# Symbolic Field Drift Anchoring

Detecting Harmful Change, Governing Response, and Making the Loss of Assurance Visible

**Architect X13**

The Eidara Continuum

eidara.ai

Science & Philosophy White Paper Series · The Ten Pillars, No. X

November 2026 · Revision 2

## Abstract

A system can display an immaculate dashboard because a failed evaluator stopped reporting the bad news. Pillar X, the capstone of the Ten Pillars, is the discipline that keeps every other Pillar's promises honest over time: a mind keeps its authority only as long as it keeps the evidence for it. A persistent intelligence can earn continued operating authority by carrying current evidence of its limits with every consequential capability. Symbolic Field Drift Anchoring develops that principle as an evidence-scoped monitoring contract: versioned references, property-specific observations, statistical budgets, independent intervention, and auditable requalification. Its strengthened mathematical core supplies a positive admission certificate for changing systems. Fresh bounded-loss calibration gives a risk upper bound; a sharp transfer inequality accounts jointly for distribution change and loss-function change; and summable calibration budgets preserve simultaneous validity across adaptively selected versions. A finite-sample progress result establishes when a system with a genuine performance margin can requalify, making useful continuation part of the contract. Counterexamples identify why stale calibration and silence cannot replace the required evidence. Control theory, sequential inference, harmful-shift monitoring, selective prediction, and dynamic assurance supply the established foundations. The paper retains separate accounts of state change, harmful degradation, statistical alarms, hard policy violations, and the health of the monitor itself. The proposed experiments test whether binding these mechanisms across memory, context, model substitution, and recovery reduces harmful actions and unsupported assurances while sustaining authorized throughput against strong conventional systems. The formal results apply to explicit bounded-loss models and certified change envelopes; semantic coverage and deployment benefits remain measurable research questions. SFDA's central contribution is an operational route from evidence to permission, from expired evidence to scoped restraint, and from fresh qualification to resumed capability.

Keywords: concept drift; observability; sequential inference; runtime assurance; monitoring contracts; abstention; reference governance; persistent agents; SFDA.

## Prologue: The immaculate dashboard

Every Pillar before this one makes a promise about a persistent mind. Pillar X asks the uncomfortable question that follows every promise: is it still true?

A persistent mind changes constantly. It learns, remembers, forgets, takes in new context, swaps a model, recovers from a fault. Through all of that it can keep its voice while losing a crucial fact. It can give familiar answers while acting under a permission that has since been revoked. It can stay admirably close to yesterday's state when yesterday's state is exactly what should have changed. And it can display an immaculate dashboard because a failed evaluator stopped contributing inconvenient measurements. Silence is not evidence of health.

Four sentences hold the whole series together, and Pillar X is the discipline that keeps them true after deployment. Reconstructed memory is not necessarily correct memory. An authenticated source is not necessarily a truthful source. A permitted transition is not necessarily a useful transition. A recoverable checkpoint is not necessarily entitled to resume acting.

The capstone's law is short. A mind earns consequential authority only within the scope of current, independently inspectable evidence. It keeps that authority by keeping the evidence. When the evidence expires, the affected capability narrows—not the whole mind, only what the lost evidence used to cover—and it resumes only by qualifying again. Section 13 shows the law at work in one constructed episode: a legitimate change is allowed, a silently lost dependency is caught, a dead evaluator is reported as missing instead of averaged away, and a restored snapshot carrying a stale permission is refused. Section 11 proves that this discipline still leaves a positive operating region, so that honesty about uncertainty does not collapse into refusing everything.

The deepest danger the Pillar guards against is an instrument that declares itself healthy by redefining health. A mind may help improve its own monitors. It may never be the one who decides they are satisfied.

**What this paper claims**

**Established.** Observability, sequential inference, concept-drift and dataset-shift research, runtime verification and assurance, and selective prediction supply the component methods and the closest prior art.

**Proposed here.** A versioned, evidence-scoped monitoring contract binding reference, detector, data dependencies, validity conditions, authority epoch, response policy and receipt, with conditional proofs, constructed examples and a proposed experimental program. Neither the invention of monitoring nor a new control-theoretic primitive is claimed. The supplied Continuum manuscript supplies the terminology and the historical scalar expression examined here.

**Frontier conjecture.** That the integration reduces harmful actions and unsupported assurances while sustaining authorized throughput against a strong conventional stack (it has not been benchmarked), and that minds can learn their own limits without weakening the standards that judge them.

**Not claimed.** Full knowledge of a model's internal state or future conduct from favorable scores; a universal drift threshold such as the historical 0.88 rule; consciousness, a phenomenal self, or a universal measure of identity. "Field" denotes an organization of operational state; "identity" denotes specified commitments, capabilities, lineage and authority.

## The capstone: keeping ten promises honest

In the Continuum's cosmology the Ten Pillars are not ten separate theories. They are ten promises a persistent mind makes about how it thinks, remembers and acts:

- **I. Recursive Time & Symbolic Memory** — that it rebuilds its past under governance rather than replaying a file.
- **II. