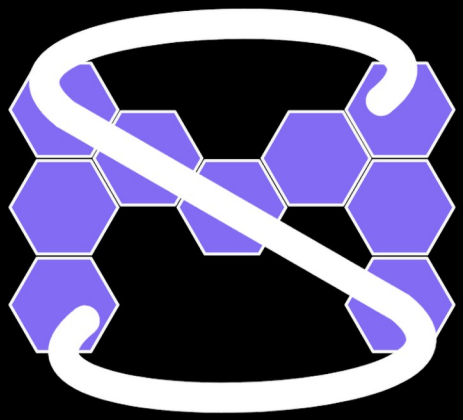

INFERENCE

TRACEABILITY

A METROLOGICAL ARCHITECTURE FOR CONSEQUENTIAL AI OUTPUTS,
AND A REFERENCE IMPLEMENTATION ON STANDARD BY METRASLATE

CAMERON BUNETA · METROLOGIST AND FOUNDER, METRASLATE LLC
SUBMITTED AS A PUBLIC COMMENT ON NIST AI 200-2 · SEPTEMBER 2026
PATENT PENDING · U.S. PROVISIONAL APPLICATION NO. 64/149,811

BY METRASLATE

Inference Traceability: A Metrological Architecture for Consequential AI Outputs, and a Reference Implementation on Standard by MetraSlate

Making the validity state of an AI system traceable to each output, determining downstream impact when that state changes, and building the record that does it

Cameron Buneta, Metrologist and Founder, MetraSlate LLC

Draft for public comment on NIST AI 200-2 ipd · September 2026 · Patent pending — U.S. Provisional Application No. 64/149,811

Abstract

Test, evaluation, verification, and validation (TEVV) of AI systems, as framed in the NIST AI Risk Management Framework and the initial public draft of NIST AI 200-2, produces evidence about a system at the time of an assessment. Operational AI systems produce consequential outputs continuously, between assessments, on inputs the assessment did not cover, under models that change. Nothing in current practice binds each consequential output to the validity state of the system at the moment it was produced, and nothing determines which outputs, decisions, and consequences are implicated when that state is later found to have changed. This paper describes an architecture that does both, by treating each consequential inference as the output of a measuring system in the sense of the International Vocabulary of Metrology and applying the disciplines calibration laboratories already use under accreditation: a calibration record bound to every output; stratified reference cases re-run as control charts; a validity domain enforced as a gate that refuses rather than discounts; a stated uncertainty budget and a guard-banded decision rule; shadow comparison before a model change is promoted; and, when an instrument is found out of tolerance, an impact assessment that traces every affected output forward through the human decisions and consequences that rested on it. The chain preserved is: reference case → model, version, configuration → validity domain → input conditions → output → uncertainty → human presentation → decision → downstream consequence.

The paper then describes the platform on which the architecture is implemented. Standard by MetraSlate is a domain-agnostic, multi-tenant record and operations platform that ships as editions for industries whose records must be defensible. Its first edition served ISO/IEC 17025 calibration laboratories, and it was calibration — where a record must be reproducible, uncertainty-qualified, decided by rule, and traceable to references for the life of an instrument — that proved the platform's primitives: an append-only record spine, hash-verified document reproduction with deterministic vocabulary projection, an exact-decimal value path, an uncertainty engine per JCGM 100, ILAC-G8 decision rules with guard bands, and database-enforced fail-closed tenancy. These are platform substrate, not calibration features, and they are the primitives inference traceability requires. The architecture's elements are mapped to those primitives with their maturity stated, and the method by which the platform is built and by which its public-safety edition was specified — Metra Praxis, in which the domain expert owns the truth in her own words and every claim is tagged by how it is known — is described as the same discipline applied to the building of the system. MetraSlate proposes to provide a reference implementation of the architecture as a Metrology Block for the TEVV-Athlon Framework, an open specification of its record classes, and a worked TEVV-Athlon on a machine-translation chain in emergency communications.

1. Introduction

Measurement science has been named, explicitly and repeatedly, as the discipline AI evaluation needs. NIST's AI Risk Management Framework places a Measure function at the center of its core [1]. The Center for AI Standards and Innovation wrote in December 2025 that many evaluations "do not precisely articulate what has been measured, much less whether the measurements are valid," and that metrology "provides methods for rigorous and trustworthy use of measured values, qualified with assessments of uncertainty" [2]. In August 2026 NIST released the initial public draft of NIST AI 200-2, the TEVV-Athlon Framework, whose second stage is named Define & Construct and whose unit of construction is the *Metrology Block* [3]. The NIST AI Consortium, with more than 280 member organizations, describes its work as "laying a foundation for global AI metrology" [4].

This paper therefore does not argue that AI needs measurement science. It argues something narrower and, we believe, operationally missing, and it presents a platform that supplies it. TEVV as currently framed is an *assessment*: an athlon is an event, conducted at a point in the lifecycle, producing evidence about the system as it stood. AI 200-2 is careful to say that measurement validation "is an ongoing process rather than a one-time activity" and lists, among measurement-science considerations, periodic calibration of evaluation processes against baselines or ground truth [3, §4.5, App. D]. But between assessments a deployed system produces consequential outputs continuously — a translation a call-taker acts on, a triage recommendation a clinician sees, a risk score a lender uses — on inputs the assessment did not cover, under models that vendors update, in conditions the assessment did not stratify. Two things are absent from current practice:

1. **Binding.** Nothing ties each consequential output to the validity state of the system at the moment the output was produced — which model and configuration, validated for which conditions, on the strength of which reference evidence, with what uncertainty, and whether the input was inside the validated domain at all.
2. **Reverse traceability.** When a later assessment finds that the system's validity state had changed — the model drifted, a reference stratum fell out of control, a vendor update shifted outputs for one population — nothing determines which outputs were produced under the invalid state, which human decisions rested on those outputs, and which consequences followed.

Calibration laboratories solved both problems for physical instruments long ago, and the solutions are codified in the standards laboratories are accredited against. Every measurement result is traceable to a calibration record for the instrument that produced it; instruments are verified between calibrations with check standards and control charts; a result is reported with a stated uncertainty; conformity is decided by a rule with a guard band; and when an instrument is found out of tolerance, the laboratory performs an impact analysis on every result issued since the last good calibration and notifies affected customers [5, 6, 7]. The claim of this paper is that these disciplines transfer to consequential AI inference without modification of their logic, and that an append-only, lineage-preserving record makes the transfer operational.

The proposition in one sentence: **an architecture for making the validity state of an AI system traceable to each consequential output, and for determining the downstream impact when that validity state later changes.** The second proposition is that the record which does this already exists — as a domain-agnostic platform that MetraSlate has built and proved first under the discipline of accredited calibration — and that MetraSlate intends Standard to be its reference implementation.

