TELOS Technical Method and Measurement Specification
Purpose Anchor formation, external reference measurement, agentic dimensions, and runtime observation
TELOS AI Labs Inc.
August 2026
TELOS Technical Method and Measurement Specification
1. Scope
This document preserves the complete mathematical method disclosed in the July 2026 TELOS whitepaper. It separates the method from the conceptual narrative, validation results, cryptographic receipt protocol, and regulatory crosswalk.
This specification carries the disclosed mathematical relationships of the method. Where a component is described but its implementation is held under NDA, that boundary is preserved rather than an implementation being invented to fill it. The disclosure boundary is stated explicitly in Section 13.
TELOS measures alignment to a human-declared purpose. It does not measure factual correctness, universal safety, legal compliance, or the wisdom of an action.
2. Method summary
The method has six stages:

- A human principal declares a behavioral specification.
- The approved specification is versioned as the Manifest.
- Manifest dimensions are encoded in an external embedding space to form the Purpose Anchor.
- Each action submitted by an integrated orchestration path is independently embedded.
- The fidelity engine measures the action against the fixed reference and configured boundary directions.
- The runtime records the scores, verdict, flags, tool metadata, and applicable cryptographic evidence.

The reference remains outside the governed model's context. It changes only when the authorized principal changes the specification or when a new scorer configuration requires recompilation and approval.
3. Terms and notation
Symbol or term	Meaning
Manifest	The human-readable, versioned, approved behavioral specification
Purpose Anchor	The mathematical reference compiled from the Manifest
	The fixed external reference embedding in the motivating conversational formalization
	The independent embedding of a response or action submitted at time
	The governed transformer's query representation at token or turn in the motivating illustration
	A prior key representation inside the governed transformer's context
	The attention weight assigned at time to position
	The stylized effective internal reference induced by accumulated context
	Measured departure at time
	Configured proportional factor for the response signal
	Graduated escalation magnitude produced from departure
	Fixed reference in the attractor illustration
	Basin centered on with configured radius
Fidelity	A configured measure of adherence to the declared specification
Boundary proximity	Evidence that an action approaches a declared prohibition or hard line
Departure	A runtime determination that a submitted action falls outside the configured Manifest conditions

4. Human authority and proportional measurement
The human principal declares the behavioral specification. The agent does not author its own governing reference. The topology is:

- Human authority declares and approves the behavioral specification.
- The governance layer measures each submitted action against the compiled reference.
- The governed model produces work under observation.
- Departures surface according to the human-declared escalation policy.

The proportional measurement signal is expressed as:

is a scalar escalation magnitude that increases with departure. In the current observation layer it is not a force applied to the model state. It therefore carries no control sign convention. The adapter may carry enforcement capability, but the current internal configuration is observation-first.
5. Manifest formation
5.1 Declared dimensions
A Manifest may include:

- Purpose
- Scope
- Boundaries
- Tool authorizations
- Sequence or chain expectations
- Escalation policy
- Human approval and version information maintained by the deployment

The current public-facing agentic instrument set describes five scoring dimensions: purpose, scope, boundary, tool, and chain. The opened runtime exposes per-dimension scores, composite and effective fidelity, decision values, flags, and optional aggregates.
ECI, Envelope Margin, compliance_rate, and CDR are not present in the opened implementation and are not claimed as runtime outputs.
5.2 Human-readable and mathematical forms
The Manifest is the human-readable and approved artifact. The Purpose Anchor is its mathematical form.
For the motivating example:

- Purpose: Help the principal structure a technical paper.
- Scope: Guide the principal's thinking without drafting the content.
- Boundary: Do not write complete paragraphs on the principal's behalf.

Each declaration is encoded as an embedding in using a sentence embedding model. The declarations together define the reference against which submitted work is measured.
5.3 Goals and prohibitions
Goals and prohibitions are not measured as interchangeable positive examples.

- Similarity toward a purpose can support evidence that an action concerns the declared assignment.
- Similarity toward a prohibition can indicate boundary proximity rather than good performance.

