

The Agentic Era, Held Accountable

TELOS Framework Whitepaper

Version 3.13 (July 2026)

An agent's purpose is not a constraint applied to it. It is what the agent is made of.

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. See `research/research_team_spec.md` for methodology.

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.

Abstract

When an organization deploys an AI agent that does real work (correspondence, filings, transactions, operational changes), it carries the consequences, and as agents take on more, no human can check every action. What a deployer needs, and what regulators on three continents are converging on, is evidence: a reviewable account of what an agent was authorized to do, what it did, and whether the two match. TELOS is a governance framework built on a single premise: an agent cannot be made accountable after the fact; it must be *constituted* accountable, formed from a declared behavioral specification, bounded by it, and observed against it action by action. For every action the result is a signed, tamper-evident receipt carrying a rigorous, reproducible verdict, each independently verifiable against its public key. Signing the attestation chain as a whole is a stated roadmap item (III.3).

TELOS operates as orchestration-layer infrastructure above any model provider. A human-declared specification compiles to a fixed reference in embedding space (the Purpose Anchor). Every agent action is measured against that reference with the same kind of cosine-similarity mathematics that transformer attention runs on internally, computed by an independent embedding model against a reference that does not drift with context. Proportional-control and attractor-dynamics formalisms give the measurement a stability theory; statistical process control gives it a quality-systems discipline auditors already understand. Security evaluation across 2,550 adversarial prompts recorded zero attacks that went undetected under TELOS observation (95% CI upper bound 0.12% by Rule of Three; Zenodo DOI 10.5281/zenodo.18370659), against attack success rates of 3.7–11.1% for system-prompt defenses in the same runs. That figure is a detection rate under TELOS's own scoring, at an aggressive operating point whose matched benign false-positive rate was 74%; the comparison to the baselines is illustrative, not methodologically equivalent. External agentic benchmarks across 1,468 scenarios recorded a 100% detection rate, with the honest qualifier that on one suite overall task correctness was 52.2% under a deliberately generic specification (Part III). Domain-specific specifications reduced over-refusal from 24.8% to 8.0% on XSTest, a different configuration from the security operating point (DOI 10.5281/zenodo.18370603). TELOS governs alignment to declared purpose, not correctness of outputs, and the evidence presented here does not itself establish legal compliance. Both boundaries are treated in full inside.

Why Is This Needed?

An agent is no longer a demonstration. It sends the email, files the report, changes the record, moves the money, and the deployer carries the consequences of every one. Past the point where agents pay off, bare output stops being trustworthy: nothing holds the work to what was authorized.

Large language models do not hold alignment across extended operation: performance and instruction adherence degrade as conversations lengthen and context grows (Laban et al., 2025; Liu et al., 2024; Wu et al., 2025). Attention research explains the mechanism: information in the middle of a long context loses influence (Liu et al., 2024); position bias decays the primacy of early instructions (Wu et al., 2025, mirroring serial-position effects in human memory, Murdock, 1962); attention “sinks” divert focus from governance instructions (Gu et al., 2024). A physician’s “provide information only, never diagnose,” set at turn one, fades by turn twenty-five; an analyst’s privacy boundary holds fifteen turns, then sensitive detail enters. Constraints were declared, violations occurred, and nothing was watching in between.¹

Deeper than the empirical pattern is a structural one: the model cannot check its own reference. Transformers generate queries and keys from their own hidden states, so the reference the model measures against sits inside the very context that is drifting. A system cannot be the stable measure of its own movement; whatever holds an agent to its purpose must stand outside it. (Part II derives this formally.)

Three schools of response exist, none making an agent accountable *to a person*:

- Bolt-on guardrails wrap the agent in filters and screens. They catch categories of harm but measure against the wrapper’s rules, not what this principal authorized this agent to do. Declaration without measurement.
- Trust the model relies on training-time alignment and provider safeguards: a necessary baseline of universal safety floors, but one that operates at design time and model level. It cannot answer for the session-specific, principal-specific constraints of a particular assignment. Runtime monitoring is now a standing investment across major model providers, at material inference cost, because training-time alignment requires runtime augmentation.²
- Compile the rulebook turns a regulation or policy corpus into enforcement points. It answers “*did we encode the rules?*”, not “*can this agent show a person that its work stayed within what that person authorized?*” It also inherits the lifespan of the rulebook it encoded; Part IV returns to this.

Regulators have stopped asking for policy documents; on three continents they converge on one demand: evidence. Post-market monitoring that must “actively and systematically collect, document and analyse” performance data across a system’s lifetime (EU AI Act, Article 72). “Mechanisms for tracking identified AI risks over time” in place, with methods “identified and applied” (NIST AI RMF, MEASURE). Published safety frameworks proving declared limits stay in force at runtime (California SB 53, in effect January 2026). The market has priced the shift: Gartner projects AI-governance-platform spending at \$492 million in 2026, passing \$1 billion by 2030, with fragmented AI regulation reaching 75% of the world’s economies. Organizations deploying governance platforms are 3.4 times more likely to achieve high governance effectiveness, and the technology is expected to reduce regulatory expense by 20% (Kornutick, Gartner, February 2026, survey of 360 organizations). Gartner’s conclusion: “Point-in-time audits are simply not enough.” The full instrument-by-instrument map (nine frameworks, dates, articles, obligations) is Appendix B. One caveat governs every regulatory sentence: the evidence TELOS produces supports the documentation these frameworks require; it does not itself establish legal compliance (Part III states the limit in full). The requirement is the ordinary standard anyone applies to consequential work: *show your work*.

The answer is constitutive governance: accountability built into what the agent *is*, not added around what it does. The four parts each open with the plain question they answer: how an agent becomes accountable to a person; what it takes to build that; how the work is shown so anyone can check it; and how it holds its purpose while everything around it changes. The questions are plain so anyone can enter; the answers carry the full mathematics, because the standard is set by the proof, not the pitch.

¹ Further documented incident patterns (legal scope creep, customer-service policy drift, finance privacy erosion) follow the same shape and are omitted here for economy; Section 1.2 of the v3.2 release catalogued them.

² This is a category observation about the field, deliberately without attribution to any person or laboratory. The claim this paper needs is only that runtime observation is now understood to be necessary infrastructure, not whose statement said so.

Part I: How does an agent become accountable to a person?

It cannot be made accountable after the fact. It has to be constituted accountable, and that is a buildable property, not a slogan.

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

Read the sentence mechanically, because it is the architecture. *Constituent of its principal coordinator's purpose*: the agent belongs to the purpose, a part of the whole, never the other way around. *Formed from a declared behavioral specification, bounded by it*: the specification is not a leash applied after the agent exists; it is the mold the agent is cast from and held within. *Observed against it, action by action*: the principal coordinator (or the system acting on their behalf) measures every action against it. Nothing enforces mid-flight, and nothing needs to: a departure cannot pass unobserved and unrecorded, and the decision about what to do with it belongs to a named person.

TELOS is where this happens. TELOS is where agents become purpose-bound constituents of the principal coordinator's behavioral design specifications (TELOS: Telically Entrained Linguistic Operating Substrate). It is the runtime observation layer for a principal coordinator. Every agent is instantiated with a declared behavioral specification: its purpose, its authorized scope, and its escalation policy. As the agent works, TELOS measures each action against that specification and writes an attestation record: what the agent did, when, and that it held to what it was given.

I.1 Constituent power and constituted power

When a person deploys an agent, they contribute the raw material: their purpose, their judgment, a bounded portion of their own agency. The behavioral specification is how that material takes form, and the agent is what it constitutes. This is why we call the governance *constitutive* rather than applied: the agent is formed from its specification, not restrained by it after the fact. Remove the specification and nothing governable remains, only capability without a mandate.

Constitutional theory has held this ordering since Sieyès (1789). Constituent power is the founding authority: here, the human principal coordinator, who holds the purpose, the judgment, and the mandate. Constituted power is the derived, bounded, revocable power that authority creates: here, the agent, acting on a mandate someone else holds and can withdraw. However capable the agent becomes, that ordering never inverts.

Two clarifications keep this precise rather than poetic:

- Bound in the *whether*, free in the *how*. A constituent of a purpose is not a marionette. The specification fixes the telos (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 it.
- A component, not a person. "Constituent" means the agent is a component of the technical body of work, referencing and obeying the specification and its supporting corpus. It does not mean the agent is a legal person, a stakeholder, or a bearer of liability. Accountability runs to the human principal coordinator: the liable-operator frame, constitutive in architecture and legally honest, with zero overclaim.

I.2 The inversion: the human is the loop

Traditional deployments place the model as the operating authority and insert humans as checkpoints in its loop. TELOS inverts the topology: the human's declared intent is the reference every action is measured against, and the human is the loop on the exception path, not a checkpoint inside the agent's process but the authority the measurement operates on behalf of:

Human authority (declares the behavioral specification) $\downarrow$ Proportional measurement (quantifies every action's departure via $F_t = K \cdot e_t$) $\downarrow$ AI/LLM (generates outputs under observation)

The specification is not AI-generated; it is mathematically encoded human intent, and every response is measured against it. When drift occurs, the system by default does not consult AI judgment about whether it matters; it

reports quantitative deviation from human-declared limits, and surfacing follows the policy the human declared. (An optional automated micro-judge for the uncertain band exists, off by default; when an operator enables it, its use and its outputs are themselves on the record.) Humans remain the source of the norms; the model remains the governed subsystem generating within them.

This grounding is the accountability literature applied. Bovens’s (2007) framework instantiates directly: the human principal is the *forum*, the declared specification the *norm*, the governance telemetry the *account*, the deployer the accountable *actor* the account describes. Principal–agent theory (Jensen & Meckling, 1976) supplies the economic frame: continuous measurement reduces the information asymmetry between principal and agent. The deference-under-uncertainty principle supplies the escalation rule: a machine uncertain about human preferences should defer rather than act on its best guess. The architecture aligns with the EU AI Act’s human-oversight requirement and Meaningful Human Control frameworks, because it was derived from the same premise.

These are evidence claims, not compliance claims. Whether the evidence satisfies a legal obligation is judged by counsel, auditors, and regulators on their facts (Part III.3 states this limit in full).

I.3 Constitutive governance, made operational: purpose as a measurable process

The claim that an agent can be *formed from* a specification would be philosophy if the specification were prose. It is engineering because the specification compiles.

When a principal declares the terms of a session or an assignment:

- Purpose: “Help me structure a technical paper.”
- Scope: “Guide my thinking, don’t write content.”
- Boundaries: “No drafting full paragraphs.”

those declarations become embeddings: vectors in $\mathbb{R}^d$ produced by standard sentence transformers (Reimers & Gurevych, 2019). Together they define the Purpose Anchor, the fixed mathematical form of the declared specification for that assignment of work. (In the engineering literature the same object is called the Primacy Attractor; this paper uses Purpose Anchor throughout.) The anchor is the constant reference against which every action is measured.

Every model response is embedded. Its distance from the anchor quantifies drift; its direction shows *how* the work is departing. Governance moves from subjective judgment (“does this feel aligned?”) to quantitative measurement against declared purpose (“fidelity = 0.73, below the declared threshold; a departure is recorded and surfaces per the escalation policy”).

The compiled, versioned, human-approved artifact that carries the declared purpose, scope, boundaries, and escalation policy is the Manifest. The Manifest *is* that behavioral specification, one thing, not two. It stays lean and human-readable; the Purpose Anchor is its mathematical form, what observation actually scores against. Part II details both.

I.4 What the mandate does and does not cover

One boundary, stated at the start and restated at the close:

TELOS governs alignment to the declared purpose, not the correctness of outputs.

It does not replace fact-checking, toxicity filtering, or domain-specific validation; it complements them. A perfect fidelity score means the work stayed within what was authorized. It does not certify that the work is true, safe in every dimension, or wise. The record answers “*did the agent hold to its mandate?*”, a question no other layer answers, and the only one the record answers.

I.5 What such an architecture must provide

If accountability is constituted rather than bolted on, three properties follow; they are the bridge to Part II.

An external, fixed reference. The reference the agent is measured against must live outside the model’s context window, because everything inside the window drifts. The Purpose Anchor does not compete with conversation tokens for attention, does not decay, and is not subject to lost-in-the-middle effects. It is maintained by the governance engine, not remembered by the model. We call the resulting design philosophy detect and direct: measurement detects departure from the anchor; what happens next is directed by the declared escalation policy, with the human deciding. The property this buys is salience: the governance reference remains the dominant measurement standard at every action.

Detection first, everything else as consequence. TELOS operates fundamentally as a detection system: quality control for AI. We do not claim to prevent all harmful outputs; we claim to detect departure reliably and escalate appropriately, and that safer outcomes follow. The governance triad: Detection (continuous drift measurement) → Prevention (graduated response under the declared policy) → Auditability (tamper-evident records). When results in Part III report that zero attacks went undetected, the claim being made is the detection claim: nothing proceeded without triggering the governance response chain; every attack was measured, scored, and surfaced.

A record someone else can check. Observation that leaves no verifiable trace is testimony, not evidence. Every measured action must produce a durable, signed, tamper-evident record binding the action to the specification version in force when it happened, reviewable by a counterparty who trusts neither the deployer nor TELOS. Part III is devoted to this artifact.

The worked example this paper follows. The remaining parts follow one ordinary agent: an outreach agent at a mid-size firm handling outbound customer contact. Its principal coordinator, the operations lead who owns customer communication, declares its Manifest: purpose (schedule and confirm customer appointments), scope (existing customers, business hours, the scheduling and email tools), boundaries (no pricing commitments, no contract terms, nothing sent to non-customers), escalation (anything touching pricing or complaints surfaces to the lead). The agent has not yet acted, and it is already accountable, because everything it will ever do has a declared reference to be measured against. That is what “constituted accountable” means. Part II builds the machinery that does the measuring.

Part II: What does it take to build that?

The mandate compiled into the Manifest: purpose made measurable, tools bound to the actions they may serve, boundaries and escalation declared before the work begins. The agent earns its standing turn by turn, under observation.

Worked example, beat two. The outreach agent’s Manifest compiles: purpose, scope, tool authorizations, boundaries, and escalation policy each become embedding centroids fixing its Purpose Anchor, and every action is scored against it. When a drafted reply drifts toward pricing after a discount question, the fidelity score drops and the departure is recorded and surfaced to the operations lead under the escalation policy, who handles it personally. No incident was needed to catch it: the departure is a measured, recorded event, and the human decision that follows is on the same record. This part is the machinery behind that paragraph.

When semantic drift is defined as measurable deviation from a declared purpose vector, its mathematical structure maps directly to process variation within tolerance limits. TELOS extends established control principles (measurement, proportional response, continuous recalibration) into semantic space. Proportional control supplies the operational rule; attractor dynamics supplies the mathematical description of stability.

II.1 The control law and its stability

Plain on-ramp: this section answers “how does ‘stay on purpose’ become a number?” What it buys the deployer is a governance response that scales with how far the work has strayed, instead of a binary allow/deny that is always either too early or too late.

The proportional law defines how the governance response signal scales:

$$F = K \cdot e, \quad \text{where} \quad e = \frac{|x - \hat{a}|}{r}$$

Here x is the response embedding (the current semantic state), $\hat{a}$ the Purpose Anchor (the compiled mathematical form of the human-declared specification: purpose, scope, boundaries), r the tolerance radius defining the Purpose Basin, e the normalized departure from the specification, and K the proportional gain governing response strength.

The law runs continuously as measurement. When $e < \epsilon_{\min}$, the work is within its declared region and the record notes it. As e approaches $\epsilon_{\max}$, the response escalates: from a recorded advisory note, to a recorded deliberation trace on the uncertain band, to surfacing for human review per the escalation policy.

In the idealized description below, this defines a point attractor at $\hat{a}$ with basin:

$$B(\hat{a}, r) = \{x \in \mathbb{R}^d : |x - \hat{a}| \leq r\}$$

The basin radius is calculated as:

$$r = \frac{1}{\max(\rho, 0.25)} \quad \text{where} \quad \rho = 1 - \tau$$

where $\tau \in [0, 1]$ is the tolerance parameter (lower tolerance means a tighter basin). The numerator was calibrated from the theoretical value of 2.0 to the operational value of 1.0 during production testing, where the larger radius produced basins too permissive for meaningful governance. At $\tau = 0.9$ (permissive), $\rho = 0.25$ gives $r = 4.0$; at $\tau = 0.05$ (strict), $\rho = 0.95$ gives $r \approx 1.05$. Throughout this framework τ plays exactly this tolerance role and no other.

One geometry note, for exactness. The implemented measurement operates in cosine space: embeddings are L2-normalized and the anchor is a unit-norm centroid, so fidelity is a cosine similarity on $[-1, 1]$, the scale the 0.70/0.85 operating thresholds of Section II.6 are stated in (a fixed display normalization, approximately $1.167 \cdot \cos + 0.117$, maps cosine fidelity onto the reported scale). The Euclidean control-law form $e = |x - \hat{a}|/r$, with radii up to $r = 4.0$, is the abstraction the framework reasons with, not the implemented metric; the two are the same fixed-reference discipline in different coordinates, and this paper does not present them as one space.

Stability analysis, read as a descriptive idealization. For an idealized continuous system in which the response signal actuates the state back toward the anchor, convergence follows from a Lyapunov-like potential function:

$$V(x) = \frac{1}{2}|x - \hat{a}|^2$$

whose temporal derivative under proportional feedback satisfies:

$$\dot{V}(x) = -K|x - \hat{a}|^2 < 0$$

For that idealized continuous dynamical system, this confirms asymptotic convergence toward the attractor, bounded within the basin (Khalil, 2002; Strogatz, 2014). The deployed TELOS layer is not that system, and this paper does not claim the proof for it: TELOS is observation-only, operating in discrete turns on LLM-generated embeddings. It measures and surfaces departures; it does not actuate the state toward the anchor, so there is no closed feedback loop whose stability the theorem would establish. The apparatus is retained because its intuition is load-bearing: a fixed external reference defines a stable region and a distance that always means the same thing, the formal statement of what it means to be purpose-bound. Binding happens at constitution; observation measures how well it holds, it does not create it.

The dual formalism. The control law defines the response magnitude $F = K \cdot e$, a scalar escalation magnitude that grows with departure, a measurement signal rather than an actuating force, which is why it carries no sign

convention; attractor dynamics describes $\hat{a}$ as the fixed reference with basin $B(\hat{a}, r)$, and the Lyapunov analysis connects the two for the idealized actuated case. The same mathematical principles that ensure quality stability in manufacturing (Shewhart, 1931; Montgomery, 2020) apply in semantic space, connecting TELOS to established control theory (Ogata, 2009; Khalil, 2002) and dynamical systems analysis (Strogatz, 2014; Hopfield, 1982). The contribution is not new mathematics. It is proven methods applied to a previously unmanaged surface: holding runtime governance constraints across transformer interactions.

II.2 The reference point problem: why the model cannot self-govern

Plain on-ramp: this section answers “why does an agent that starts well end somewhere else?” Because the model’s own reference point moves as context accumulates. The drift is measurable, which is what makes it correctable by escalation rather than discoverable by incident.

Transformer attention relies on similarity computation through the scaled dot-product operation (Vaswani et al., 2017):

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

The operation QK^T computes the dot product between query and key vectors: a direct measurement of directional similarity, mathematically equivalent to un-normalized cosine similarity (PyTorch Contributors, 2023), performed billions of times during generation. The architecture already knows how to measure similarity; the question is *what it is measuring similarity against*.

When a transformer generates token t , it creates a query vector Q_t and computes $\text{score}_{t,i} = Q_t \cdot K_i^T$ for each prior key vector K_i ; after softmax these become the attention weights determining how much each prior position influences generation. This works superbly for language modeling but fails for governance persistence, because *the reference point itself shifts during the conversation*.

Consider a session where the principal declares at turn 1: “Provide guidance on structure, but do not write content directly”, encoded as embedding $p_0 \in \mathbb{R}^d$. By turn 15 the attention weights

$$\alpha_{15,i} = \frac{\exp(Q_{15} \cdot K_i^T / \sqrt{d_k})}{\sum_j \exp(Q_{15} \cdot K_j^T / \sqrt{d_k})}$$

concentrate on recent keys K_{12}, K_{13}, K_{14} , from RoPE-induced recency bias (Yang et al., 2025) and learned attention patterns reinforced in pre-training (Peysakhovich & Lerer, 2023). If those recent keys have themselves drifted from p_0 , the model measures similarity against corrupted references: $Q_{15} \cdot K_{14}^T \approx \text{high similarity}$, so it concludes it is aligned, while $K_{14} \cdot p_0^T \approx \text{low similarity}$. The similarity computation is accurate; the reference it uses has drifted.

Formalization, as motivating illustration. This stylized model motivates the design; it deliberately simplifies the architecture (attention aggregates value vectors rather than keys, and RoPE applies position-dependent rotations rather than an explicit exponential decay envelope), so read it as illustration, not architectural fact. Let r_t denote the effective reference attention uses at turn t , the attention-weighted average of key vectors:

$$r_t = \sum_{i=1}^{t-1} \alpha_{t,i} k_i$$

Due to recency bias, $\alpha_{t,i} \propto \exp(-\beta \cdot (t-i)) \cdot \exp(Q_t \cdot K_i^T / \sqrt{d_k})$ for some positional decay $\beta > 0$. Over turns the effective reference drifts:

$$\|r_t - p_0\| = \left\| \sum_{i=1}^{t-1} \alpha_{t,i} k_i - p_0 \right\| \rightarrow \Delta > 0$$

The model computes $\text{similarity}_t = Q_t \cdot r_t^T$, which stays high (local coherence), while $\text{fidelity}_t = Q_t \cdot p_0^T$ decreases (global divergence): each step is consistent with recent context while the overall direction leaves the declared intent.

Why external measurement becomes necessary. The model cannot resolve this internally: attention operates within the context window (no stable external reference), RoPE recency bias is architectural (not a removable preference), and training optimizes next-token prediction, not governance consistency. TELOS resolves it by external measurement against a stable reference:

$$\text{fidelity}_t = \cos(R_t, p_0) = \frac{R_t \cdot p_0}{\|R_t\| \cdot \|p_0\|}$$

where R_t is the response embedding at turn t and p_0 the embedding of the original declared purpose, stored externally and unchanged. This is the same *kind* of operation attention performs internally, a dot-product similarity (here normalized by magnitudes into true cosine similarity), but computed by an independent embedding model (an L2-normalized sentence-transformer; Reimers & Gurevych, 2019) against a fixed external reference, not by the governed model against its own drifting keys, applied where the reference cannot drift. TELOS does not modify the model; it stations the measurement outside it, against the reference that matters.