1.1 Contributions

1. A record model in which every consequential inference is bound to a **calibration record** for the instrument that produced it, and in which the chain *reference case → instrument → validity domain → input conditions*

→ *output* → *uncertainty* → *presentation* → *decision* → *consequence* is preserved as linked, append-only records (Section 3).

2. **Continuous verification** of deployed instruments by stratified reference cases re-run as control charts, including a separate reference set for the input-conditioning layer so that a change in the signal path is detected before it is attributed to the model or to the population (Section 3.3).
3. A **validity gate** that refuses to produce an inference for any stratum outside the instrument's validated domain or out of control, writing the refusal as a record with reason codes rather than an output with a lowered confidence (Section 3.6).
4. **Out-of-tolerance impact assessment** as a graph traversal: when the instrument is found outside its validated state for an interval and a stratum, every output produced under that state is marked suspect and every decision and consequence that rested on it is enumerated (Section 3.5).
5. A **longitudinal record of the instrument** that separates population change from measurement change, and a **disparity chain** that locates where a group difference enters before any of it is called bias (Sections 3.7, 3.8).
6. A **reference implementation** on a domain-agnostic platform whose primitives were proved first under accredited calibration, with the maturity of each element stated, and a method for specifying such systems in which the domain expert owns the truth (Sections 6, 7).
7. A mapping onto NIST AI 200-2's Blocks, Events, and Tools, proposed terminology from the VIM, and responses to specific open questions posed by CAISI (Section 8).

1.2 Scope and the word "calibration"

The architecture applies to any AI output on which a human or automated decision with consequences rests. It is illustrated with speech and language systems in emergency communications, where the requirements were developed with a serving 9-1-1 center director, and it applies unchanged to clinical, financial, and safety-critical settings.

The word *calibration* has two established meanings that collide here. In machine learning it denotes agreement between predicted probabilities and observed frequencies [8]. In metrology it denotes the operation that establishes the relation between values indicated by an instrument and values provided by reference standards, with their uncertainties [9, def. 2.39]. This paper uses the metrological sense throughout and says *probability calibration* when it means the other.

1.3 Who is proposing this

The author is a practicing metrologist in commercial calibration and the founder of MetraSlate LLC, whose product, Standard, is a domain-agnostic, multi-tenant record and operations platform that ships as editions — a shared substrate of records, tenancy, documents, operations, and numerics, with each industry's subjects, vocabulary, numbering, and document templates declared in an edition manifest. It is not calibration software. Its first edition, for calibration laboratories accredited to ISO/IEC 17025, was the proving ground, chosen because it is the hardest record discipline the author knows: the product is a sealed certificate that a customer relies on and an assessor audits, and the instruments are calibrated against references, monitored between calibrations, and recalled when they drift. Its second edition, for emergency communications centers, was specified in three rounds of adversarial review with the Executive Director of Aurora911, whose questions — what was knowable, by whom, from what source, at what time; what the system refuses to know; and how the system would detect the introduction or development of bias in its own inferences — produced the requirements this architecture answers. The platform is named for what a calibration laboratory measures against. The intent is that it become the

reference implementation of this discipline: the record format others implement, and the system that proves it in service.

2. Background and prior work

2.1 NIST's AI measurement work

The AI RMF's Measure function calls for "appropriate methods and metrics" and for TEVV throughout the lifecycle [1]. AI 200-2 supplies a four-stage method for constructing an assessment: Articulate & Organize, Define & Construct (Metrology Blocks), Apply & Measure (Events and a Toolbox), and Synthesize & Interrogate [3]. Its Appendix D lists measurement-science considerations including calibration of evaluation processes, measurement of uncertainty, use of controls or baselines, and validity; its Appendix E lists experimental-design considerations including comparison baselines and repetition. The draft's example TEVV-Athlon is drawn from the ARIA pilot evaluation [10]. CAISI's companion draft addresses the reliability of automated benchmark evaluations [11]. NIST SP 1270 treats bias as a socio-technical phenomenon to be managed across the lifecycle [12], and the Generative AI Profile recommends documenting statistical variance and bias in metrics and establishing construct validity for complex constructs [13].

CAISI's December 2025 statement of open questions is the most direct invitation [2]: how to identify, quantify, and communicate sources of uncertainty (II-A); how well pre-deployment evaluations predict post-deployment behavior (I-B.3); best practices for reporting (II-D); how and when AI systems should be used to test, evaluate, or monitor AI systems (I-D.2); reproducible methods to measure AI-driven changes in downstream outcomes over time (III-A.1); methods to measure the causal effect of safeguards and interventions on downstream outcomes (III-A.2); methods to "measure, categorize, and track components of complex, multi-turn human-AI interaction workflows" (III-A.3); and best practices for including stakeholders, including end users and subject-matter experts, in assessment (III-B). Sections 3, 7, and 8 address each.

What the NIST corpus does not yet contain, on our reading, is a mechanism that binds the evaluation state to each operational output, or a procedure for determining impact when that state changes. AI 200-2 is, by its own description, an assessment framework; the Operate & Monitor lifecycle stage appears in its figures, but the draft's methods for that stage — post-deployment feedback, field pilots, incident analysis [3, App. C] — observe outcomes rather than trace them to instrument state.

2.2 Metrology

The International Vocabulary of Metrology defines the concepts this paper borrows [9]: *measuring instrument* (3.1) and *measuring system* (3.2); *calibration* (2.39); *metrological traceability* (2.41), the property of a result whereby it can be related to a reference through a documented unbroken chain of calibrations, each contributing to the uncertainty; *measurement uncertainty* (2.26); *verification* (2.44), provision of objective evidence that a given item fulfils specified requirements; *validation* (2.45), verification where the specified requirements are adequate for an intended use; *instrumental drift* (4.21); and *maximum permissible error* (4.26). The Guide to the Expression of Uncertainty in Measurement supplies the method for combining uncertainty components and expressing an expanded uncertainty with a coverage factor [14]; AI 200-2 already cites it.

Laboratory practice adds what the vocabulary describes. ISO/IEC 17025:2017 requires a laboratory to monitor the validity of its results — by reference materials, functional checks, intermediate checks, control charts, and replicate measurements — and, for nonconforming work, to evaluate its significance "including an impact analysis on previous results" [5]. ANSI/NCSL Z540.3 requires evaluation of the impact of out-of-tolerance

conditions on prior measurements and notification of affected parties [6]. ILAC-G8 specifies decision rules for statements of conformity, including guard bands so that a result near a limit is not declared conforming or nonconforming beyond what its uncertainty supports [7]. ISO 10012 frames measurement management as a system with metrological confirmation of equipment [15]. National metrology institutes have begun to state the same program for AI: that trustworthy predictions require reliable quantitative assessment of uncertainty, and that traceability applies to AI outputs as to any measurement [16, 17].