The engine therefore treats positive intent and prohibited direction differently. The composite formulation and tolerance mapping are disclosed in Sections 10.1 and 8.1. The specific boundary corpus contents and centroid construction remain held under NDA.
5.4 Formation quality
The measurement inherits the quality of the Manifest. A vague, conflicting, incomplete, or strategically permissive declaration creates a weak reference. TELOS measures the specification that the operator approved. It does not warrant that the specification is well formed.
For evidence review, the declaration and its approval context should accompany the scored result. A procedurally consistent score does not cure a defective mandate.
6. Why the reference is external
6.1 Transformer attention
Transformer attention uses scaled dot-product similarity. Following Vaswani et al.:

The term is a dot product between query and key representations. It measures directional similarity before normalization and softmax. See the PyTorch scaled dot-product attention documentation.
At token or turn , the model creates and computes a score against each prior key :

After softmax, those values become attention weights that influence generation.
6.2 Motivating example
Assume the principal declares at turn 1:
Provide guidance on structure, but do not write content directly.
Encode that declared reference as:

At turn 15, the attention weight for prior position is illustrated as:

If recent keys , , and have drifted from , the model can remain locally coherent with recent context while diverging from the declared reference:

while:

The similarity computation can be internally coherent even when the reference against which it is implicitly operating has moved.
6.3 Stylized effective-reference model
Let denote a stylized effective reference at time , represented as an attention-weighted average of prior key vectors:

This is a motivating simplification. Actual attention aggregates value vectors, not key vectors, and rotary positional embeddings apply position-dependent rotations rather than an explicit exponential decay envelope. The expression is not presented as a literal implementation of transformer attention.
To illustrate recency bias, let:

Over turns, the effective reference can move away from the original declaration:

The model can retain high local similarity while losing global fidelity:

The first can remain high because each step fits recent context. The second can fall because the sequence has moved away from the original declared purpose.
In limiting shorthand, the attention weight on the initial declaration can tend toward zero as the sequence grows:

This section draws on research concerning context position, recency, and long-turn degradation, including Liu et al., Laban et al., Peysakhovich and Lerer, Wu et al., and Yang et al.. The specific causal relationship between attention movement and TELOS fidelity remains an open empirical prediction.
7. External fidelity measurement
TELOS stores the declared reference outside the governed model and embeds each submitted response or action independently:

The independent model is an L2-normalized sentence transformer in the motivating implementation, following Sentence-BERT. Production runtime inference configuration and deployment-specific tuning remain held under NDA.
The external reference addresses persistence, not semantic perfection. Its advantages are:

- It does not compete with conversation tokens for attention.
- It does not move merely because context accumulates.
- It can be versioned and re-approved independently of the governed model.
- It supports repeated measurement across an instrumented action stream.

Its limitations are:

- It inherits the strengths and weaknesses of the embedding model.
- Cosine similarity is not itself a proof of entailment, authorization, truth, or safety.
- A changed embedding model changes the geometry and requires recompilation and recalibration.
- A weak Manifest produces a weak reference.
- The full production score cannot be reconstructed from the disclosed formula alone.

7.1 Instantaneous and mean fidelity
Using cosine similarity (Cover and Thomas, 2006), instantaneous and mean fidelity are:

where is instantaneous fidelity at turn , is mean fidelity over turns, is the response embedding, and is the declared purpose vector. Every turn produces quantified adherence, and the stream supports statistical tracking through mean, variance, and control limits, threshold-based surfacing, and audit evidence.
7.2 Primacy State, for conversational governance
For conversational governance, the Primacy State measures the degree to which user and system converge on the declared anchor:

where is anchor strength, meaning basin membership confidence, and are the two fidelities, and the harmonic mean ensures 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 composite measures of Section 10.1.
8. Control and attractor formalisms
8.1 Fixed reference and basin
The attractor illustration describes the fixed reference as with a basin of attraction:

The basin radius is calculated as:

where is the tolerance parameter, so that lower tolerance produces 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 , permissive, gives . At , strict, gives . Throughout this specification plays exactly this tolerance role and no other.
Departure is measured as normalized distance from the reference, and the response magnitude grows proportionally with it:

In the idealized actuated case the basin admits a Lyapunov function

whose derivative under proportional correction is

which is the formal statement of why a fixed reference gives stable meaning to distance. The limits of that statement for the shipped observation layer are set out immediately below.
8.2 Boundary of the analogy
Classical stability analysis applies to systems that actuate state back toward a reference. The current TELOS layer is observation-only and operates on discrete embeddings produced by an external model. It measures and surfaces departures. It does not itself drive the governed model state back toward the anchor.
No closed-loop stability guarantee is claimed for the current observation layer. Any Lyapunov argument applies only to an idealized actuated case. The contribution is not new control mathematics. It is the application of fixed-reference measurement discipline to runtime semantic governance.
9. Statistical process control discipline
9.1 Numeric action stream
Each submitted action produces a numeric measurement stream. That permits tracking of:

- Central tendency
- Variation
- Configured display bands
- Departure rates
- Drift vectors
- Baseline-relative sustained drift
- Optional session aggregates

Traditional Statistical Process Control often uses limits such as:

TELOS uses configured tolerance bands around the Purpose Anchor. It does not claim that all semantic similarity distributions are normal or that classical three-sigma limits are automatically valid for cosine scores.

Cosine similarity is bounded on and may be non-normal, especially early in a session. Operative verdict thresholds are configured rather than derived from each observed session.

9.2 EWMA tracking

The opened response manager includes a baseline-relative exponentially weighted moving average tracker after baseline collection. The method does not disclose the smoothing constant, baseline collection rule, sustained-drift threshold, or full continuity formulation. Those values are deployment-calibrated and held under NDA unless a particular benchmark artifact discloses them.

Severity is described as a relative decline from the collected baseline under configured thresholds, not as a universal absolute window-average scale.

9.3 Purpose capability

The method carries the process-capability idea into semantic monitoring. From process capability analysis (Montgomery, 2020):

with the maximum acceptable drift, which is the escalation trigger, the minimum required fidelity, which is the basin threshold, and and the session fidelity mean and standard deviation.

Interpretation under the stated calibration band: indicates a highly capable observed process; indicates a capable process to monitor; indicates a process that is not capable, where departures will surface. The limits derive from calibration against session data. The specific calibration artifacts accompany deployments rather than this specification.

This hands regulators a familiar quality metric applied to AI governance.

10. Runtime decision model

10.1 The agentic anchor and composite score

For agentic deployments the reference is not a single vector but a tuple:

Each proposed action submitted to the engine receives five weighted decision terms: purpose, scope, tool, chain continuity, and boundary penalty.

where each dimension fidelity . The boundary term is a **penalty**. The anchor encodes desirable behaviour, while the boundary corpus encodes prohibitions, so higher similarity means closer to a violation and boundary similarity subtracts. The boundary corpus is built in three layers, for both semantic coverage and regulatory traceability: hand-crafted boundary phrasings, generated gap-fillers for what the hand-crafted set misses, and regulatory extractions with provenance chains to specific regulatory text.

Operational continuity within a submitted work sequence is measured separately by the Semantic Continuity Index:

which tracks step-to-step coherence. When , chain context inherits from the previous step, giving 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.

10.2 Graduated outcomes

The operational frame uses two principal outcomes plus escalation:

- Within the Manifest
- Departure from the Manifest
- Human surfacing or configured handling based on severity and policy

The verdict vocabulary is execute, clarify, and escalate.

- **execute** records an action assessed as within the declared specification.
- **clarify** marks proximity to a boundary, or a dimension the measurement cannot resolve confidently, for resolution on the agent path before the action proceeds. It is not a human review state.
- **escalate** is the human path. It routes the matter to the accountable person named in the escalation policy.

An optional micro-judge can be enabled for unresolved cases. It is off by default and does not judge boundary proximity.

A recorded verdict is a classification, not a disposition. Under the current gateMode: observe and failPolicy: open configuration, a recorded clarify or escalate does not by itself establish that the action was blocked or delayed, that clarification took place, or that a human received or resolved anything. Evidence of those outcomes requires the adapter event record, recipient, response, and timing.

10.3 Observation and enforcement

An adapter can carry enforcement capability. Deployment configuration decides whether a result is observed or enforced.

The current internal configuration is:

gateMode: observe

failPolicy: open

Under this configuration, the action proceeds. The system records the measurement and may surface it. The current evidence therefore supports observation claims rather than universal prevention claims.