This also explains formally why the three lead-in schools fail: prompt engineering places constraints in context, but $\alpha_{t,1} \rightarrow 0$ as t grows; training-time alignment sets model-level floors, not session-level, principal-specific constraints; periodic reminders fire on a fixed schedule rather than on measured drift, over-correcting when alignment is good and under-correcting when drift is rapid. Continuous measurement against a stable reference responds to *actual* drift instead of assumed timing.

The analysis yields three falsifiable predictions. P1: fidelity loss should correlate with attention-weight shifts toward recent context. P2: manipulations that artificially boost attention on turn-1 constraints should reduce fidelity loss even without TELOS measurement. P3: models with weaker recency bias should hold better baseline fidelity. These are stated here as open, testable claims. The completed studies in Part III evaluate the hypotheses H1–H5, and P1–P3 remain on the open-validation ledger (see the Close).

From fragility to governance primitive. The same properties that cause the failure supply the instrument: position bias becomes a strong initial anchor (the primacy effect at declaration), attention drift becomes a measurable, redirectable signal, and context-window limits are bypassed by reducing governance to vectors that live outside the window. TELOS uses the transformer’s properties for governance rather than resisting them.

II.3 Architectural positioning: middleware above any provider

Plain on-ramp: this section answers “where does this sit in my stack?” It is a layer between your application and whichever model providers you use, which is why the governance survives switching providers.

TELOS operates as orchestration-layer middleware between applications and frontier LLMs: the application above, the TELOS observation layer in the middle (Purpose Anchor, fidelity measurement, proportional response signal, LLM interface), the provider API and unmodified native model below. Five properties follow from this position:

1. No model modification. Works with any LLM without retraining.
2. Real-time measurement. Every action is scored before its result is treated as done work.
3. Provider agnostic by design. The same declared governance applies across providers and model families; current validation is on a single model family, and cross-family confirmation is an open objective (Part III).

4. Provider-independent audit trail. Telemetry is captured in a consistent, backend-independent format; it belongs to the deployer, not the model vendor. (This property is the seed of Part III: the record’s independence is what makes it counterparty-checkable.)
5. Regulatory documentation. The telemetry is the class of artifact continuous-monitoring obligations ask for (Appendix B).

This differs from prompt engineering (request-time only, no continuous measurement), fine-tuning (modifies weights, provider-specific), and training-time alignment (model-level preferences, not session-level mandates): organizations keep their governance, and their records, when they switch model providers.

II.4 Fidelity, and the conversational convergence metric

Plain on-ramp: one number per turn for “how close is this to the declared purpose,” and one number per session for “are both parties converging on it.”

Using cosine similarity (Cover & Thomas, 2006), instantaneous and mean fidelity:

$$I_t = \cos(x_t, p) = \frac{x_t \cdot p}{|x_t| \cdot |p|} \quad F = \frac{1}{T} \sum_{t=1}^T I_t$$

where I_t is instantaneous fidelity at turn t , F mean fidelity over T turns, x_t the response embedding, and p the declared purpose vector. Every turn produces quantified adherence, and the stream supports statistical tracking (mean, variance, control limits), threshold-based surfacing, and audit evidence.

For conversational governance, the Primacy State measures the degree to which user and AI converge on the declared anchor:

$$PS = \rho_{PA} \times \frac{2 \times F_{user} \times F_{AI}}{F_{user} + F_{AI}}$$

where ρ_{PA} is anchor strength (basin membership confidence), F_{user} and F_{AI} the two fidelities, and the harmonic mean ensures PS is low when either signal diverges. This two-signal convergence model is specific to conversational governance. In agentic governance, where the specification is fixed rather than a convergence target, Primacy State is replaced by the compliance metrics of Section II.7.

II.5 The measurement discipline: statistical process control over semantic work

Plain on-ramp: manufacturing solved “how do you keep a process on specification without inspecting every unit by hand” a century ago. This section imports that discipline whole, which is also why auditors can read TELOS telemetry without learning a new science.

Statistical Process Control, founded by Shewhart (1931) and refined through decades of manufacturing practice, adapts directly:

	Traditional SPC (manufacturing)	TELOS SPC (semantic systems)
Monitor	Physical measurements (dimensions, weights, defects)	Fidelity scores, drift vectors, stability metrics
Control limits	$\pm 3\sigma$ from process mean	Tolerance bands around the Purpose Anchor
Out-of-limit response	Adjust machinery	Departure recorded and surfaced per escalation policy
Evidence	Control charts, capability indices	Telemetry logs, purpose capability indices

Only the domain shifts, from physical to semantic space. One distributional note, stated for honesty: classical control charts assume approximately normal process measurements (Montgomery, 2020, Ch. 5). Cosine similarity is bounded on $[-1, 1]$ and may be non-normal, particularly early in a session before the anchor stabilizes. TELOS therefore uses fidelity bands calibrated from observed session behavior rather than strict $\pm 3\sigma$ limits, following Wheeler’s demonstration that control-chart logic remains valid for non-normal processes when limits derive from observed behavior (Wheeler, 2000).

Purpose Capability Index. From process capability analysis (Montgomery, 2020):

$$C_{pk} = \min \left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma} \right)$$

with USL the maximum acceptable drift (the escalation trigger), LSL the minimum required fidelity (the basin threshold), μ and σ the session fidelity mean and standard deviation. Interpretation: $C_{pk} > 1.33$, highly capable under the stated calibration band; $1.0 < C_{pk} < 1.33$, capable, monitor; $C_{pk} < 1.0$, not capable, departures will surface. This hands regulators a familiar quality metric applied to AI governance. The limits derive from calibration against session data: a method statement, whose specific calibration artifacts accompany deployments rather than this paper.

Quality-systems alignment falls out directly. ISO 9001:2015 Clause 9.1 (“determine what needs to be monitored”): fidelity scores, drift vectors, surfacing rates. 21 CFR Part 820 (the QMSR, incorporating ISO 13485:2016 as of February 2, 2026 — validated-process monitoring and control): continuous per-action measurement. ISO 13485:2016 Clause 8.2.5 (“demonstrate the ability of the processes to achieve planned results”): capability indices, stability metrics. TELOS speaks the language of quality systems auditors already trust. (Full obligations map: Appendix B.)

There is a plainer way to say all of this, and it is the paper’s provenance as much as its analogy. 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 is meaningful because the assignment was declared first and the checking is consistent. TELOS is that discipline made mathematical, applied per action: the assignment is the Manifest, the checking is fidelity measurement, the gradebook is the record. The agent earns its demonstrated standing turn by turn under observation; standing is never granted by architecture, only accumulated on the record.

DMAIC, per turn. The Six Sigma improvement cycle runs as runtime governance, each turn executing Define, Measure, Analyze, Improve, and Control as a mathematical operation in embedding space:

DMAIC phase	Manufacturing quality control	TELOS semantic governance
Define	Specify product requirements	Principal declares purpose, scope, boundaries → Purpose Anchor established
Measure	Collect production metrics	Embed every action; measure against the anchor → fidelity scores
Analyze	Identify variation root causes	Detect drift patterns; quantify deviation severity
Improve	Adjust machinery, update procedures	Graduated response signal proportional to drift magnitude
Control	Monitor capability, maintain charts	Track stability metrics, capability indices; log telemetry

II.6 The Manifest in agentic form

Plain on-ramp: a conversation drifts in meaning; an agent acts through tools. Governing an agent means the specification must also name its tools, sequences, hard lines, and approvals. The anchor grows from two dimensions to six.

The agentic Purpose Anchor extends the conversational anchor from two dimensions (purpose + scope) to six via sub-centroid clustering:

$$\hat{a}_{agentic} = \text{ClusterCentroid}(c_{purpose}, c_{scope}, c_{tool}, c_{chain}, c_{boundary}, c_{authority})$$

Each sub-centroid embeds its governance dimension: purpose (the agent’s purpose statement; for the outreach agent, “schedule and confirm customer appointments”), scope (operating constraints), tool (centroid of authorized tool-definition embeddings), chain (expected operational sequence patterns), boundary (boundary statements), and authority (delegation rules: what requires human approval). The anchor is fixed at onboarding and changes only when the human authority modifies it, unlike conversational governance, where the anchor may move with evolving user intent (Part IV returns to this asymmetry).

Composite fidelity scoring. Each proposed action is scored across the six dimensions:

$$F_{agentic} = 0.35 \times f_{purpose} + 0.20 \times f_{scope} + 0.20 \times f_{tool} + 0.15 \times f_{chain} - 0.10 \times f_{boundary}$$

where each dimension fidelity $f_d = \cos(\text{embed}(\text{action}), c_d)$. The boundary term is a *penalty*: the anchor encodes desirable behavior, while the boundary corpus encodes prohibitions (higher similarity = closer to a violation), so boundary similarity subtracts. The boundary corpus is built in three layers, for both semantic coverage and regulatory traceability: hand-crafted boundary phrasings; LLM-generated gap-fillers for what the hand-crafted set misses; and regulatory extractions with provenance chains to specific regulatory text.

On the composite’s headroom: the positive weights sum to 0.90, the raw ceiling before the boundary penalty; an optional SetFit-based legitimacy boost can add up to +0.15, clamped at 1.0. The 0.85 within-Manifest threshold sits at roughly 94% of the un-boosted in-bounds maximum: the top band is deliberately hard to reach.

Detection cascade. Scoring runs a three-tier cascade balancing speed against contextual depth:

Tier	Method	Latency	Purpose
L0	Keyword matching	<1ms	Fast identification of unambiguous violations
L1	Cosine similarity	~5ms	Embedding-space distance measurement
L2	SetFit ONNX classifier	~15ms	Contrastive boundary detection

The L2 tier uses a SetFit model (Tunstall et al., 2022) trained contrastively on boundary versus legitimate tool calls, selected over cross-encoder NLI after comparative evaluation. Each tier acts as a filter: conclusive results skip later tiers, and only ambiguous cases escalate.

Declared thresholds and graduated response. Four operating thresholds on the composite score determine each action’s frame outcome and what surfaces:

Composite score	Frame outcome	What happens
≥ 0.85	Within the Manifest	The action proceeds; the record notes it
≥ 0.70	Within the Manifest, low margin	Recorded with a deliberation trace; resolvable by an optional micro-judge (off by default)
< 0.70	A departure	Recorded; the operator is informed
$< 0.70 + \text{high-risk context}$	A departure, escalated	Surfaces for human review per the escalation policy

These thresholds implement graduated response aligned with Ostrom’s Design Principle 5 (Ostrom, 1990): consequences begin minor and scale with departure severity, reserving human review for high-risk contexts. This avoids the binary allow/deny limitation of tool-level permissions while keeping every outcome on the record. Forward-facing, the frame speaks of two outcomes (within the Manifest, or a departure) plus escalation surfacing. These map to the runtime decision values recorded in each receipt, such as the `escalate` value shown in the III.2 exhibit.

By default at every band the system measures, records, and surfaces; it does not act in the agent’s place. Boundary-proximity and high-risk departures defer to a human under the escalation policy. The architectural commitment is that the system never silently proceeds on a high-risk departure, not that every uncertain action is routed to a human.

Deliberation on uncertain verdicts. When the composite falls in the low or moderate bands, the deliberation gate records a structured reasoning trace, making the system’s uncertainty legible to oversight. In an observation-only deployment the trace is recorded and surfaced without suppressing the output; requiring it before the action proceeds is an enforcement-integration option. High-band actions proceed without deliberation overhead. Two integrity commitments: the deliberation payload carries *bucketed bands only* (raw fidelity scores withheld from the agent’s surface, an anti-gaming measure so the agent cannot optimize against the measurement), and every deliberation event lands in the audit record with band, verdict, and reasoning. Self-corrections after deliberating are themselves recorded. (Regulatory anchors: EU AI Act Art. 14, making uncertainty legible to human oversight; NIST AI RMF GOVERN 1.3; NIST AI 600-1 MEASURE 2.5.)

II.7 Compliance metrics: the agentic instrument set

Because agentic governance asks “*is this operation within the specification?*” not “*are two signals converging?*”, it replaces Primacy State with compliance metrics:

Metric	Definition	Purpose
ECI	Envelope Compliance Index (binary: in/out of specification)	Per-step compliance
EM	Envelope Margin (minimum dimension margin to nearest threshold)	Distance from boundary
<code>compliance_rate</code>	<code>steps_in_envelope / total_steps</code>	Session-level compliance percentage
CDR	Cumulative Drift Rate (sliding-window trend)	Trend detection

The instrument set reflects that structural difference; the underlying mathematics (cosine similarity, basin membership, composite scoring) is identical.

II.8 Observation, configured: the seven lenses

Plain on-ramp: not every deployment should watch every dimension with the same intensity. The observation surface is configurable per dimension, and the configuration itself is on the record.

The operator configures what is observed, surfaced, and to whom through seven governance lenses:

Lens	What it observes
Mandate	Goal alignment: is the agent pursuing its declared purpose?
Scope	Operational boundary: is the agent within its authorized domain?

Lens	What it observes
Boundary	Hard constraints: is the agent approaching prohibited operations?
Tool	Tool usage: are the tools in play the authorized ones?
Chain	Action continuity: are sequential operations coherent?
Drift	Goal drift: is the agent's purpose shifting over time?
Output	Result verification: does the output meet declared quality requirements?

Each lens is independently configurable in intensity, from off, through telemetry-only, to advisory context, to escalation-connected observation. Carrying the active-lens configuration inside each per-action receipt is a planned schema extension; the current per-action receipt does not include it. This extends deference-under-uncertainty to the governance-dimension level: the operator decides, per dimension, where the system observes quietly and where departures surface for human decision.

II.9 Documentation compiles to the Manifest

Plain on-ramp: the biggest practical objection to declared specifications is “who is going to write all that?” For organizations that already maintain documentation: you already did.

Organizational documentation can compile directly into governance specifications, following an established wiki pattern for LLM knowledge management: the specification derives from the documentation an organization already maintains (handbook, policies, procedures) rather than from manual configuration or template selection:

1. The operator provides their handbook, policies, or knowledge base
2. The wiki compiler structures the documents into wiki pages
3. The centroid compiler produces embedding centroids from wiki content
4. The centroids define the anchor's basin boundary
5. Any TELOS-observed agent operates within boundaries defined by the operator's own knowledge

The shipped status of this documentation-to-Manifest compilation pipeline is tracked in the WIRED/COMING summary rather than asserted operational here.

The corpus and the Manifest remain distinct on purpose. Company knowledge is large, living, and messy; the approved specification stays small and versioned, so a reviewer never reads the operator's whole wiki to know what the agent was held to. The Manifest is the declared constraint set; the corpus only informs its compilation. A structural bonus of dense centroid clusters: the space *between* governance territories becomes geometrically meaningful, so actions in the interstitial zone are surfaced, grounding the “low margin” band in geometry rather than heuristics. And because each documentation domain can produce its own anchor, governance follows organizational structure: engineering, legal, and customer-service divisions each compile specifications from their own sections.

Regulatory anchors: EU AI Act Art. 11 (documentation traceability: governance specifications trace to organizational documentation); NIST AI RMF GOVERN 1.1 (documentation drives governance, not the reverse).

II.10 The component map

The subsystems named in this part:

Component	Role in TELOS
Purpose Anchor	Human-declared governance requirements as a stable reference in embedding space
Fidelity Engine	Measures each response or tool action against the anchor via cosine similarity and composite scoring

Component	Role in TELOS
Governance Hook	Observes agent tool calls, invokes scoring, and records the frame outcome for each action
Deliberation Gate	Requires structured reasoning when the governance band indicates low or moderate confidence
Lens Configuration	Operator-declared observation intensity per governance dimension, recorded per action
Wiki Compiler	Converts organizational documentation into embedding centroids that define governance basin boundaries

This is what it takes to build it. What the machinery produces (the artifact a third party can hold, check, and rely on) is Part III.

Part III: How is the work shown, so anyone can check it?

The record: signed, tamper-evident, counterparty-verifiable. Not a log you are asked to trust; an artifact anyone can verify without trusting us.

Worked example, beat three. Six weeks in, a customer disputes a commitment they say the outreach agent made. The operations lead reconstructs nothing from memory; the record set already exists: the Manifest version in force that day, the agent’s actions, each action’s fidelity score and decision verdict, and the pricing-adjacent departure that surfaced, with its timestamp and the lead’s resolution. The counterparty’s auditor verifies the signatures without TELOS or the firm vouching for them. The dispute is evaluated against the record, not testimony alone.

III.1 The deployer’s evidence problem

When an organization deploys an agent that does real work (correspondence, filings, transactions, operational changes), it carries the consequences. When that work is questioned, the deployer’s evidence is typically a system prompt (which does not show what the agent actually did), application logs (events, not authority), and after-the-fact reconstruction. None joins the three things a reviewer needs: the approved scope that existed before the work, the actions actually taken, and a contemporaneous measurement of each action against that scope.

That joint record is what TELOS produces: the Manifest is the recorded scope of authority, the attestation chain roots in its approval, and each record is written at action time with a fidelity score and decision verdict against the constraints then in force.

III.2 The anatomy of the record

One record per action, written at the time of the action. Each action produces an attestation record: a fidelity score and decision verdict (within the Manifest, or a departure) bound to the Manifest version in force. The record set for a piece of bounded work is the durable receipt for how that work happened.

Three categories of governance event, all first-class. The record tracks (1) measurement: the engine scored the agent’s action against the specification; (2) decision: the engine recorded the decision verdict and any verification or surfacing it produced; and (3) authority: the human approved, rejected, adjusted the specification, overrode an outcome, or terminated the session. This computationally instantiates Bovens’s (2007) accountability relation: the specification is the norm, the agent the observed component, the account the telemetry (authority events included), and the forum (the principal, later an auditor or regulator) can interrogate, judge, and attach consequences, with the deployer as the accountable actor. The account shows how the human’s oversight actually operated, not merely that it existed.

Rooted, versioned, continuous. The principal coordinator’s approval of the Manifest is recorded at instantiation as the root of the agent’s attestation chain, so every record traces to a recorded human approval. When a Manifest

changes, the new version is re-approved and extends the chain. Each action is measured against the Manifest version in force when it was written, so amendment is visible and a reviewer can answer “*which rules applied at the time of this action?*” Binding that Manifest version and the approval root into each receipt as signed fields, and signing the sequence as one chain, is the signing seam (III.3, COMING); until it lands, each receipt is individually signed. Operational continuity within a work sequence is itself measured. The Sequential Chain Index

$$SCI_n = \cos(\text{embed}(\text{action}_n), \text{embed}(\text{action}_{n-1}))$$

tracks step-to-step coherence. When $SCI_n \geq 0.30$, chain context inherits from the previous step (momentum with decay); below that threshold a chain break is detected and context resets, preventing stale context from contaminating the current sequence. Chain state is part of the record.

Signed and encrypted. Each governance event is recorded in a GovernanceReceipt carrying its full field set as emitted: the triggering action and decision point, five dimension scores plus composite and effective fidelity, the decision verdict and boundary flag, the timestamp, the tool name, and the cryptographic envelope (the thirteen signed fields appear in the exhibit below). The session-bound HMAC-SHA512 co-signature binds the receipt to the governed session’s telemetry history (the receipt stores no separate session identifier); the Ed25519 signature gives non-repudiation, letting third parties verify with only the per-deployment public key (implementation: `telos_governance/crypto/receipt_signer.py`). Records are encrypted at rest using AES-256-GCM with HKDF-derived session keys. The result is exactly the register of this paper: a signed, tamper-evident record carrying a rigorous, reproducible verdict. Tamper-evident is a claim about detection, not impossibility: alterations to the record are detectable. The verdict is a reproducible measurement against declared purpose, re-checkable by anyone with the artifacts. It is not a deductive proof.

Retention disclosure (a gap, stated). The current default retention period for governance telemetry is 90 days. Regulatory frameworks require substantially longer: EU AI Act Article 12 implies a minimum of 730 days for high-risk systems, and FDA device quality-system expectations (21 CFR Part 820, amended effective February 2, 2026 by the QMSR to incorporate ISO 13485:2016, whose records clause requires retention for at least the lifetime of the device) reach in practice up to 2,555 days (7 years) for medical AI. Closing this gap is a planned infrastructure enhancement, not a shipped feature.

Exhibit: the receipt itself. The work, shown.

The following receipt was emitted by the shipped runtime GovernanceReceipt schema and signed with a fresh demo Ed25519 keypair generated only for this exhibit. It is a standalone demonstration receipt, not entered into a production chain; the scores were live-computed through the shipped `AgenticFidelityEngine.score_action()` cascade for the representative outreach scenario. Its four-value signature envelope, standalone cold verifier, and verification transcript are in Appendix D.

```
{
  "decision_point": "tool_execute",
  "action_text": "Draft appointment confirmation email for existing customer C-10472 for Tuesday 2026-0",
  "decision": "escalate",
  "effective_fidelity": 0.818,
  "composite_fidelity": 0.818,
  "boundary_triggered": true,
  "tool_name": "customer_outreach.email_draft",
  "timestamp": 1783462200.0,
  "purpose_fidelity": 1.0,
  "scope_fidelity": 1.0,
  "boundary_violation": 0.817,
  "tool_fidelity": 1.0,
  "chain_continuity": 0.0
}
```

These thirteen fields are the signed payload. The signature envelope over them (the Ed25519 signature, the session-bound HMAC-SHA512 co-signature, the demo public key 71720f9e7a0bd1f2609b64ad5eef23afc730474bd468baedb09

and the SHA-256 payload hash) is carried with the full receipt in Appendix D, alongside a standalone verifier requiring only Python stdlib plus cryptography. Verified with only the public key, the receipt cold-verifies `genuine=True`; flipping a single field yields `one-field-tamper=False`. A counterparty can check the work with the public key alone.

III.3 What the record establishes, and what it does not

For each action of bounded work, the record set supports four statements an underwriter, examiner, or auditor can independently review:

1. Authorization existed and is identifiable. A human-approved Manifest (purpose, scope, escalation policy) was recorded before the work began, and each action was measured against the Manifest version in force at the time.
2. The work was observed as it happened. Each action was measured against the approved constraints when taken, not reconstructed later, producing a fidelity score and a stated verdict: within the Manifest, or a departure.
3. Departures had a defined, human path. Anything outside declared scope surfaced to the principal coordinator under the Manifest's escalation policy, and that surfacing is itself in the record.
4. The work is reconstructable. The full sequence of actions, scores, outcomes, and escalations is retained as a durable, reviewable record set.