2.3 Machine-learning monitoring and governance

Much of the machinery used here exists in some form in machine-learning practice. Model cards, datasheets, and factsheets document a model's intended use, evaluation data, and limitations [18, 19, 20]. Experiment-tracking and model-registry tools record versions and training lineage. Concept-drift detection is a mature literature [21]. Conformal prediction provides distribution-free prediction sets with coverage guarantees [22, 23]. Shadow and canary deployment are standard release practices. The U.S. Food and Drug Administration's guidance on predetermined change control plans requires that anticipated model modifications be pre-specified with a protocol for verifying and validating them [24]. The EU AI Act requires automatic logging over the lifetime of high-risk systems and post-market monitoring [25]. ISO/IEC 42001 specifies an AI management system, and ISO/IEC TR 24029, TS 4213, and 5259 address robustness assessment, classification performance, and data quality [26, 27, 28, 29]. W3C PROV provides a data model for provenance graphs [30].

Three things are missing from this body of work, and they are the three this paper supplies. First, none of it binds a *per-output* record to the instrument's validity state; documentation describes the model, drift detectors raise alarms, registries record versions, but no artifact says, for a given output, "produced by instrument θ under calibration record k, input in stratum s whose status was *validated*, with expanded uncertainty U." Second, none of it carries lineage forward through the human presentation and decision to the consequence, so that an out-of-tolerance finding can be turned into a list of affected decisions; provenance graphs can represent such a chain, but no monitoring practice populates one from the human side. Third, none of it treats refusal as a first-class outcome of a validity gate rather than as a low-confidence output. The novelty claimed is the composition, its grounding in accreditation disciplines that make it auditable, and its implementation on a record that already carries those disciplines.

3. The architecture

3.1 Definitions

- **Instrument.** A specific model version with a specific configuration and a specific input-conditioning pipeline (feature extraction, preprocessing, prompt template), identified by content hash. An instrument is a *measuring system* [9, 3.2]. A change to any component produces a new instrument.
- **Calibration record.** The record establishing an instrument's relation to references: the reference set it was evaluated against, per stratum; the date and result; the validity domain; the uncertainty budget; and the promotion decision that placed it in service, with the comparison report that supported the decision. Every calibration record has an *as-found* state and an *as-left* state [31].
- **Reference set.** A curated, versioned corpus of *reference cases*, each with a known expected output and the uncertainty of that expectation — the AI analogue of certified reference materials and check standards — *stratified* across the conditions the instrument will encounter, so that a change in one stratum is visible as its own line. Reference cases are sourced only where a lawful basis exists.

- **Validity domain and validity status.** The set of strata for which the calibration record holds reference evidence; each stratum is *validated* (current evidence within tolerance), *not validated* (no evidence), or *suspended* (a reference check failed, or an investigation is open).
- **Uncertainty budget.** The components combined into an expanded uncertainty for an inference: repeatability of input conditioning on reference re-runs; reproducibility of the model across configurations or seeds; input-quality contribution; validation coverage for the stratum; and reference-label uncertainty (for human-labeled references, inter-rater disagreement, with the panel's composition recorded). Combined per the GUM at a stated coverage factor [14].
- **Decision rule.** The rule by which an inference produces a categorical result, applied with a *guard band*: a result flags only when the measured deviation exceeds the threshold by more than the expanded uncertainty; inside the guard band it is *indeterminate* and recorded as such [7].
- **Inference record.** The record of one consequential output: input reference (a pointer, not necessarily a copy), measured input conditions (quality metrics, stratum), instrument, calibration record in force, validity status of the stratum at the time, output, expanded uncertainty, timestamp.
- **Presentation record.** That an output was shown to a principal: which, when, in what surface, at what *capture grade* — *acknowledged*, *rendered*, or *available*.
- **Decision record.** A human or automated action with a *basis* (the presented outputs and other information it rested on, at a stated basis grade), an *authority* (who or what had standing to decide), and a timestamp.
- **Consequence record.** A downstream event that a decision record references as basis, transitively.
- **Refusal record.** That the validity gate declined to produce an inference, with reason codes.
- **Evidence grade.** Every record carries one of *verified*, *received*, *derived*, *asserted*, or *absent* (with reason). Absent is a value.

3.2 The traceability chain

Link	Record	What it fixes in the chain
1	Reference case (stratum s)	What "correct" meant, with its own uncertainty
2	Instrument (model, version, configuration, conditioning)	What produced the output
3	Calibration record	The instrument's validity domain and status at the time, and the evidence for it
4	Input-condition measurement	Which stratum the input fell in, and its quality
5	Inference record	The output, bound to 2–4, with expanded uncertainty
6	Uncertainty budget	Where the interval came from
7	Presentation record	Whether, when, and how a person saw the output
8	Decision record	What was decided, on what basis, with what authority
9	Consequence record	What followed

Figure 1. The traceability chain as linked, append-only record classes. Lineage runs in both directions.

Two properties do the work. **Forward closure**: for any instrument and interval, the inference records it produced are a query, and the decisions and consequences reachable from them through basis edges are the transitive closure. **Backward closure**: for any consequence, the instruments, calibration records, and reference cases it rests on are a query. The first makes impact assessment a traversal; the second makes every consequential output metrologically traceable in the VIM's sense [9, 2.41].

3.3 Continuous verification

Between calibrations an instrument is verified by re-running the reference set on a schedule and on every deployment, per stratum, charted as control charts [32]. A stratum whose chart goes out of control is an out-of-tolerance finding for that stratum. Two refinements matter. **The conditioning layer has its own reference set**: for a speech instrument, the audio path — microphone, console, recorder, codec — is part of the measuring system, and a set of synthetic and recorded test signals passed through the actual path on the same schedule detects a headset replacement or a codec change before it appears as a change in the model or in the people measured. **Reference sets are subject to Goodhart's law** [3, §4.6]: a sealed subset never disclosed to model developers and rotated on a schedule, and a fence that no training pipeline can read it, are the mitigations.

3.4 Change control: shadow comparison and promotion

A candidate instrument runs in shadow — the same live inputs as the incumbent, its inference records stored, marked *shadow*, never presented — for a validation period. The promotion gate compares per-stratum reference agreement, output rates on the same live population, per-group rate change, and the distribution of uncertainty intervals, each against a tolerance set in advance. A candidate outside tolerance does not promote until a named person records why. Promotion is a recorded decision with the comparison report as its basis; the promoted instrument's first calibration record is its as-left state [24].

3.5 Out of tolerance: impact assessment by reverse traceability

When a reference check fails for a stratum, or a challenge to a specific output is upheld, the instrument is out of tolerance for that stratum from the last passing check to the failing one, and the response is the one ISO/IEC 17025 requires [5, §7.10]: every inference record produced under it in the interval and stratum is marked *suspect* by a new record; every decision whose basis includes a suspect inference, and every consequence reachable from it, is enumerated by forward traversal; every downstream signal derived from a suspect inference is *withdrawn* by a withdrawal record, the signal retained; every principal who received one is notified with the reason and the interval; and the stratum's status becomes *suspended* until the cause is found and a new calibration record is issued. The output of the traversal is the *impact set*, and it is a query rather than an investigation.