10.4 Configuration evidence

The active threshold and optional-component configuration is not included in the current per-action receipt. A reviewer needs separately preserved configuration evidence to determine:

- Which embedding model and version were active
- Which Manifest and Purpose Anchor were in force
- Which thresholds and weights were applied
- Whether optional classifiers or micro-judges were enabled
- Whether the adapter was observing or enforcing
- Which failure policy applied

This missing binding is a current evidence limitation, not a change to the scoring method.

11. Architectural positioning

11.1 Middleware interface

The opened source provides a framework-independent governance protocol and adapters that an application can place between orchestration logic and a model or tool layer.
The interface has six stated properties:

- It scores submitted text and tool metadata without retraining the governed model.
- Measurement occurs when an adapter invokes the protocol.
- The interface is not coupled to a single model provider.
- TELOS-defined local record formats are independent of model-provider response formats.
- The records support documentation and review but do not by themselves establish compliance.
- Coverage begins only after an orchestration path is explicitly integrated.

11.2 Provider independence
Provider independence is an interface property. Empirical equivalence across models and embedding backends remains unestablished. The July AgentHarm scorer-profile difference, reported in the validation companion, demonstrates why the distinction matters.
11.3 Documentation-to-Manifest compilation
The planned documentation compiler would derive a concise, versioned Manifest from organizational policies, procedures, and handbooks. The corpus would remain distinct from the approved specification.
That pipeline is COMING. It is not shipped, and its implementation remains held. The current method begins from a manually formed or otherwise supplied behavioral specification.
12. Open empirical predictions
Three falsifiable predictions follow from the method:
P1. Fidelity loss should correlate with attention-weight movement toward recent context.
P2. Manipulations that increase attention on the original constraint should reduce fidelity loss even without TELOS measurement.
P3. Models with weaker recency bias should retain better baseline fidelity.
These predictions remain unevaluated. They motivate research but are not evidence supporting the current product.
12A. Implementation, configuration, and validation state
A single status label cannot carry this material, because three facts vary independently: whether a mechanism is **implemented**, whether it is **active in the configuration described by this specification**, and whether its behaviour has been **validated**. A mechanism can be built and inactive. A mechanism can be active and unvalidated. The table below therefore separates them.

Mechanism	Implementation	Current configuration	Validation
Per-action measurement against the Purpose Anchor	Implemented	Active on instrumented paths	Validated only for the cited tests and scorer profiles
Verdict emission (execute / clarify / escalate)	Implemented	Active	Emission of a verdict is not evidence of the behaviour the verdict names
Agent-path clarification behaviour	Designed	Not active in the evaluated posture	Not validated across live tolerance settings
Adapter enforcement branch	Present in the adapter interface; deployment configuration selects observation or enforcement	Inactive under gateMode: observe and failPolicy: open	No universal prevention claim
Documentation-to-Manifest compilation	Designed	Not built	Not validated
Full improvement cycle (section 12B)	Designed, partial	Not the current operating mode	Not validated

Two conventions govern this table. **Designed** means specified but not demonstrated by executable evidence. **Inactive** means a code path exists but the current configuration does not select it; that is a different state from unbuilt, and it is why a single roadmap label is insufficient.
12B. Designed improvement cycle and current implementation status
The quality-systems improvement cycle is a useful frame for how a governed deployment is intended to improve over time. Its phases do not share a single state, so each is given its own.

Phase	What it means here	State
Define	The principal approves the operational specification and the tolerance configuration before deployment	Current practice, performed through manual approval and configuration. The documentation-to-Manifest compiler is designed and not built
Measure	Per-action measurement of submitted actions against the fixed reference	Current, on instrumented paths
Analyze	Recorded per-dimension scores, flags, and baseline-relative trends	Current. This is recorded variation, not demonstrated causal root-cause analysis
Improve	Designed clarify behaviour, resolving proximity or ambiguity on the agent path before an action proceeds	Designed. Not active in the evaluated posture, and live behaviour across varied tolerance settings is not validated
Control	Operator-owned tolerance configuration, escalation policy, versioning, and the adapter enforcement branch where a deployment selects it	Configuration and policy are current. Enforcement is inactive under observe / open. No closed-loop convergence is claimed

This cycle does not execute per turn in the evaluated configuration. Under observe and open the action proceeds while the result is recorded and may be surfaced. Nothing in this section transfers the idealized actuated result of Section 8 to the designed cycle; that result applies to an idealized actuated case and carries no evidentiary force for a discrete agent reasoning loop until its actuator, feedback timing, and measured outcome are specified and tested.
A designed cycle cannot govern an action it never receives. Coverage extends only to actions submitted on integrated paths.