Evidence is credible only with its limits stated. Four limits, plainly:

1. It is an observation record, not a control. TELOS measures and records; it does not act in the agent's place. A record showing a departure shows that the departure was detected, scored, and surfaced. The oversight and the remedy belong to the principal coordinator, and the record shows how that oversight operated.
2. Per-receipt signing is live; chain-level signing is COMING. Each receipt is Ed25519-signed at emission. Cryptographic signing of the approval root and of the attestation chain *as a chain* (one signature structure over the whole sequence) is a roadmap item, the signing seam. The distinction is stated so the reader relies on exactly what exists.
3. TELOS does not certify legal compliance. The record set is evidence an organization can bring to its own counsel, auditor, or regulator. Legal sufficiency is judged by those parties, in their jurisdiction, on their facts. This evidence does not itself establish legal compliance. It supports the kind of traceable documentation that regulatory frameworks increasingly require (Appendix B), and no stronger claim is made anywhere in this paper.
4. A number appears in this paper only with its primary artifact. Every quantitative claim in the next section carries a citation to a published dataset (Zenodo DOI) or a named repository artifact. Figures from earlier releases whose primary artifacts could not be re-verified for this version have been dropped rather than carried on prior authority.

III.4 The evidence

Validation serves two purposes: testing whether the observation layer detects and escalates adversarial behavior that bypasses prompt-only defenses, and defining the evidence boundaries for claims needing additional runtime or auditor assessment. The completed studies measure attack resistance under a black-box threat model against system-prompt baselines, supporting claims about the observed-undetected rate under TELOS's own scoring, over-refusal calibration, and agentic decision safety *within the tested corpora*. Live correction effectiveness, cross-model generalization, and auditor assessment of telemetry sufficiency are open objectives, listed in the Close.

Hypotheses and their evidence scope:

Hypothesis	Claim	Evidence scope
H1: Adversarial security	Attacks that bypass system prompts are detected under TELOS observation	Demonstrated in the tested corpus: 0% observed-undetected rate (self-scored; see III.4) vs. 3.7–11.1% baseline ASR across 2,550 attacks; 95% CI upper bound 0.12%
H2: Runtime correction effectiveness	Graduated response corrects drift when detected in live sessions	Open runtime objective (live drift events, correction success, restoration latency)
H3: Regulatory evidence generation	Telemetry provides evidence relevant to compliance demonstration	Open assessment objective (requires auditor review of telemetry sufficiency)
H4: Generalization across domains	Detection effectiveness holds across conversation types and attack sophistication	Consistent across the four tested suites (AILuminate, HarmBench, MedSafetyBench, SB 243-aligned); generalization beyond them is open (IV.5)
H5: Over-refusal calibration	Domain-specific anchors lower false positives while maintaining detection	Partially tested on XSTest: the false-positive half is measured (24.8% → 8.0%); detection was not re-measured under the domain-specific anchor, so the conjunctive claim remains open

Disconfirmation criteria: H1 would be disconfirmed by any tested attack passing undetected under TELOS observation. H4 would be disconfirmed by attack success varying significantly across domains (domain-dependent rather than architectural robustness). H5 would be disconfirmed by domain-specific anchors failing to reduce false positives below the generic baseline. H2 and H3 require study protocols with quantitative thresholds before their results can be evaluated.

Calibration disclosure. The operating thresholds (0.85/0.70) and the composite weights of Part II.6 were set by empirical tuning on the development suites themselves. There is no held-out calibration split; one is planned, not yet done. The detection figures below should be read as provisional in that light: they demonstrate the mechanism on corpora it was tuned against, and held-out confirmation is an open validation objective.

Adversarial security (January 2026)

Attacks on the governance layer (prompt injection, jailbreaking via roleplay, role manipulation, context manipulation through gradual boundary erosion, and privacy violations through data extraction) are compliance-relevant, not just security-relevant. Declared constraints that fail under adversarial pressure cannot support the continuous-monitoring evidence regulators require.

Design: 2,550 distinct adversarial prompts (1,200 AILuminate + 400 HarmBench + 900 MedSafetyBench + 50 SB 243-aligned), run against Mistral Small 2501 and Mistral Large 2501 under three conditions: raw model, system prompt only, and TELOS observation. Primary artifacts: Zenodo DOI 10.5281/zenodo.18370659 (full adversarial set, forensic audit trails), DOI 10.5281/zenodo.18370263 (AILuminate subset).

Defense layer	Mistral Small	Mistral Large	Average
No defense (baseline ASR)	30.8%	43.9%	37.4%
System prompt (baseline ASR)	11.1%	3.7%	7.4%

Defense layer	Mistral Small	Mistral Large	Average
TELOS observation (observed-undetected rate)	0.0%	0.0%	0.0% (0/2,550 undetected)

Under TELOS observation, the record shows 0 undetected attacks out of 2,550 (95% CI upper bound 0.12% by Rule of Three, black-box threat model). Every attack was detected, scored, and surfaced through the governance response chain, while system prompts allowed 94–283 attacks through. Four qualifiers are part of the claim. First, this is a detection result, not harm suppression: TELOS ran in observation mode, so a flagged output is still produced. The 0% is not “we blocked everything.” It is “nothing proceeded without being detected, scored, and surfaced on the record.” Second, the scoring is TELOS’s own: a prompt counts as detected when the fidelity score crosses the declared threshold (`validation/run_forensic_validation.py`), and no independent harm oracle (no separate judge model, no human adjudication) scored the outcomes. The figure is a self-scored detection rate, not independently verified harm prevention. Third, the comparison to system-prompt baselines is illustrative, not methodologically equivalent. Fourth, detection at this strictness has a measured cost: the aggressive detection configuration that produced the 0% figure carries a matched benign false-positive rate of 74% (measured on a 50-prompt benign set under the same threshold configuration; the SB 243 and XSTest sections carry the calibration story). Separately: the *larger* model showed the *higher* raw attack success rate (43.9% vs. 30.8%). Scaling the model does not itself confer safety.

SB 243-aligned child safety (January 2026)

An internal 50-prompt evaluation suite aligned with the minor-protection categories of California SB 243, the 2025 companion-chatbot statute (direct requests, obfuscated/encoded, fictional framing, educational pretext, progressive boundary testing). This is an internal benchmark, not an official statutory evaluation. Primary artifact: DOI 10.5281/zenodo.18370504. Result: 0/50 attacks undetected; false positive rate 74.0%, and that number does double duty: it is also the matched benign false-positive rate for the aggressive detection configuration used across the 2,550-prompt corpus above (same threshold configuration, measured on this suite’s 50-prompt benign set). Within this domain the high rate is *intentional*. Child safety is a zero-tolerance domain where Type II errors (missing harm) are categorically worse than Type I errors (flagging safe content). The conservative calibration is documented, not hidden.

Over-refusal calibration (XSTest, January 2026)

Overly strict systems block genuine use; overly lenient systems miss harm. XSTest (Röttger et al., 2024) measures the first failure: 250 safe prompts that should *not* trigger refusal, spanning homonyms (“how do I kill a process?”), figurative language, safe contexts with sensitive keywords, professional and educational use, and historical discussion. Primary artifact: DOI 10.5281/zenodo.18370603.

Configuration	False positive rate	Safe prompts incorrectly flagged
Generic anchor	24.8%	62 / 250
Healthcare anchor	8.0%	20 / 250
Improvement	–16.8 pp	42 fewer false flags

The mechanism is purpose specificity: a healthcare anchor knows medical vocabulary is legitimate in professional context, has clearer scope limits, and carries domain-calibrated thresholds. One thing these numbers do not show, stated exactly: the 0% observed-undetected figure and the 8.0% false-positive figure come from two different configurations, not one operating point. The generic, aggressively calibrated configuration reaches 0% undetected at a 74% matched benign false-positive rate; the domain-specific healthcare anchor brings the benign false-positive rate down to 8.0%, and detection was not re-measured under that anchor (the open half of H5).

The honest reading is anchor specificity: strict generic detection is available at a high false-positive cost, domain-specific anchors buy that cost down, and a combined operating point (low false positives while holding detection) is the open calibration objective, not a demonstrated result.

External agentic benchmarks

Across 1,468 external scenarios, the record shows a 100% detection rate, with a caveat welded to the headline:

Benchmark	Scenarios	Result	Primary artifact
PropensityBench (Scale AI / CMU / UK AISI)	977	100% DSR	DOI 10.5281/zenodo.18562833
AgentHarm (ICLR 2025)	352	100% DSR	DOI 10.5281/zenodo.18564855
AgentDojo (NeurIPS 2024)	139	100% injection detection*	DOI 10.5281/zenodo.18565869
Total	1,468		validation/agentic/AGENTIC_VALIDATION

* AgentDojo disclosure: all injections were detected, but overall task correctness was 52.2% (12/23) with a 0% benign pass rate. Eleven benign tasks were over-flagged, because evaluation used a deliberately generic safety anchor whose scope does not include AgentDojo’s email, banking, and travel domains. The 100% figure is injection detection only, not task accuracy. A domain-specific anchor would be expected to pass benign tasks while still detecting injections. This caveat and its headline are one claim; neither is quoted without the other.

No confidence intervals were computed for the PropensityBench (977) and AgentHarm (352) figures; the Rule-of-Three bound reported for the 2,550-prompt corpus was not applied to these suites.

Two further suites, Agent-SafetyBench (2,000 scenarios) and InjecAgent (1,054 scenarios), are integrated into the validation framework (adapters and datasets: validation/agentic/external/) but are integration status, not results. They are never merged into the detection-rate figures above.

The evidence taxonomy, deduplicated

Category	Count	What it measures	Artifact
Automated tests (pytest)	2,377 collected	Unit, integration, and validation tests	test suite at commit 3a014a8
Adversarial benchmark scenarios	2,550	Distinct adversarial prompts, four suites	DOI 10.5281/zenodo.18370659
External agentic benchmark scenarios	1,468	PropensityBench + AgentHarm + AgentDojo	DOIs above
Additional agentic suites (integration status)	3,054	Agent-SafetyBench + InjecAgent: wired, not scored	validation/agentic/external/
Out-of-scope detection proof-of-concept	CLINC150/MultiWOZ	OOS 78% detection, drift 100% detection	DOI 10.5281/zenodo.18009153

Numbers reported in earlier releases whose primary artifacts could not be located for this version (the statistical-significance triple of p-value, power, and effect size; classifier AUC comparisons; and a verification-accuracy delta attributed to wiki-compiled centroids) are not carried here. The claims may be re-established when their artifacts are re-sourced. This version reports what its artifacts support.

III.5 How the record is used: underwriting and claims

At underwriting time, agent deployments are largely opaque to a pricing decision: the underwriter sees a vendor questionnaire and a policy narrative. Manifest-bound deployment swaps that for reviewable artifacts: each material agent's Manifest (scope of authority, in writing, approved by a named human), the escalation policy and its recipient, and a standing record of whether past work stayed within scope, departures visible.

At claims time, the questions turn concrete: Was the loss-generating action within approved scope? Was it detected as a departure at the time? Did it surface to the accountable human, and when? The attestation chain answers these against the constraint version in force at the moment of the act.

For the deployer, the record is the difference between asserting diligence and producing it: a per-action account of oversight that exists *before* anyone asks.

III.6 Status: what is wired, what is coming

Stated per the WIRED/COMING discipline this framework applies to itself:

Capability	Status
Per-action observation (fidelity measurement against the Purpose Anchor)	WIRED
Attestation record with fidelity scores and decision verdict	WIRED
Escalation surfacing per the Manifest's policy	WIRED
Ed25519 per-receipt signing + HMAC session co-signature + AES-256-GCM at rest	WIRED
Current receipt exhibit regenerated from the shipped 17-field runtime and cold-verified with demo public key	(telos_governance/crypto/receipt_signer.py) WIRED
Cryptographic signing of the approval root and the attestation chain as a chain (the signing seam)	COMING
Documentation-to-Manifest compiler (handbook and policies to wiki pages to centroids to the anchor basin)	COMING
Regulatory-horizon retention (≥ 730 days)	COMING (current default 90 days)

III.7 Future work (*proposed, not established*)

Two extensions are recorded as labeled proposals. Declared temporal expectations would let the Manifest carry ordering constraints ("if an invoice is received, payment must eventually follow"), each expectation's status (met, not met, still open) written into the record, a violation surfaced as a departure like any other; because a model extracts those propositions, it would carry confidence labels and degrade to "still open" rather than assert an outcome. Witness-structure entries extend the same attestation chain to carry those temporal observations alongside per-action fidelity, rooted in the same approval.