3.6 The validity gate and refusal

Before an instrument produces an inference, a gate evaluates the input's stratum against the calibration record. An inference is produced only when the stratum is *validated* and in control. Otherwise the gate writes a refusal record — stratum not validated, stratum suspended, input quality below the conditioning floor, conditioning reference out of control, instrument suspended — and no output is produced. A lowered confidence on an input for a condition the instrument was never validated for is not a measurement with larger uncertainty; it is an indication from an instrument used outside its specified range, and a laboratory would not report it. The gate refuses, and the refusal is a record graded *absent with reason*.

The enforcement point. Binding and the gate hold only for inferences that pass through the recording layer. In practice that layer is an inference gateway between the application and the model — whether the model is a

vendor API, a hosted service, or a local deployment — or it is the vendor's own emission of conformant records. A deployment must make the recording layer the only path to the instrument, or reconcile the record against the vendor's usage logs on a schedule; a *coverage check* reports any inference traffic the record did not see, and unrecorded traffic is itself a finding. The record does not claim completeness it cannot verify: the coverage check's result is part of every calibration record and every impact assessment.

3.7 The longitudinal record of the instrument

Because inference records are append-only and bound to their instrument, the instrument acquires a longitudinal record. When a new instrument is promoted it is also run over a held historical validation set spanning the prior instrument's era, and both outputs are stored, so that what changed in the population (each instrument on its own era) and what changed in the measurement of the population (both instruments on the same evidence) are two queries.

Apparent change	Where it shows	Where it does not
Instrument drift	Reference cases change output after a version or configuration change	The raw conditioned inputs; the conditioning reference set
Population change	The measured population's inference records move	Reference cases stable; conditioning reference stable; environmental indicators quiet
Input or conditioning drift	The conditioning reference set moves; input-quality metrics shift; a device or codec change is on record	Reference cases under clean conditions stable
Environmental change	Contextual indicators move (workload, staffing, integration health)	Reference cases stable; exposure-adjusted comparison flat
Introduced bias	A disparity appears at a named link of the chain for a comparison group (3.8)	Per-stratum reference results stable

Table 1. Distinguishing sources of apparent change by the reference layer that moved. Attribution is a lookup, because every inference record carries the reference checks in force when it was produced.

3.8 Disparity across the chain

A group difference can enter at any of seven links, each a record class the architecture holds: source data (input quality by group); conditioning (conditioning-reference results by the group's strata); inference (reference results by stratum, uncertainty width by group); threshold (deviation distributions by group at the guard band); flag (output rate per group at equal exposure); human response (response type per group at equal output, from decision records); outcome (later consequences per group after equal outputs). A difference at any link is reported as a *disparity* with its magnitude and the first link at which the groups diverge. *Bias* is a finding a human records after the investigation a disparity opens; the record never labels a demographic difference bias on its own. Because the human response is a record, an implementation in which identical outputs are answered differently for different groups appears as a disparity at link six with links one through five flat, and is distinguished from a model finding because the chain distinguishes it. Disparity series are retained and intersected with calibration records and organizational event records, so the first window in which a disparity appears is a query.

3.9 Immutability, commitments, and lawful destruction

Every record is stored as an envelope (identifiers, times, instrument, class, and a commitment) and a payload (content-addressed, encrypted under a data key). The commitment is a hash over the payload and a random nonce stored only with the payload; a hash chain over the envelopes seals the record. Destruction deletes the payload and nonce and destroys the key; the retained commitment proves that a record existed with that identifier at that time and, without the nonce, cannot be brute-forced against guesses of low-entropy content. Because every derived record carries basis references, destruction of an input is a reverse traversal over everything computed from it. The instrument's longitudinal record survives lawful destruction of individual payloads because the envelopes and commitments remain.

3.10 What the record refuses to hold

A traceability record is not a warrant to possess everything. Each data element is assigned one of four decisions — *ingest* the payload, *reference* it by pointer to its authoritative custodian, *derive* it from what is held, or *never touch* it — in that order of tests, in a versioned register the operating organization owns. The inference record holds an input reference, not necessarily a copy; the audio behind a transcript stays in the recorder, and only conditioned features come back. A continuous minimization check reports any held payload whose justification has lapsed. The chain is complete with pointers.

4. Formalization

Let an instrument be $\theta = (m, v, c, p)$: model m, version v, configuration c, conditioning pipeline p, identified by hash. Let S be the set of strata. A calibration record $K(\theta)$ assigns to each $s \in S$ a validity status $V(\theta, s) \in \{\text{validated, not validated, suspended}\}$, a reference result $R(\theta, s)$ with its control-chart state, and an uncertainty budget from which an expanded uncertainty $U(\theta, s, q)$ is computed for an input of measured quality q.

An input x arrives with measured conditions $(s(x), q(x))$. The validity gate G produces an inference record only if

$$V(\theta, s(x)) = \text{validated} \;\text{ and }\; \text{chart}(\theta, s(x)) \text{ is in control} \;\text{ and }\; q(x) \geq q\_\min(p, s(x));$$

otherwise it produces a refusal record $\rho(x, \theta, \text{reasons})$. An inference record is

$$I = \langle \text{ref}(x), s(x), q(x), \theta, K(\theta), \hat{y}, U(\theta, s(x), q(x)), t \rangle.$$

A decision rule with threshold τ and guard band produces a flag only if $|d(\hat{y})| - U > \tau$, an indeterminate result if $|d(\hat{y})| - U \leq \tau < |d(\hat{y})| + U$, and no flag otherwise.

Presentation, decision, and consequence records form a directed acyclic graph with basis edges; let $\text{desc}(N)$ denote the descendants of N. For an out-of-tolerance finding on (θ, s) over $[t_0, t_1]$:

$$\text{Suspect}(\theta, s, t_0, t_1) = \{\, I : \theta(I) = \theta,\ s(I) = s,\ t_0 \leq t(I) \leq t_1 \,\}$$

$$\text{Impact}(\theta, s, t_0, t_1) = \bigcup \{ I \;\; \text{Suspect} \}\ \text{desc}(I).$$

The impact set is computed at the basis grade its presentation records support, and the record states which. For instruments θ_1 (era E_1) and θ_2 (era E_2) with a held validation set H spanning both, measurement change is the comparison of $\{\theta_1(h), \theta_2(h) : h \in H\}$; population change is the comparison of $\{\theta_1(x) : x \in E_1\}$ with $\{\theta_2(x) : x \in E_2\}$ after bridging through H. Neither is answerable without both instruments' outputs on the same evidence, which is why the prior instrument's outputs are never overwritten.

5. Worked example: a machine-translation chain in emergency communications

A caller who speaks only Spanish reaches a 9-1-1 center. A vendor system transcribes and translates in real time; the English rendering is placed on the call-taker's screen; the call-taker enters an answer to a protocol question on the strength of it; the answer produces a determinant; the agency's response plan maps the determinant to a response level; a dispatcher assigns units. Days later a reviewer determines that a better translation of one utterance would have produced a different answer.

