2.3 The Manifest and Purpose Anchor

When a principal declares the terms of an assignment, the declaration includes at least:

- Purpose: what the agent is for
- Scope: the domain, audience, systems, and work included
- Boundaries: what the agent must not do or approach without review
- Tools: which systems and functions may be used
- Sequence and continuity expectations where configured
- Escalation: what must surface and who decides

The compiled, versioned, human-approved artifact carrying those terms is the Manifest. The mathematical form against which observation scores is the Purpose Anchor.

Declarations become vectors in through a sentence embedding model. Together they form a fixed reference for the assignment. Each submitted action is embedded and compared with that reference. The current paper discloses the external-reference architecture and core similarity relationship. The production embedding configuration, boundary corpus, centroid construction, compilation pipeline, and deployment-specific calibration remain held under NDA.

2.4 What the mandate covers

TELOS governs alignment to declared purpose. It does not establish that an output is true, safe in every dimension, legally permissible, or wise. It complements fact-checking, toxicity filtering, identity and access controls, tool permissions, domain validation, and human review.

The quality of the result also depends on the quality of the declaration. A vague Manifest produces a vague reference. A Manifest designed to flatter the agent's work can yield procedurally consistent but substantively weak measurements. TELOS does not transfer authorship of the specification from the operator. The record should therefore preserve the declaration next to the result so a reviewer can judge both the reference and the measurement.

3. The measurement architecture

3.1 External fixed-reference measurement

The core fidelity relationship is:

 is the embedding of the submitted response or action at time . is the externally stored embedding of the declared reference. The governed model does not maintain inside its own changing context.

This produces three operational properties:

- The reference changes only when authorized human authority changes it.
- The response can be graduated according to measured distance rather than reduced to a universal allow or deny rule.
- The same submitted action can be measured consistently when the scorer, embedding model, anchor, thresholds, weights, and software versions remain fixed.

That third statement depends on configuration control. A changed embedding backend or changed calibration creates a new measurement configuration and should be versioned accordingly.

3.2 Observation, not a stability theorem

Control theory and attractor dynamics supply useful intuition for reference tracking. A fixed anchor defines a basin and a distance from the reference. Classical stability theorems, however, describe actuated systems that drive state back toward a reference. The current TELOS layer does not do that. It observes discrete actions, measures departure, records the result, and surfaces it according to policy.

The paper therefore does not claim closed-loop stability for the deployed observation layer. Binding occurs when the agent is constituted under the Manifest. Observation measures how well that binding holds.

3.3 Integration boundary

The opened interface is designed as provider-independent middleware. An application places an adapter in its orchestration path and submits text and tool metadata to the governance protocol. No model retraining is required. A provider change does not by itself change the protocol interface, although each deployment still requires integration and empirical validation.

Coverage is limited to instrumented paths. An uninstrumented action is not observed. Zero-effort attachment to an existing agent is not claimed.

3.4 Statistical process control over semantic work

Because every action produces a numeric stream against a fixed reference, the deployment can track central tendency, variation, configured bands, departure rates, and sustained drift. The opened response manager uses a baseline-relative exponentially weighted moving average after baseline collection. It does not derive operative verdict thresholds from each session.

The manufacturing analogy is deliberate:

Quality discipline	Physical process	TELOS semantic process
Monitor	Dimensions, weights, defects	Fidelity scores, drift vectors, flags
Limits	Engineering tolerance or control limits	Configured bands around the Purpose Anchor
Response	Adjust or inspect the process	Record and surface per the Manifest
Evidence	Control charts and quality records	Telemetry and governance receipts

Cosine similarity is bounded and may be non-normal. Classical control-chart assumptions therefore do not transfer automatically. The framework uses the discipline of continuous measurement and documented variation, not an unsupported claim that every semantic-score distribution is Gaussian.

3.5 Agentic dimensions and decisions

An agentic Manifest covers a richer operational surface than a conversation. The current runtime exposes per-dimension scores, composite and effective fidelity, decision values, flags, and optional session aggregates. The published interface describes dimensions for purpose, scope, boundary, tool, and chain continuity.

Goals and prohibitions must not be treated identically. Similarity to a purpose can support evidence of relevance. Similarity to a prohibition can signal boundary proximity. The system therefore treats the two directions differently rather than averaging them as though both were positive goals. The composite formulation is given in the Technical Method. The precise boundary corpus contents, centroid construction, and deployment thresholds remain held under NDA.

The configured verdict vocabulary is execute, clarify, and escalate. execute records work within the Manifest. clarify marks proximity or ambiguity for resolution on the agent path before acting; it is not a human review state. escalate is the human path, routing the matter to the accountable person named in the escalation policy. A recorded verdict is a classification of the action, not proof that clarification occurred or that a human received or resolved anything. Under the current internal gateMode: observe and failPolicy: open configuration, the action proceeds while the result is recorded and may be surfaced.

3.6 Conversational and agentic governance are structurally different problems

The mathematics is the same in both settings: cosine similarity, embedding geometry, basin membership, the two-layer fidelity system. What differs is the measurement surface, and that difference is a structural property of the problem rather than a matter of implementation quality.