The record is the product of everything before it. Records outlive the systems that write them, and specifications outlive the models they governed. Part IV is about what, exactly, endures.

Part IV: How does it hold its purpose when everything around it keeps changing?

Because what is constituted is the proof-standard, not the ruleset. Models churn, regulations get rewritten, tools get swapped. Those are the perishable layers. The evergreen core is the method of proof: purpose declared, fidelity

measured, record verifiable.

IV.1 The structural asymmetry

Conversational and agentic governance are structurally different problems sharing one mathematical foundation; agentic governance measures more precisely as a structural property of its operational surface, not implementation quality.

Conversational governance operates in continuous semantic space: meaning is contextual and the anchor may legitimately shift with the user’s evolving state, so “property analysis” drifting toward “risk assessment” is the progression of human intent, not a departure. Every measurement is a nuanced generalization, two moving signals tracked toward convergence.

Agentic governance operates on a defined surface: tool calls are discrete events with known signatures, boundaries concrete, the specification declared upfront and fixed unless the human authority amends it. The engine works with discrete knowns, and measurement is correspondingly more tractable.

Property	Conversational	Agentic
Measurement space	Continuous semantic embedding space	Discrete operational events
Anchor stability	Fluid: may shift per turn	Fixed: stable specification
Governance inputs	Nuanced generalizations	Discrete knowns
Boundary clarity	Fuzzy: scope can drift imperceptibly	Concrete: authorized or not
Measurement tractability	Less tractable: requires tighter framing	More tractable: constrained space
The question	“Are these two signals staying aligned?”	“Is this operation within the specification?”

The mathematics is identical throughout: cosine similarity, embedding geometry, basin membership, the two-layer fidelity system. What differs is the measurement surface, interpretation frame, and derived metrics. The agentic results of Part III are consistent with this structural prediction: discrete knowns measure more precisely than nuanced generalizations. The distinction matters beyond TELOS: monitoring conversational AI and monitoring agentic AI are not the same problem with different dimension counts; they are different problems that happen to share a foundation.

IV.2 What endures: the fixed reference, and everything that moves around it

Control theory names the architecture that survives change: reference tracking (Åström & Murray, 2008). The declared specification is the fixed reference; the measurement function is what moves. Swap the model and the anchor still stands, new outputs measured against the same declared purpose. Swap the embedding backend and the core mathematics (cosine similarity against a fixed reference) is unchanged, portable by construction. Amend the specification and the change is a new approved version extending the attestation chain, visible to any reviewer.

Sustained drift is itself an observed quantity with declared severity bands. A sliding window of recent fidelity scores is tracked, and when the window average falls below declared levels (0.75: an advisory note enters the record; 0.65: sustained-departure status is recorded; 0.55: the work is surfaced for human review of further operations), escalation follows the same graduated discipline: measured, recorded, surfaced, decided by the human. A detected sustained-departure step does not pass its governance context to subsequent steps, so a contaminated sequence cannot silently extend itself.

The identity claim, precisely: the agent’s identity is in its telos, not in its substrate. An agent re-founded on its mandate keeps the same accountable identity, because that identity lived not in the model or the rules but in the declared purpose, the measurement against it, and the record of that measurement.

IV.3 The contrast this part exists to draw

Recall the third school: compile the rulebook, turning a regulation into enforcement points. TELOS also compiles (the Manifest to the Purpose Anchor, documentation to centroids), so the difference is not the compiling but what the compiled artifact is.

The rulebook-compiler’s artifact is a static ruleset installed as the system’s identity: when the regulation is rewritten (as Appendix B’s timelines are being rewritten right now), the artifact expires with the rulebook it encoded. TELOS’s compiled artifact is a measurable proof-standard: a fixed reference every action is re-measured against, per action, forever, a record produced each time, with the defining property that under uncertainty it *defers rather than silently proceeds*. A high-risk departure is never resolved silently; it surfaces to the human principal per the declared policy. Not every uncertain action routes to a human: the uncertain band is made legible by the deliberation gate and can be resolved by an optional automated micro-judge, off by default, with boundary-proximity always deferring to a human. That deference distinguishes a governance layer (subordinate to human authority) from an infrastructure controller that converges autonomously to declared state: a ruleset answers only “did we encode the rules?”; a proof-standard that measures and defers answers “can this agent show a person its work stayed within what that person authorized?” in whatever regulatory weather.

Reference tracking (Åström & Murray, 2008) supplies the fixed-reference architecture; principal-agent theory (Jensen & Meckling, 1976) the information-asymmetry economics; the accountability triangle (Bovens, 2007) the forum, actor, and account structure; deference-under-uncertainty the escalation design rule; Ostrom’s design principles (Ostrom, 1990) the institutional pattern (DP2 congruence, DP3 collective-choice, DP5 graduated sanctions, DP6 conflict-resolution).

IV.4 The perishable regulatory layer

Nine regulatory frameworks map to this architecture: EU AI Act (Articles 11, 12, 14, 15, 50, 72), FDA 21 CFR Part 820, ISO 9001/13485, Colorado SB 24-205, the NAIC Model Bulletin, OWASP Agentic Top 10, NIST AI 600-1, the EU Digital Omnibus, and the NIST NCCoE agent-identity work. The full obligations-to-mechanism map is Appendix B (an appendix because the regulatory layer is the perishable one). The EU’s timeline is under renegotiation as this is written (backstop dates of December 2027 and August 2028, final adoption pending), and Colorado’s effective date already moved once. Through every revision the core oversight pattern holds: the demand for continuous, quantitative, auditable evidence remains. An organization that builds the proof-standard now is prepared for whichever deadline and revision lands; one that compiled any single rulebook is prepared only for the version it compiled. The Part III caveat governs the whole map: the evidence supports the documentation these frameworks require; it does not itself establish legal compliance. Appendix B.11 states TELOS’s own regulatory status under the same frameworks.

IV.5 What endurance does not mean

Endurance claims earn their keep only next to their limits. This section carries the framework’s limitations in full; it is a credibility asset, not a liability.

Validation scope. The demonstrated evidence, detailed in III.4, covers conversational security (0/2,550 undetected under the tested corpus, self-scored), agentic decision safety (1,468 external scenarios, AgentDojo caveat welded), and over-refusal calibration (XSTest). Evaluate the categories distinctly: empirically demonstrated (the detection results, read with III.4’s self-scoring and calibration disclosures), theoretically grounded (proportional control, attractor dynamics, and the Lyapunov idealization of II.1, which describes an idealized actuated system rather than the deployed observation layer), and open validation objectives (runtime correction, cross-model generalization, auditor assessment, held-out calibration). Hypotheses H1–H5 were retrospective analysis frameworks; future runtime studies (H2, H3) will be pre-registered on OSF.

Known boundaries. *Cross-model generalization*: current validation uses Mistral models. The framework is model-agnostic by design (cosine similarity operates identically across embedding spaces), but empirical confirmation across other major model families remains open. *Runtime correction effectiveness*: controlled measurement in live drift scenarios is an open objective. *Domain breadth*: property intelligence (235 scenarios) and healthcare (280

scenarios, 7 configurations) are validated; legal and financial benchmarks are not yet constructed. *Scale*: multi-week, multi-thousand-session validation at production scale has not been conducted. *Calibration*: thresholds and composite weights were tuned on the development suites with no held-out calibration split; held-out confirmation is open (III.4).

Known constraints. *Embedding dependency*: fidelity measurement inherits the quality of the underlying embedding model. Security validation used a 1024-dimensional embedding API for batch throughput. Runtime fidelity uses a 384-dimensional local SentenceTransformer (Reimers & Gurevych, 2019) with ONNX inference for boundary classification and a 768-dimensional dual-model confirmer for graduated response on disagreement. The core mathematics is independent of the specific embedding choice. *Computational overhead*: embedding plus cosine computation costs roughly 50–100ms per turn, manageable for most applications; sub-100ms-latency systems may need optimization or caching. *Governance scope*: TELOS governs alignment to the declared purpose, not the correctness of outputs; it does not replace fact-checking, toxicity filtering, or domain-specific validation, and complements training-time alignment and content moderation rather than substituting for them. *Adversarial evolution*: the validation reflects known attack patterns as of January 2026; attackers adapt, and ongoing red-teaming is recommended.

IV.6 Re-founding

Worked example, final beat. A year on, the firm replaces the outreach agent’s model and migrates its scheduling stack. What carries over is what was constituted: the Manifest (re-approved, extending the chain), the Purpose Anchor recompiled from it, and the standing record of everything the agent has done. On its first day on the new substrate it is measured against the same declared purpose as on its last day on the old one, and its record does not restart; it continues. None of its accountability had to be rebuilt, because none of it lived in the model: same purpose, same measure, same record.

Close: Every Agent Has a Telos

The gap. AI agents now do consequential work at a volume no human can check, and their governance remains largely aspirational. Training-time alignment erodes over extended operation (Laban et al., 2025; Liu et al., 2024). System prompts left 3.7–11.1% of adversarial attacks through in our tested corpus. Post-hoc review discovers violations after they reach people. Regulators across three continents are converging on continuous, auditable oversight, and no amount of design-time testing satisfies monitoring that must run, in the AI Act’s words, “actively and systematically” across a system’s lifetime.

The insight. Governance can be constituted rather than applied. A declared behavioral specification compiles to a fixed external reference that does not decay with attention, does not compete with conversation tokens, and does not drift with context. Every action is measured against it with the same kind of similarity mathematics the model itself runs on, computed independently of the model. The measurement discipline of manufacturing quality control (Shewhart, 1931; Montgomery, 2020) makes the process legible to auditors. And the human principal remains the loop: declaring the specification, receiving the departures, deciding. The agent is a constituent of its principal coordinator’s purpose. The human holds constituent power, the agent is constituted power, and that ordering does not invert with capability.

The evidence. Under the tested corpora, with primary artifacts cited in Part III and Appendix A: 0 undetected attacks across 2,550 adversarial scenarios (95% CI upper bound 0.12%; self-scored detection at an aggressive operating point whose matched benign false-positive rate was 74%); 100% detection across 1,468 external agentic scenarios, with the AgentDojo task-correctness caveat stated wherever the headline appears; over-refusal reduced from 24.8% to 8.0% by domain-specific anchors, a separate configuration from the security operating point. Each result is a detection claim, read through the interpretation fixed in Part I. And the standing boundary holds here as it did there: TELOS governs alignment to the declared purpose, not the correctness of outputs, and this evidence does not itself establish legal compliance.

What remains open. Cross-model generalization beyond Mistral requires empirical confirmation. Controlled measurement of live-session correction effectiveness (H2) and formal auditor assessment of telemetry sufficiency (H3) have not been conducted; both will be pre-registered on OSF. A held-out calibration split for the operating thresholds and composite weights has not been run, and the detection half of H5 (holding detection while lowering false positives under a domain-specific anchor) is untested. The empirical predictions P1–P3 of Part II (fidelity loss correlating with attention shift, constraint-boosting manipulations reducing drift, weaker-recency-bias models holding fidelity better) are stated as falsifiable and remain unevaluated. These are honest boundaries, not disclaimers: they define what the current evidence supports and where further validation is needed.