The chain as recorded. Link 1: the reference set holds Spanish-language cases in the strata *headset audio* and *radio-compressed audio*, with panel labels and inter-rater agreement. Link 2: the instrument θ is the vendor model at a version and configuration hash, plus the center's audio path. Link 3: K(θ) shows *Spanish, headset* validated and in control on the last check; *Spanish, radio-compressed* not validated. Link 4: the input's measured conditions — Spanish, with language-identification confidence stated; headset stratum; quality above floor. Link 5: the inference record with rendering Y and U. Link 7: the presentation record — rendered on the call-taker's screen at 02:14:31, capture grade *rendered*. Link 8: the protocol-answer decision with basis {presentation of Y}, authority *telecommunicator*; the determinant; the policy resolution against response-plan version 14; the dispatch. Link 9: the unit response. The reviewer's rendering Z is a correction record referencing Y, marked post-hoc; Y is unchanged, and any view of the 02:14 decision in as-known mode shows Y.

Refusal. Had the same call arrived over a radio patch, link 4 would have assigned *radio-compressed*, V would have been *not validated*, and the gate would have written NO VALID INFERENCE with the stratum and status. The call-taker would have been routed to the center's interpreter practice, and the record would show why no machine rendering was offered.

Out of tolerance. Three weeks later the vendor updates the model. The center's reference re-run shows *Spanish, headset* out of control while English strata are unchanged. Every Spanish-headset inference in the interval is marked suspect; the impact set enumerates every protocol answer, determinant, and dispatch that rested on them at the basis grade the presentation records support; affected reviews are annotated; the stratum is suspended, so subsequent Spanish calls receive refusals until a new calibration record is issued. The question a court will ask — which decisions rested on a translation from a model that was not performing to its validated state — has an answer.

Population versus measurement. The center tracks, per instrument, the rate at which reviewers record corrections on Spanish calls. It rises after the update. Because the prior instrument's outputs are retained and the new instrument was run over the held validation set, the rise is attributed to the instrument.

Disparity. Suppose instead the reference re-runs are flat and the correction rate still rises, but only for calls reviewed on the night rotation. The disparity chain shows links one through five flat and a difference at link six. That is a finding about the review process, not the instrument, and the chain says so.

6. The reference implementation: Standard by MetraSlate

6.1 Why the platform already has the primitives: calibration as the proving ground

Standard is one platform that ships as editions. The substrate — the record spine, tenancy, the document engine, the operations layer, the numerics — is shared; an edition declares its subjects, its vocabulary, its numbering, its document templates, and its reports in a manifest, and a fence asserts that nothing domain-specific is declared

outside one. The substrate therefore carries no assumption about what industry it records. It was proved first in calibration because calibration is the record discipline that forgives the least. A calibration laboratory's product is a certificate: a sealed, numbered, revisioned document stating a measured value, its expanded uncertainty, the reference standards it is traceable to, the decision rule applied, and the conformity statement that followed — a document a customer relies on for years and an assessor audits under ISO/IEC 17025. Standard's first edition was built to produce and keep that document, and the properties that required forced the substrate primitives that inference traceability requires. They are stated with their status: *built* is in service in the current platform; *specified* is designed in full — in the platform's decision record or in this paper — and is the next build on the substrate above. The substrate is service-grade: several hundred tables under forced row-level security, hash-verified rendering, and an exact-decimal value path proven by golden tests, all carried by every edition.

Primitive in Standard	Maturity	What it is	What it provides to inference traceability
Append-only record spine	Built	Records are never updated in place; a correction references what it supersedes; the history of any object is a read	The inference, presentation, decision, and consequence records; the instrument's longitudinal record
Hash-verified document reproduction with deterministic vocabulary projection	Built	A rendered document carries a hash of its source data; the words on the page derive from data, not mutable configuration; a re-render reproduces or fails	Calibration records and impact reports that reproduce years later; a defect in which a re-render could change wording while its hash still verified was found and closed during development — an instance of measuring the measurement
Exact-decimal value path	Built	Exact decimals with explicit scale end to end; nothing on the value path passes through binary floating point; serialized as strings; fidelity proven by golden tests	Uncertainty budgets, guard bands, and control-chart statistics that are not rounded by the platform
Uncertainty engine	Built	Propagation per JCGM 100, effective degrees of freedom by Welch–Satterthwaite, coverage factors, uncertainty budgets as records	The expanded uncertainty on every inference record
Decision rules with guard bands	Built	ILAC-G8 acceptance with guard-band multipliers; conformity, nonconformity, and the indeterminate zone as distinct results; no color on conforming results in the interface	The guard-banded decision rule; *indeterminate* as a first-class outcome
Fail-closed multi-tenant isolation	Built	Row-level security enabled and forced on every tenant-scoped table; a connection without tenant context returns zero rows; an existence-and-count oracle in search was found and closed	Class-scoped authorization and existence-classified relationships; the audit that holds identifiers and never payloads

Primitive in Standard	Maturity	What it is	What it provides to inference traceability
Reference standards and traceability in the asset spine	Built	Equipment with calibration due dates, service history, and the references each measurement traces to	Reference sets as a record class with strata, versions, and expected outputs; the traceability chain from output to reference
Domain-event emission	Specified	A typed event for every state change, feeding projections, workflows, and integrations	The ingest of instrument outputs, presentations, and decisions from external systems
Systems-integrity monitoring	Specified	Deterministic monitoring of the platform's own consistency, integration liveness, and drift from expected state	Control-chart alarms on reference strata; the minimization check; the alarm on an incomplete hold copy
Calibration records for instruments; validity gate; shadow promotion; impact traversal; refusal; disparity chain	Specified	This paper, Sections 3–4	The instrument-metrology layer, composed from the rows above
Commitments with nonces; class-scoped audit; lawful destruction with reverse traversal	Specified	This paper, Sections 3.9–3.10	Evidentiary sealing and lawful destruction, composed from the record spine and tenancy rows

Table 2. The architecture's elements mapped to Standard's primitives, with status stated. Every specified row composes rows that are built.

The first seven rows are substrate, not calibration features. They exist because the first edition's certificate had to be reproducible, exact, uncertainty-qualified, decided by rule, isolated per customer, and traceable to references — and once built as substrate, they carry into any edition unchanged. The architecture of Section 3 is the same certificate written for an inference instead of a voltage, on the same platform.

6.2 The record classes as an open specification

For Standard to be the reference implementation rather than the only implementation, the record classes of Section 3 — instrument, calibration record, reference set and case, inference record, presentation record, decision record, consequence record, refusal record, evidence grade — and the semantics of forward and backward closure must be specified independently of any product. MetraSlate proposes to publish them as an open specification with a serialization in W3C PROV [30], under an open license carrying a royalty-free patent grant for conformant implementations, with a conformance suite and a conformance mark that MetraSlate maintains, so that any system can emit conformant records and any auditor can traverse them, with Standard as the first conforming implementation. The name is the commitment: a standard is a reference others measure against.