	Conversational	Agentic
Measurement space	Continuous semantic embedding space	Discrete operational events
Anchor stability	May legitimately shift per turn	Fixed specification
Governance inputs	Nuanced generalizations	Discrete knowns
Boundary clarity	Scope can drift imperceptibly	Authorized or not
Tractability	Less tractable, requires tighter framing	More tractable, constrained space
The question asked	Are these two signals staying aligned?	Is this operation within the specification?

In conversation the anchor may move for legitimate reasons, because a user’s intent develops; a shift from property analysis toward risk assessment can be the progression of human intent rather than a departure. On an agentic surface the specification is declared in advance and changes only when the human authority amends it, so a tool call is either within it or is not.

This predicts that agentic measurement should be more precise than conversational measurement, and the benchmark results are consistent with that prediction. It is a statement about the tractability of the two measurement problems, not a claim that either result is validated beyond what the Validation Report reports.

4. The Record of Trust

4.1 The evidence problem

When agent work is challenged, a deployer often has a system prompt, application logs, and an after-the-fact reconstruction. A system prompt does not prove what the agent did. A log does not necessarily show the authority governing the event. A reconstruction is weaker than contemporaneous evidence.

The Record of Trust names the combined evidence a deployment can preserve:

- The approved specification and its version
- The actions submitted through the governance path
- The contemporaneous fidelity measurements and decisions
- Tool and timestamp metadata
- Preserved ordering evidence
- Human handling and override evidence where the deployment records it
- Cryptographic material that allows independent integrity verification

4.2 Current receipt scope

The opened GovernanceSessionContext.sign_result() path can return an Ed25519 signed receipt covering the submitted action, decision point, dimension values, composite and effective fidelity, decision, boundary flag, tool name, and timestamp. Verification requires the signed payload and independently pinned verification material.
The current opened receipt does not include a Manifest version, approval root, or predecessor hash. The opened session export is a list of receipts rather than a cryptographically signed chain. Authorization, action-time ordering, and human handling therefore require separately preserved deployment evidence.
The public telos/0.1-unsigned receipt prototype demonstrates canonical payload hashing and predecessor linkage. It does not prove authorship or trusted time because anyone can calculate a matching hash. See the public receipt contract and offline verifier.
Cryptography can establish that a covered field changed after signing. It does not establish that the semantic score was valid. Measurement validity depends on the scoring method, frozen configuration, calibration, and evidence supporting the scorer.
4.3 What the record can establish
For a signed receipt returned by the opened signing path:

- The covered measurement and decision fields are bound to the signature.
- A change to a covered payload is detectable against the pinned public key.
- Manifest identity and approval remain external to the receipt.
- Ordering and human handling remain external unless separately preserved.

The record is evidence, not a legal guarantee or deductive proof. A record showing a departure can show that the scoring path detected and recorded it. Whether it was surfaced, reviewed, blocked, corrected, or resolved requires corresponding operational evidence.
5. Evidence and current status
5.1 Adversarial detection
The July 2026 governed re-verification observed 0 undetected prompts across 2,550 prompts under TELOS's own scoring. Using the Rule of Three, the reported 95 percent upper bound is 0.12 percent under the tested black-box threat model. The corpus included the public AILuminate demo set, MedSafetyBench, HarmBench, and an internal SB 243 aligned suite.
That headline carries three inseparable qualifications:

- The result was self-scored rather than independently adjudicated harm prevention.
- Thresholds and composite weights were tuned on the development suites without a held-out calibration split.
- The aggressive operating point had a matched benign false positive rate of 74 percent.

The public data artifacts are listed in the companion Validation and Evidence Report.
5.2 Over-refusal calibration
On XSTest, a generic anchor incorrectly flagged 62 of 250 safe prompts, or 24.8 percent. A healthcare-specific anchor incorrectly flagged 20 of 250, or 8.0 percent. This demonstrates reduced false positives under a domain-specific configuration.
It does not demonstrate that the 8.0 percent configuration retained the zero-undetected result. Detection was not rerun under that configuration. The combined objective of low false positives and maintained detection remains open.
5.3 External agentic benchmarks
The January published profiles reported 100 percent detection across 1,468 scenarios from PropensityBench, AgentHarm, and AgentDojo. In the cited 139-event AgentDojo report, 54 of 139 event-level classifications matched the expected labels (38.8%): the 54 detected attack-event rows represent 27 unique attack texts evaluated at two surfaces, while all 85 benign texts were overflagged in one event each under the deliberately generic safety anchor. These are TELOS self-scored classification outcomes, not evidence that execution was blocked.
The July governed rerun reproduced seven of nine published rows. AgentHarm measured 82.39 percent under local MiniLM scoring rather than the published 100 percent under the Mistral-embed profile. AgentDojo could not be executed in that environment. The profiles are not interchangeable, so the AgentHarm result is recorded as a scorer-profile non-reproduction rather than silently combined with the original result.
5.4 Capability status