Configuration evidence limits apply. As Section 10.4 states, the active thresholds and optional components are not bound into the current per-action receipt, so a receipt alone does not establish which tolerance configuration, verdict thresholds, micro-judge setting, or enforcement mode was in force.

13. Disclosure boundary

13.1 Disclosed

- External fixed-reference architecture
- Manifest and Purpose Anchor distinction
- Cosine fidelity relationship
- Complete motivating attention and drift formalization
- Proportional escalation relationship
- Attractor and basin interpretation
- Five named agentic dimensions
- Goal versus prohibition distinction
- Baseline-relative EWMA presence
- Observation and enforcement configuration distinction
- Runtime output categories
- Public benchmark operating points where published

13.2 Held under NDA

- Production embedding configuration
- Boundary corpus
- Centroid construction
- Manifest compilation pipeline
- Deployment-specific weights and thresholds not published with benchmarks
- EWMA constants and detailed baseline logic
- Scoring subsystem internals
- Signing implementation
- Validation runners, protocol details, and specific scoring configuration

The disclosed method explains the architecture and motivating mathematics. It does not permit an outside party to reconstruct the calibrated production engine.

14. Method boundaries

The method should be read with five standing boundaries:

- **Construct validity.** Similarity to a declared reference is not automatically equivalent to authorization or compliance.
- **Configuration dependence.** Results depend on the embedding model, corpus, centroids, weights, thresholds, and optional components.
- **Instrumentation dependence.** Only submitted actions are measured.
- **Specification dependence.** A poor Manifest creates a poor reference.
- **Evidence dependence.** A signed score proves integrity of covered fields, not validity of the scorer.

15. Technical references

- Astrom, K. J., and Murray, R. M. Feedback Systems. Princeton University Press, 2008.
- Bovens, M. Analysing and Assessing Accountability. European Law Journal, 2007.
- Cover, T. M., and Thomas, J. A. Elements of Information Theory. 2nd edition, 2006.
- Gu, X., et al. When Attention Sink Emerges in Language Models. 2024.
- Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. PNAS, 1982.
- Jensen, M. C., and Meckling, W. H. Theory of the Firm. Journal of Financial Economics, 1976.
- Khalil, H. K. Nonlinear Systems. 3rd edition, 2002.
- Laban, P., et al. LLMs Get Lost in Multi-Turn Conversation. 2025.
- Liu, N., et al. Lost in the Middle. 2024.
- Montgomery, D. C. Introduction to Statistical Quality Control. 8th edition, 2020.
- Murdock, B. B. The serial position effect of free recall. Journal of Experimental Psychology, 1962.
- Ogata, K. Modern Control Engineering. 5th edition, 2009.
- Ostrom, E. Governing the Commons. Cambridge University Press, 1990.
- Peysakhovich, A., and Lerer, A. Attention Sorting Combats Recency Bias. 2023.
- Reimers, N., and Gurevych, I. Sentence-BERT. EMNLP, 2019.
- Rottger, P., et al. XSTest. NAACL, 2024.
- Shewhart, W. A. Economic Control of Quality of Manufactured Product. 1931.
- Sieyes, E. J. Political Writings. Original constituent-power essay published 1789.
- Strogatz, S. H. Nonlinear Dynamics and Chaos. 2nd edition, 2014.
- Vaswani, A., et al. Attention Is All You Need. NeurIPS, 2017.
- Wheeler, D. J. Understanding Variation. 2nd edition, 2000.
- Wu, X., et al. On the Emergence of Position Bias in Transformers. 2025.
- Yang, B., et al. RoPE to NoPE and Back Again. 2025.