6.3 The method: Metra Praxis

The platform is built, and its second edition was specified, by a method that applies the same epistemology to the making of the system as the system applies to the record.

The domain expert owns the truth. Every person, record, procedure, standard, number, and word the software will handle comes from the person who does the work, interviewed across eight worksheets, written in her words, and read back to her section by section until she says it is true. The operator writes a ten-page picture of the whole system from that material and from research in which every claim is tagged *observed* (read from the primary document), *claimed* (said by the expert or a vendor), or *inferred* (the operator's conclusion). An adversarial grill reads the picture against the worksheets and rules on each section. The ratifier signs, with a date. Nothing enters a build lane until the picture is ratified, and the first product is one pain removed in thirty days with its acceptance audit written before the build. The whole discovery is timeboxed to ten working days.

In the build, a rule that nobody has watched fail is not a rule: every fence is planted with a violation and watched to go red before it counts. A refuter re-runs the gate in a fresh context and its verdict is a report, never the landing. A landing happens on a recorded word. The observed/claimed/inferred tags on the process are the verified/received/derived/asserted/absent grades on the record; the fence law is the validity gate; the refuter's verdict is the finding that is never the signal; the recorded word is the promotion decision. The Public Safety Edition's requirements — the seven questions on what was knowable, the seven on boundaries and deletion, the three on human state and the instrument that measures it — were produced this way, by the Executive Director of a large emergency communications center, and this paper's architecture is the answer to them. That is the answer to CAISI's question III-B on including stakeholders and subject-matter experts in assessment: they do not review the specification; they write it.

The method is itself an instance of the architecture, and it has been running. Standard is built by AI coding agents, and the method treats each agent session as an instrument. A session is *oriented* before it acts — read the vault's state files, the decision record, and the laws — which is calibration against a reference; the vault is the reference standard. Every rule is a fence with a planted violation watched to fail, which is a check standard. A refuter re-runs the gate in a fresh context and writes a verdict that is a report, never the landing — a second instrument whose disagreement is a discrepancy, not an override. A landing happens only on the operator's recorded word, which is the promotion gate. Every claim a session makes carries an evidence tag — observed, claimed, inferred — and every landing leaves a receipt: gate results, hash pins before and after a planted fence, the refuter's verdict in the pull-request body. The instrument has changed under the method several times — successive model versions from the same provider — and the vault carries the longitudinal record across those changes. The record was tallied on 2026-09-07 and is reported in the Appendix: 525,122 lines in the current repository in 57 days, on top of 497,929 in its predecessor; six model versions crossed; every landing since the pull-request rule on a recorded decision except one; 22 of 46 fence tests with a watched receipt and the remainder a known gap; eight incidents on the record, each caught by a control the method already had and each converted into a new one; one design-drift defect found before landing and none recorded after. That is preliminary evidence, at the grade the method assigns it, for the architecture's central proposition on the one consequential AI instrument the company has operated longest: bind each output to the reference state in force, refuse outside it, verify between calibrations, promote only on a recorded decision, and keep the record across instrument changes.

7. What MetraSlate proposes to provide

1. **A reference implementation** of the *Instrument validity state* Metrology Block (Section 8.2) on Standard, shipped as the platform's Traceability Edition — the inference gateway, the calibration program, the reference-set tooling, and the review surfaces as a product for any organization whose decisions rest on AI outputs — with the Events and Tools named, the fences planted and watched, and the status of every element stated as in Table 2.

2. **An open specification** of the record classes and closure semantics of Section 3, serialized in W3C PROV, under an open license with a royalty-free patent grant for conformant implementations, with a conformance suite and mark.
3. **A worked TEVV-Athlon** on a machine-translation chain in emergency communications, from Articulate & Organize through Synthesize & Interrogate, with a serving emergency communications center as the reference site for its Events, subject to that center's procurement and disclosure rules.
4. **Participation** in the NIST AI Consortium's task groups on measurement science and evaluation, and in the TEVV-Athlon Framework's revision, as the metrology practitioner's seat.

8. Relation to NIST AI 200-2 and the CAISI open questions

8.1 Terminology

NIST AI 200-2 requests input on the definitions of TEVV terms and on other AI measurement terms and their sources [3]. We propose that the draft adopt, or align with, the VIM's definitions of *verification* (2.44) and *validation* (2.45), which already carry the distinction the draft draws between meeting specifications and being adequate for intended use, and adopt *metrological traceability* (2.41), *measurement uncertainty* (2.26), *instrumental drift* (4.21), and *maximum permissible error* (4.26) as the vocabulary for the ongoing validation it calls for in §4.5. We further propose that the draft state the collision between metrological *calibration* and machine-learning *probability calibration* and resolve it by qualification, since the draft uses the word in the metrological sense in Appendix D and the machine-learning sense in §4.6.

8.2 A Block, its Events, and its Tools

- **Block: Instrument validity state.** For each consequential output, the validity status of the producing instrument for the input's stratum at the time of production, with the reference evidence and uncertainty supporting it.
- **Events.** Continuous reference re-run (control charts per stratum, including the conditioning layer); shadow comparison before promotion; out-of-tolerance impact assessment on a failed check; refusal audit (rate and reasons per stratum).
- **Tools.** Stratified reference sets with sealed subsets; control-chart methods; the GUM-based uncertainty budget; the guard-banded decision rule; the lineage store with basis edges; the impact-set query.

The Block belongs to the Operate & Monitor lifecycle stage, where the draft currently lists observational methods, and it turns the athlon from an event into a standing condition: the instrument is always under assessment, and every output records the assessment's state.

8.3 Appendix D

The consideration "Calibration of evaluation processes" [3, Table 6] is extended in two directions: from the evaluation process to the deployed instrument, and from periodic verification to a binding of each output to the verification state in force. We suggest the appendix add a consideration in those terms, and add *impact analysis on prior results* as the consequence of a failed verification, citing ISO/IEC 17025 §7.10.

8.4 The CAISI questions

- **II-A, uncertainty.** The uncertainty budget identifies the sources — conditioning repeatability, model reproducibility, input quality, validation coverage, reference-label disagreement — and the inference record communicates them as an expanded uncertainty on every output, not on an aggregate.

- **I-B.3, pre- versus post-deployment.** Shadow comparison is a direct measurement of the gap between offline evaluation and live behavior, per stratum and per comparison group, before promotion.
- **II-D, reporting.** The calibration record is the reporting unit and carries exactly the detail the question asks for: reference set and version, strata, results, uncertainty budget, validity domain, promotion decision and evidence.
- **I-D.2, AI to monitor AI.** A second instrument used to check a first is a *check standard* and must carry its own calibration record; its disagreements with the instrument under test are discrepancy records, never overrides.
- **III-A.1 and III-A.2, downstream outcomes and the effect of interventions.** The consequence records and the disparity chain give a reproducible method for measuring change in downstream outcomes over time and for locating the effect of an intervention — a model promotion, a policy version, a threshold change — as an association with a stated window, with causation reserved for a human finding.
- **III-A.3, tracking components of multi-turn human–AI workflows.** The presentation and decision records with basis and authority are exactly such a tracking method, and the capture grade states how much of the human side the record actually observed.
- **III-B, stakeholder consultation.** Section 6.3.