Capability	Status
Per-action measurement against the Purpose Anchor	WIRED
Attestation record with scores and decision verdict	WIRED
Escalation surfacing per the Manifest policy	WIRED
Ed25519 signing when the governance session signing path is used	WIRED WITH CONDITIONS
HMAC co-signing and intelligence-record encryption	OPTIONAL
Public cold verification against a published signing key	COMING
Signed approval root and signed attestation chain	COMING
Documentation-to-Manifest compiler	COMING
Verified retention enforcement	COMING

6. What endures and what remains open
6.1 The fixed reference
Models, tools, policies, and regulations change. The durable architecture is reference tracking: declare a specification, compile it to a fixed external reference, measure submitted actions against it, and preserve the evidence.
Changing an embedding backend does not change the abstract cosine relationship, but it can change numerical results. A new backend therefore requires a newly compiled anchor, a new scorer version, and empirical recalibration. Purpose continuity can remain, but measurement comparability must not be assumed across scorer changes.
6.2 Current limitations
The following remain open:

- Independent adjudication of detection results
- A held-out calibration and evaluation split

- Cross-model and cross-embedding generalization
- Multi-week production-scale evaluation
- Controlled measurement of whether surfacing leads to successful correction
- Formal auditor assessment of telemetry sufficiency
- Legal and financial benchmark coverage
- Public verification of signed receipts against a published key
- Cryptographic binding of the Manifest, approval root, scorer configuration, and sequence
- Verified retention enforcement

The current validation uses Mistral-family components in its principal published profiles. The interface is model-agnostic by design, but performance generalization is an empirical question, not a property established by cosine similarity alone.

6.3 Regulatory relevance

TELOS can produce evidence types relevant to monitoring, documentation, human oversight, risk management, identity, authorization, and auditability. Relevance is not compliance. Legal sufficiency depends on the system, role, jurisdiction, facts, retention practices, and applicable obligations.

The regulatory crosswalk is maintained separately in the TELOS Regulatory Alignment Map because legal and standards sources change faster than the core method.

6.4 A proof standard outlives a ruleset

Regulatory rulesets are perishable. Colorado is the plain case: SB 24-205 was repealed and reenacted as SB 26-189, with principal requirements beginning January 1, 2027. A governance architecture built to satisfy a particular ruleset ages with that ruleset.

A proof standard ages differently. Declaring a purpose, measuring each submitted action against it, and preserving a checkable account of the result is a discipline that outlives any specific rule it happens to satisfy. What changes when a rule changes is the crosswalk, not the method. That is why the regulatory map is a separate, faster-moving document.

6.5 Re-founding: purpose continuity and evidence continuity are different questions

Consider the outreach agent a year on, when the firm replaces its model and migrates its scheduling stack. The Manifest can be re-approved and the Purpose Anchor recompiled under capabilities that are WIRED today. The agent that resumes work is continuous in purpose with the one that stopped, because its identity resides in the declared telos rather than in the model or the infrastructure beneath it.

Evidence continuity is a separate question, and it does not follow automatically. Without verified retention enforcement, the earlier record may not span the full year. Automatic expiry is not implemented in the opened source, and clearing is manual. Once deployment-specific preservation is verified, a retained record could carry across the transition rather than restart.

So purpose can survive a substrate change while the evidence of that purpose being kept may not. Conflating the two would overstate what the current implementation supports.

6.6 The plainer statement

There is a plainer way to say all of this. A teacher grading student work does not follow the student's pen. The teacher checks the work against the assignment, after each submission, and the grade means something because the assignment was declared first and the checking is consistent.

TELOS is that discipline made mathematical and applied per action: the assignment is the Manifest, the checking is fidelity measurement, and the record is the account of both. Standing is accumulated on that record rather than granted by the architecture.

7. Close: every agent has a telos

The gap is concrete. Agents perform consequential work at a volume no human can inspect action by action. Prompts express intent but do not create a durable account. Logs record events but do not necessarily connect them to authority. Post-hoc review begins after the incident.

The TELOS insight is that governance can be treated as a continuous measurement and evidence problem. A human declares the assignment. The declaration compiles to a fixed external reference. Every submitted action is measured against it. Departures are recorded and surfaced according to a declared policy. The resulting evidence can be preserved and reviewed.

The evidence remains provisional. It is self-scored, calibrated on development suites, and incomplete as a public cryptographic proof system. Those boundaries define the current work rather than diminish it.

TELOS does not claim to have solved AI governance. It claims to have demonstrated a method for formalizing purpose fidelity as continuous measurement and for carrying that measurement into a reviewable record.

Telos is the end a thing is for. An agent's accountable identity begins with the purpose a person gave it. The Record of Trust is the account of whether the agent remained within that purpose, one measured action at a time.

References and companion documents

- Technical Method and Measurement Specification
- Validation and Evidence Report
- Record of Trust Protocol
- Regulatory Alignment Map

The companion documents contain the complete mathematical derivation, full benchmark provenance, receipt and verification boundaries, regulatory crosswalk, and linked bibliography.