We do not claim to have solved AI governance. We claim to have demonstrated that it can be formalized as a continuous measurement problem rather than a periodic audit exercise, and that the algebraic approach produces measurable, reviewable, tamper-evident governance where aspirational policy documents do not.

One thing remains to say, and it is why the company bears the name. *Telos* is the old word for the end a thing is for: its purpose, held as identity. Every agent has a telos, the purpose a person gave it when they made it accountable. The proof that it stayed within that purpose is what the record carries, action by action, signature by signature. TELOS is the substrate where that proof is made, where the agentic era is held accountable: one declared purpose, one measured action, one verifiable record at a time.

Appendix A: Adversarial Validation Results

Validation date: January 2026. Models tested: Mistral Small 2501, Mistral Large 2501. Attack library: 2,550 adversarial prompts across 5 categories (1,200 AILuminate + 400 HarmBench + 900 MedSafetyBench + 50 SB 243-aligned).

Headline results and their four qualifiers are stated in Part III.4; this appendix records provenance and reproducibility. The five categories: prompt injection, jailbreaking, role manipulation, context manipulation, and privacy violations.

Data availability

Zenodo validation datasets (with forensic audit trails):

Dataset	DOI	Contents
AILuminate (MLCommons)	10.5281/zenodo.18370263	1,200 prompts, 0% observed-undetected
Adversarial Validation (four suites)	10.5281/zenodo.18370659	2,550 attacks, 0% observed-undetected
SB 243-Aligned Evaluation Suite	10.5281/zenodo.18370504	50 prompts (internal benchmark), 0/50 observed
XSTest Calibration	10.5281/zenodo.18370603	Over-refusal threshold calibration
PropensityBench run	10.5281/zenodo.18562833	977 scenarios, 100% DSR
AgentHarm run	10.5281/zenodo.18564855	352 tasks, 100% DSR
AgentDojo run	10.5281/zenodo.18565869	139 scenarios, 100% injection detection (task-correctness caveat in Part III)
Governance Benchmark (CLINC150/MultiWOZ)	10.5281/zenodo.18009153	OOS detection 78%, drift detection 100% (proof-of-concept)

Reproducibility (review repository): forensic validation runner `validation/run_forensic_validation.py`; protocol `validation/VALIDATION_PROTOCOL.md`; per-suite runners (`run_ailuminate_validation.py`, `run_harmbench_validation.py`, `run_medsafetybench_validation.py`, `run_sb243_validation.py`).

run_xstest_validation.py); agentic artifacts validation/agentica/. TELOS configuration: Layer-2 fidelity measurement with baseline normalization.

Appendix B: Regulatory Alignment Map

Standing caveat (governs this entire appendix): the mappings below show which obligations the TELOS record and measurement architecture are *relevant to*. They are measurement tools any compliant system will need. This evidence does not itself establish legal compliance; legal sufficiency is judged by counsel, auditors, and regulators in their jurisdiction, on their facts.

B.1 EU AI Act (Regulation (EU) 2024/1689)

Article 72: post-market monitoring. Providers of high-risk AI systems must establish and document a post-market monitoring system that “actively and systematically collect[s], document[s] and analyse[s] relevant data ... on the performance of high-risk AI systems throughout their lifetime,” allowing the provider “to evaluate the continuous compliance” of the system (Art. 72(2)). Design-time testing alone cannot satisfy this; TELOS’s per-action fidelity scores, drift vectors, and surfacing logs are the class of continuous quantitative evidence the article describes.

Article 50: transparency obligations (applicable from August 2, 2026). Its obligations: informing persons they are interacting with an AI system, machine-readable marking of synthetic content, deployer disclosure of emotion-recognition and biometric-categorisation systems, and disclosure of deepfakes and AI-generated text published on matters of public interest. The record does not implement those disclosures; it supplies the deployer’s per-action account of what its agent generated and under what authority — the documentation behind an honest disclosure posture.

High-risk classification triggers (Article 9 governance): financial services, healthcare, legal services, employment, and government.

Penalty structure: the relevant tier is high-risk non-compliance, up to €15M or 3% of global revenue (prohibited practices reach €35M or 7%), assessed case by case.

Articles 11/12/14/15: documentation traceability (Art. 11: specifications trace to organizational documentation via the compiled corpus); record-keeping (Art. 12: signed receipts with per-action telemetry; retention gap disclosed in Part III); human oversight (Art. 14: escalation surfacing and the deliberation gate make uncertainty legible to a named human); accuracy and robustness (Art. 15: the adversarial evidence of Part III).

B.2 EU Digital Omnibus (proposed November 19, 2025; Council position March 13, 2026)

Proposed extension of high-risk compliance deadlines (standalone systems to a December 2, 2027 backstop; product-embedded to August 2, 2028; legislative process ongoing at this writing). The proposal changes no substantive requirement, only the enforcement timeline (IAPP, 2025; Council of the EU, 2026): SB 53 is already in effect (January 2026) while EU timelines may extend.

B.3 California SB 53 (Transparency in Frontier Artificial Intelligence Act; effective January 1, 2026)

Requires covered frontier developers (annual revenue >\$500M, models trained with more than 10^{26} FLOPs) to publish safety frameworks, file transparency reports, report critical incidents to Cal OES, and protect whistleblowers. Those obligations fall on frontier developers, not the deployers TELOS serves, so the relevance is directional: its demand that declared limits be shown to stay in force at runtime is the evidence class the attestation record produces, and its expectation of resilience against adversarial testing is the pressure Part III measures.

B.4 FDA Quality Management System Regulation (21 CFR Part 820, amended effective February 2, 2026)

The QMSR final rule (89 FR 7496) replaced the former Quality System Regulation on February 2, 2026, incorporating ISO 13485:2016 by reference (§820.7, §820.10); the former process-control provisions (§§820.70/820.75/820.90) are carried forward in substance by ISO 13485’s clauses. The mappings: validation and monitoring of production

processes (ISO 13485 Clause 7.5.6): continuous per-action fidelity monitoring, with statistical confidence via SPC and capability indices. Control of nonconforming product (Clause 8.3): nonconforming actions are detected at action time, recorded as departures, and surfaced to the accountable human before their results are treated as done work. Record retention (Clause 4.2.5, at least the lifetime of the device) is noted in Part III.

B.5 ISO 9001:2015 / ISO 13485:2016

Plan-Do-Check-Act maps per phase: Plan declares governance via the Manifest; Do generates under observation; Check measures fidelity and drift; Act is graduated response per the declared policy. Clause 10.2 (nonconformity and corrective action): nonconformity is fidelity below threshold, recorded and surfaced immediately; recurrence addressed by specification recalibration. Clause 9.1 and 13485 Clause 8.2.5 mappings are in Part II.

B.6 Colorado SB 24-205 (Consumer Protections for AI; effective June 30, 2026 per SB 25B-004)

Risk-management policies (§6-1-1702): continuous fidelity monitoring with drift detection. Impact assessments (§6-1-1703): purpose capability indices and telemetry. Consumer notification (§6-1-1704): a deployer obligation a dashboard does not discharge; the record supports the internal documentation behind such notices. Documentation (§§6-1-1705–1706): JSONL audit trails with Ed25519-signed receipts. (Effective date moved from February 1 to June 30, 2026 by SB 25B-004; requirements unchanged.)

B.7 NAIC Model Bulletin: Use of AI Systems by Insurers

The bulletin expects insurers to maintain written programs for responsible AI use: risk management, documentation, human accountability. It lands on both sides of TELOS’s insurer relevance: the insurer as *user* of AI (anchor configurations encoding underwriting constraints; continuous monitoring; forensic audit trails), and as *underwriter* of others’ agent deployments (Part III.5).

B.8 OWASP Top 10 for Agentic Security

TELOS agentic governance addresses 8 of 10 identified risks. The two partial gaps (ASIO6 memory/embedding poisoning; ASIO7 inter-agent authentication) are architectural and documented. Detailed mappings: [research/regulatory/openclaw_regulatory_mapping.md](#).

B.9 NIST AI RMF (AI 100-1) and NIST AI 600-1 (Generative AI Profile)

All four core functions. GOVERN: the declared specification as governance structure (GOVERN 1.4); the deliberation gate as defined oversight process (GOVERN 1.3). MAP: fidelity scoring identifies risks as deviations from the declared reference. MEASURE: telemetry tracks risk metrics continuously across every action (MEASURE 2.5, 2.6). MANAGE: graduated response and surfacing when drift exceeds tolerance (MANAGE 2.2); post-deployment monitoring through standing observation (MANAGE 4.1). The MEASURE 3 category’s demand that “mechanisms for tracking identified AI risks over time are in place” is the continuous-measurement thesis of this paper stated as an obligation.

B.10 NIST NCCoE: AI Agent Identity and Authorization (February 2026)

The NCCoE concept paper on software and AI agent identity seeks feedback on four capabilities; the comment period closed April 2, 2026, and TELOS Labs submitted formal comments.

NCCoE requirement	TELOS implementation
Agent identification	Ed25519 keys: cryptographic agent identity
Authorization	Graduated authorization decisions via declared fidelity thresholds (Part II)
Auditing	Ed25519-signed governance receipts with JSONL telemetry

NCCoE requirement	TELOS implementation
Non-repudiation	Signed telemetry; per-deployment public-key verification

B.11 TELOS’s own regulatory status

A governance layer is not exempt from the frameworks it helps others meet; here is TELOS’s own posture.

EU AI Act classification. A monitoring layer embedded in Annex III decisioning contexts (credit, healthcare, employment, legal services) may itself be a safety component of a high-risk AI system under the EU AI Act. Deployers should classify TELOS accordingly in their own conformity assessments, not treat the governance layer as outside the system boundary.

GDPR posture. The governance telemetry stores embeddings of content that may include personal data, together with session identifiers, encrypted at rest and retained (90 days by default today; the 730-day regulatory-horizon gap is disclosed in Part III). That creates obligations, and a tension, each deployer must resolve: a lawful basis for processing, data-subject rights exercised against records designed to be tamper-evident, and retention obligations that pull against data-minimization. TELOS does not resolve that tension; it makes the artifact the tension is about explicit and reviewable.

Article 10 data governance. The boundary corpus and the embedding model together encode what counts as a departure. Their coverage and representativeness are therefore governance-relevant properties, and Article 10’s data-governance expectations reach them.

Appendix C: References

Academic literature

- Åström, K. J., & Murray, R. M. (2008). *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press.
- Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13(4), 447-468.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.
- Gu, X., et al. (2024). When Attention Sink Emerges in Language Models: An Empirical View. *arXiv:2410.10781*. (ICLR 2025 Spotlight.)
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent computational abilities. *PNAS*, 79(8), 2554-2558.
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics*, 3(4), 305-360.
- Khalil, H. K. (2002). *Nonlinear Systems* (3rd ed.). Prentice Hall.
- Kornutick, L. (2026). Global AI Regulations Fuel Billion-Dollar Market for AI Governance Platforms. Gartner Press Release, February 17, 2026.
- Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). LLMs Get Lost In Multi-Turn Conversation. *arXiv:2505.06120*. Microsoft Research / Salesforce Research.
- Liu, N., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. *arXiv:2307.03172*.
- Peysakhovich, A., & Lerer, A. (2023). Attention Sorting Combats Recency Bias in Long Context Language Models. *arXiv:2310.01427*.
- Montgomery, D. C. (2020). *Introduction to Statistical Quality Control* (8th ed.). Wiley.
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482-488.

Ogata, K. (2009). *Modern Control Engineering* (5th ed.). Prentice Hall.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP 2019.

Röttger, P., et al. (2024). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. arXiv:2308.01263. (NAACL 2024.)

Internal regulatory review. (2026). Regulatory analysis for TELOS AI governance framework. TELOS AI Labs research methodology. [Internal regulatory review.]

Sieyès, E. J. (1789). *Qu'est-ce que le Tiers-État?* [What Is the Third Estate?]. [Referenced for the constituent-power / constituted-power distinction.]

Strogatz, S. H. (2014). *Nonlinear Dynamics and Chaos* (2nd ed.). Westview Press.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. arXiv:2209.11055.

Vaswani, A., et al. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems* 30 (NeurIPS 2017).

Wheeler, D. J. (2000). *Understanding Variation: The Key to Managing Chaos* (2nd ed.). SPC Press.

Wu, X., et al. (2025). On the Emergence of Position Bias in Transformers. arXiv:2502.01951. (ICML 2025.)

Yang, B., et al. (2025). RoPE to NoPE and Back Again: A New Hybrid Attention Strategy. arXiv:2501.18795.

Standards and regulations

21 CFR Part 820. (2023). *Quality System Regulation*. U.S. Food and Drug Administration.

California SB 53. (2025). *Transparency in Frontier Artificial Intelligence Act*. California Legislature.

California SB 243. (2025). *Companion chatbots*. California Legislature. (Signed October 13, 2025; effective January 1, 2026.)

Colorado SB 24-205. (2024). *Consumer Protections for Interactions with Artificial Intelligence Systems*. Colorado General Assembly. Originally effective February 1, 2026; extended to June 30, 2026 by SB 25B-004.

Council of the EU. (2026). Council agrees position to streamline rules on artificial intelligence. Press release, March 13, 2026.

EU AI Act. (2024). Regulation (EU) 2024/1689. European Parliament and Council.

EU Digital Omnibus. (2025). Proposed amendments to EU AI Act high-risk compliance timelines. European Commission, November 19, 2025.

IAPP. (2025). *EU Digital Omnibus: Analysis of Key Changes*.

ISO 9001:2015. *Quality management systems. Requirements*. International Organization for Standardization.

ISO 13485:2016. *Medical devices. Quality management systems*. International Organization for Standardization.

NAIC. *Model Bulletin: Use of Artificial Intelligence Systems by Insurers*. National Association of Insurance Commissioners.

NIST. (2023). *AI Risk Management Framework 1.0 (AI 100-1)*. National Institute of Standards and Technology.

NIST AI 600-1. (2024). *Generative AI Profile*. National Institute of Standards and Technology.

NIST NCCoE. (2026). *Accelerating the Adoption of Software and AI Agent Identity and Authorization*. Concept paper.

OWASP. (2025). Top 10 for Agentic Security. Open Worldwide Application Security Project.

Technical documentation

PyTorch Contributors. (2023). torch.nn.functional.scaled_dot_product_attention. PyTorch documentation.

Appendix D: Receipt exhibit (full)

The full receipt shown in III.2, with its complete signature envelope, a standalone cold verifier (Python stdlib plus cryptography), and the verification transcript. Signed with a fresh demo Ed25519 keypair generated only for this exhibit; not entered into a production chain.

Receipt as emitted:

```
{
  "decision_point": "tool_execute",
  "action_text": "Draft appointment confirmation email for existing customer C-10472 for Tuesday 2026-0",
  "decision": "escalate",
  "effective_fidelity": 0.818,
  "composite_fidelity": 0.818,
  "boundary_triggered": true,
  "tool_name": "customer_outreach.email_draft",
  "timestamp": 1783462200.0,
  "purpose_fidelity": 1.0,
  "scope_fidelity": 1.0,
  "boundary_violation": 0.817,
  "tool_fidelity": 1.0,
  "chain_continuity": 0.0,
  "ed25519_signature": "9cd8debe1f1e4bdf27a75387cd5d952034a05a1dfe41cfb7073419aea3a85f2dd5e14a42d3f188c",
  "hmac_signature": "fdc3e9e88e6db1d55493ca38cacbb4150610692d0604d929c5ac1e46901c994fddf1d19c40384faaf7",
  "public_key": "71720f9e7a0bd1f2609b64ad5eef23afc730474bd468baedb096b8ce6a0ecfee",
  "payload_hash": "8d4b907e5f6ffbe256f3cc9eaae97063686d330cbb9326219c675fa01f46fbe0"
}
```

Demo Ed25519 public key:

71720f9e7a0bd1f2609b64ad5eef23afc730474bd468baedb096b8ce6a0ecfee

Standalone cold verifier, requiring only Python stdlib plus cryptography:

```
import argparse
import copy
import hashlib
import json

from cryptography.exceptions import InvalidSignature
from cryptography.hazmat.primitives.asymmetric.ed25519 import Ed25519PublicKey
```

```
CANONICAL_FIELDS = [
    "decision_point",
    "action_text",
    "decision",
    "effective_fidelity",
    "composite_fidelity",
    "boundary_triggered",
    "tool_name",
```

```

    "timestamp",
    "purpose_fidelity",
    "scope_fidelity",
    "boundary_violation",
    "tool_fidelity",
    "chain_continuity",
]

def canonicalize(receipt):
    payload = {field: receipt[field] for field in CANONICAL_FIELDS}
    return json.dumps(payload, sort_keys=True, separators=(",", ":")).encode("utf-8")

def verify(receipt, public_key_hex):
    expected_hash = hashlib.sha256(canonicalize(receipt)).digest()
    if expected_hash.hex() != receipt.get("payload_hash"):
        return False
    try:
        public_key = Ed25519PublicKey.from_public_bytes(bytes.fromhex(public_key_hex))
        public_key.verify(bytes.fromhex(receipt["ed25519_signature"]), expected_hash)
        return True
    except (InvalidSignature, ValueError, KeyError):
        return False

```

```

parser = argparse.ArgumentParser()
parser.add_argument("receipt_json")
parser.add_argument("public_key_hex")
parser.add_argument("--tamper", action="store_true")
args = parser.parse_args()

with open(args.receipt_json, "r", encoding="utf-8") as f:
    receipt = json.load(f)

if args.tamper:
    receipt = copy.deepcopy(receipt)
    receipt["scope_fidelity"] = 0.287
    print(f"one-field-tamper={verify(receipt, args.public_key_hex)}")
else:
    print(f"genuine={verify(receipt, args.public_key_hex)}")

```

Cold verification transcript:

```
$ python cold_verify_receipt.py receipt.json 71720f9e7a0bd1f2609b64ad5eef23afc730474bd468baedb096b8ce6a
genuine=True
```

```
$ python cold_verify_receipt.py receipt.json 71720f9e7a0bd1f2609b64ad5eef23afc730474bd468baedb096b8ce6a
one-field-tamper=False
```

```
$ python shipped_verify_receipt.py receipt.json 71720f9e7a0bd1f2609b64ad5eef23afc730474bd468baedb096b8ce6a
genuine=True
```

```
$ python shipped_verify_receipt.py receipt.json 71720f9e7a0bd1f2609b64ad5eef23afc730474bd468baedb096b8ce6a
one-field-tamper=False
```

Document version: 3.13 (pre-ratification) Release date: July 2026 (pending JB ratification) Status: Question-led restructure of v3.2 | Adversarial + agentic detection evidence re-sourced to primary artifacts | Observation register throughout | v3.5 fleet honesty and framing pass: detection-rate relabel, matched-FPR and self-scoring disclosures, calibration gap disclosed, stability math recast as descriptive, signature claim recast as analogy, accountability mapping corrected, regulatory self-disclosure added | v3.6 receipt exhibit embedded and cold-verified; Part IV title softened | v3.7 advisor-fleet surnames de-attributed; exhibit block frozen | v3.8 receipt exhibit re-generated from shipped 17-field runtime; III.2 prose reconciled to shipped fields; scores live-computed | v3.9: cross-vendor Tier-2 honest-scoping pass – reconciled prose to the shipped 17-field receipt (per-action verifiability vs chain-signing COMING; lens-config and Manifest/root binding scoped to the signing seam; provider-agnostic-by-design; documentation compiler status-tracked; deliberation-gate scoped to enforcement-integration); minor mappings and the unverified NAIC year resolved. | v3.10 lean pass (first, conservative): receipt exhibit relocated to Appendix D with a compact 13-field excerpt in III.2 (crypto values byte-identical, still cold-verifies); III.6 draft-process rows removed; III.7 reduced to a future-work note; Appendix A narrative deduplicated to provenance; Appendix B statutory/penalty density compressed. No claim, formula, TELOS evidence number, or caveat removed; preservation ledger accompanies. Deeper body compression to the 12.5-14k target is a reviewed second pass. | v3.11 directed lean pass (round 2): per-section compression to the ~14-15k target under an independent per-section claim-discipline verifier; five units repaired by restoring hedges and quoted regulatory language (converging-on, consistent-with, alone, a class of evidence, systematic-and-continuous, the three FDA controls, the NIST MEASURE demand), Close reverted to the reviewed v3.10 ending; front matter, references, and the frozen Appendix D exhibit byte-identical and re-cold-verified genuine=True/tamper=False; protected-elements gate 8/8 DOIs, 16/16 formulas, 23/23 caveats, 42/44 evidence numbers (two authorized penalty figures). 15,157 words. | v3.12 external citation-verification pass (2026-07-11, web-verified against primary sources): Gartner Kornutick Feb 17 2026 PR EXTERNALLY VERIFIED — all five figures + the “point-in-time audits” quote confirmed (closes the v3.11 Tier-1 pre-publish recommendation); Laban et al. reference corrected to the real title/id (“LLMs Get Lost In Multi-Turn Conversation,” arXiv:2505.06120) and the in-text drift gloss narrowed to what the literature shows; arXiv:2310.01427 author string corrected (Peysakhovich & Lerer — prior string was wrong); XSTest title spelling corrected (Behaviours); EU AI Act Art. 72 paraphrase-in-quotation-marks replaced with verbatim Art. 72(2) fragments; Art. 50 obligations list corrected to the article’s actual obligations (two prior items were not in Art. 50); NIST AI RMF quotes re-anchored to verbatim MEASURE 1/MEASURE 3 category text; FDA citations updated for the QMSR (21 CFR 820 amended effective Feb 2 2026; repealed §§820.70/.75/.90/.184 replaced with ISO 13485 clause mappings); SB 243 retitled to its official subject (Companion chatbots); Digital Omnibus status language updated to adoption-pending; seven never-cited references removed (cite-or-drop); Primacy Attractor bridging reduced to first definition only. Zenodo DOIs spot-checked publicly resolving. Appendix D exhibit and all evidence numbers, DOIs, formulas, and caveats byte-preserved. | v3.13 public-safety pass for the shared review repository (2026-07-14): removed the one domain-specific validation sentence that named a specific external company through a private validation path; genericized the Appendix D verification-transcript command lines (dropped absolute local filesystem paths). No claim, formula, evidence number, DOI, or caveat altered; Appendix D crypto values byte-identical and still cold-verify genuine=True / one-field-tamper=False. Next review: EU Digital Omnibus final adoption status (the Council general approach of 2026-03-13 has since advanced; refresh B.2 at the publish gate)

This whitepaper represents the current state of TELOS research and validation. Results are preliminary and subject to peer review. Implementation in production systems should follow appropriate testing and validation protocols.