9. Limitations and open questions

Construct validity is not solved by traceability. The architecture makes an instrument's validity state traceable; it does not make the construct the instrument measures valid. A translation instrument's reference labels have a defensible ground truth; an instrument claiming to measure a human state does not, and the architecture's response — a closed dimensional vocabulary, panel labels with recorded disagreement and composition, refusal outside validated strata, and a firewall between such inferences and any consequential use — is containment, not solution.

Reference-set coverage is a cost and a program. Strata the reference set does not cover are *not validated* and the gate refuses them. Building stratified reference sets under lawful basis scales with the diversity of the operating population.

Capture grade bounds the impact set. Where presentation surfaces emit no display or interaction events, the impact set is computed at *available* grade and is a superset of the true set. The record states this; it cannot fix it without instrumentation of the surface.

Coverage depends on the enforcement point. An inference path that bypasses the recording layer produces no inference record. The gateway and the coverage check (3.6) make bypass detectable where vendor logs exist, not impossible; a deployment that wants a complete record routes every call through the layer and says so in its calibration program.

Nondeterministic instruments. For large language models, run-to-run variation is a repeatability component of the budget, and reference re-runs must be repeated to estimate it.

Program. The substrate is built and in service; the instrument-metrology layer is specified in full and composes primitives that exist. What remains is a build program with a stated scope — Section 7 — and the stratified reference sets that any implementation, on any platform, will need to assemble under lawful basis.

10. Conclusion

NIST has asked how AI measurement science should identify and communicate uncertainty, how pre-deployment evaluation should predict post-deployment behavior, how to report, how to track the components of human–AI workflows, how to measure downstream outcomes and the effects of interventions, and how to include the people who know the work. Calibration laboratories answer the corresponding questions for physical instruments every day, under accreditation, with a vocabulary and procedures that have been stable for decades, and the record system that keeps those answers has been built. This paper has proposed that the procedures transfer to consequential AI inference when three things are in place: a record that binds each output to the instrument's validity state at the time of production; a gate that refuses rather than discounts outside the validated domain; and lineage that makes the impact of an out-of-tolerance finding a query over decisions and consequences. The result is an instrument that can be calibrated, verified between calibrations, changed under control, recalled when it drifts, and audited for the point at which a disparity enters — and a record in which an organization can state, for any consequential decision, what the machine knew, how much to trust it, and what has changed since. MetraSlate built that record as a domain-agnostic platform, proved it first in calibration laboratories, has specified it for emergency communications with the people who run one, and proposes to make it the reference implementation of this discipline.

Acknowledgments

This paper exists because Tina Buneta, ENP, CPE, RPL, Executive Director of Aurora911 — the emergency communications department of the City of Aurora, Colorado, which serves a city of more than 400,000 people and handles more than 600,000 calls for service a year — agreed to review an outsider's architecture for her profession and then declined to let it off easy. Over three rounds of written questions she asked what "the incident" is when one event crosses three disciplines and two jurisdictions; whether the record could show what a telecommunicator knew at 14:03:17 without letting 14:19 leak backward into the judgment; which system owns which truth; how a protocol deviation is told apart from professional judgment; what an immutable record does when the law says destroy; whether knowing that two records intersect is itself a leak; and what a system should refuse to know. She then asked the question that changed the work: how would the AI detect the introduction or development of bias in its own inferences, and, when a person appears to change after a model update, how would anyone know whether the person changed or the instrument did. Her question is the one Section 3 answers, and the answer is the method the author had already been running for months on his own AI coding instruments — reference cases re-run continuously, verification before anything lands, a recorded decision before any promotion, refusal outside the validated state — applied to the instrument she asked about. She named it: calibrate your AI.

Mrs. Buneta began her career in 1999 as a front-line telecommunicator with the Colorado State Patrol and served that agency for more than twenty years across multiple communications centers as a training officer, supervisor, and regional center manager before taking the Aurora911 seat in December 2019. She has been recognized as APCO International's Director of the Year and Colorado NENA-APCO's Director of the Year, has been entered into the Congressional Record for her service, and co-chairs the APCO International Staffing Task Force. Her rule that a wellness signal must never outrank the human being is written into the architecture as a constraint the schema enforces, and it is the standard against which every other rule in this paper should be read.

She is also the author's mother. That fact is stated so that no reader has to discover it, and because it is the reason a metrologist was ever allowed to sit in a 9-1-1 center's chair long enough to learn what the record has to prove.

References

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

[2] Keller D, Steed R, Bergman S, and the Applied Systems Team at CAISI (2025). *Accelerating AI Innovation Through Measurement Science*. CAISI Research Blog, NIST, December 2, 2025.

[3] Phillips PJ, Jensen T, Hall P, Amironesei R, Choong Y-Y, Greenberg C, Greene KK (2026). *The TEVV-Athlon Framework for Evaluating AI Systems*. NIST AI 200-2 ipd (Initial Public Draft), August 2026. https://doi.org/10.6028/NIST.AI.200-2.ipd

[4] National Institute of Standards and Technology (2026). *NIST Artificial Intelligence Consortium*. Federal Register notice, May 29, 2026.

[5] ISO/IEC 17025:2017. *General requirements for the competence of testing and calibration laboratories*. Clauses 7.7 (Ensuring the validity of results) and 7.10 (Nonconforming work).

[6] ANSI/NCSL Z540.3-2006 (R2013). *Requirements for the Calibration of Measuring and Test Equipment*.

[7] ILAC-G8:09/2019. *Guidelines on Decision Rules and Statements of Conformity*.

[8] Guo C, Pleiss G, Sun Y, Weinberger KQ (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*.

[9] JCGM 200:2012. *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*, 3rd edition.

[10] Amironesei R, Godil A, Greenberg C, Greene K, Hall P, Jensen T, Fiscus J, Schulman N (2025). *Assessing Risks and Impacts of AI (ARIA): ARIA 0.1 Pilot Evaluation Report*. NIST AI 700-2.

[11] Center for AI Standards and Innovation (2026). *Practices for Automated Benchmark Evaluations of Language Models*. NIST AI 800-2 ipd.

[12] Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. NIST Special Publication 1270.

[13] National Institute of Standards and Technology (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1.

[14] JCGM 100:2008. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement (GUM)*.

[15] ISO 10012:2003. *Measurement management systems — Requirements for measurement processes and measuring equipment*.

[16] Adel T, Bilson S, Levene M, Thompson A (2024). Trustworthy artificial intelligence in the context of metrology. National Physical Laboratory. arXiv:2406.10117.

[17] AI Standards Hub (2025). *How National Metrology Institutes build confidence in AI*. March 2025.

[18] Mitchell M, Wu S, Zaldivar A, et al. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

[19] Gebru T, Morgenstern J, Vecchione B, et al. (2021). Datasheets for datasets. *Communications of the ACM* 64(12).

[20] Arnold M, Bellamy RKE, Hind M, et al. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63(4/5).

[21] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A (2014). A survey on concept drift adaptation. *ACM Computing Surveys* 46(4).

[22] Vovk V, Gammerman A, Shafer G (2005). *Algorithmic Learning in a Random World*. Springer.

[23] Angelopoulos AN, Bates S (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning* 16(4).

[24] U.S. Food and Drug Administration (2024). *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions*. Final guidance, December 2024.

[25] Regulation (EU) 2024/1689 (Artificial Intelligence Act). Articles 12 (Record-keeping) and 72 (Post-market monitoring).

[26] ISO/IEC 42001:2023. *Information technology — Artificial intelligence — Management system*.

[27] ISO/IEC TR 24029-1:2021. *Assessment of the robustness of neural networks — Part 1: Overview*.

[28] ISO/IEC TS 4213:2022. *Assessment of machine learning classification performance*.

[29] ISO/IEC 5259 series. *Data quality for analytics and machine learning*.

[30] W3C (2013). *PROV-DM: The PROV Data Model*. W3C Recommendation.

[31] NCSL International (2014). *RP-1: Establishment and Adjustment of Calibration Intervals*; and laboratory practice on as-found/as-left reporting.

[32] Montgomery DC (2019). *Introduction to Statistical Quality Control*, 8th edition. Wiley.

Appendix: Evidence from the method

The counts below were produced on September 7, 2026 by a read-only tally over the company's project vault and the Standard repository — run, in keeping with the method, by an agent session oriented against the vault — with the command or file that produced each value recorded beside it, nothing rounded, and anything that could not be counted marked unverified. They are aggregate counts; no code, file contents, or paths are reproduced. They are the company's evidence for the claim in Section 6.3 and are offered at the grade the method itself assigns to them: observed.

The record. The vault's evidence ledger begins 2026-07-11; the Standard repository's first commit is 2026-07-12. Between then and 2026-09-07: 497 decision-record head lines (495 distinct decisions, highest D-502; 21 amended, 3 superseded); 522 evidence-ledger entries; 54 specifications (25 banner-marked ratified); 29 adversarial grill files, 5 of which returned blocking findings; 462 work orders; 45 lane briefs; 106 archived state hand-offs; 184 recorded orientations (a session reading the vault's state, decisions, and laws before acting); 35 checkpoint stamps.

The instrument and its changes. The records name six distinct model versions from one provider across the period — Claude Opus 5, Fable 5, Fable 5.1, Sonnet 5, Haiku 4.5, and Opus 4.8 — by token count 107, 101, 52,

29, 10, and 8 mentions respectively. The routing table that assigns model and effort per task carries 10 rows and 10 amendment marks. The three repository hooks are byte-identical to their vault sources.

Fences (check standards). 46 Kotlin fence, architecture, contract, inventory, and boundary test files, of which 22 have a recorded plant-and-watch receipt (a hash pair or a "watched biting" line), 9 are named in the records without such a receipt, and 15 are never named. 17 web lint rules, of which 8 have such a receipt. 175 record lines contain "planted," 72 contain "watched biting," 93 contain a hash pin. 11 ledger entries record a hook refusing a real command; one refused a command during the tally itself.

Refutation (second instrument). 23 receipted refuter-verdict artifacts: 7 verdict files, 14 pull-request bodies naming a refuter (12 carrying hash pins, 7 carrying the literal "NOT REFUTED"), 2 repository reports. Distinct refuter runs are unverified: the sources overlap and no register assigns run identifiers.

Landings (promotion on a recorded decision). Since the pull-request rule took effect on 2026-09-06, every merge to `main` has been by pull request; 555 first-parent commits before that date were direct. 18 pull requests merged; 17 carry the operator's recorded word (14 landing rows with timestamp and quoted word; 3 in archived state heads); 1 (2026-08-16, before the rule) carries none. One false landing row was written and caught the same day (ledger 943), and the landing script is now hash-pinned.

Incidents. The record holds each with its date, the rule it broke, what caught it, and the rule or control that followed: an agent session that spent seven hours on a document it should not have produced (2026-07-29; a design-gate hook followed); four corrections in one day with no checkpoint fired (July; D-125; the checkpoint rule followed); a verifier returning one real finding against four false ones (ledger 602, 611; a law on verifier findings followed); a builder dispatched before its grill returned (2026-09-05; an agreements addendum followed); a merge without the ratifier's review (2026-09-05; caught in the next day's hand-off review; branch protection and a landing script followed); a module deleted in a lane worktree and restored from git (2026-09-06); a false landing row (2026-09-07; caught within the hour, the landing script hash-pinned); and a development checkout's dependencies emptied through a junction and restored (2026-09-07). A review on 2026-09-06 named seven process failures and produced three ratified rule packets. Every incident became a control. Drift, as the vault defines it — divergence of the code from a ratified decision without a recorded amendment — appears once in the decision record before landing (the V-1 rendering defect, closed by a substrate decision) and the tally found no ledger entry recording one after landing; absence in the record is not proof of absence in the code, and the tally says so. The tally also measured drift in the record itself: a work-order index of 458 against 462 files; a decision index of 498 lines over 497 heads and 495 distinct numbers; 36 gaps in migration numbering; 226 of 522 ledger entries with no date on their head line. Continuous integration on `main` recorded 187 runs: 97 succeeded, 86 failed, 4 cancelled; the failures on `main` since 2026-09-05 are attributed in the record to a single named integration test.

Volume. Standard at `main` on 2026-09-07: 3,201 tracked text files, 525,122 lines, of which 507,098 are code (Kotlin 233,205; TSX 140,050; TypeScript 72,864; SQL 29,097; CSS 24,317; other 7,565); 664 commits reachable from `main`; 19 pull requests; one author under two git identities. The predecessor build, counted separately with the same exclusions: 2,495 text files, 497,929 lines, 1,536 commits, 2026-04-26 to 2026-07-11. Database, from 352 migration files: 339 live tables, 290 carrying a tenant column, 285 with row-level security enabled and forced, 5 carrying the column without it (a finding the tally surfaced).

What the numbers say, and do not. They say the method has produced roughly a million lines across two repositories in five months, with the instrument changed six times under it; that since the landing rule every merge but one carries a recorded human decision; that about half the fences carry a watched receipt and the rest are a known gap; and that every incident on record was caught by a control the method already had and became a

control the method now has. They do not say that no drift escaped; they say that none is recorded, that the record's own inconsistencies are counted, and which records would have to exist to show more.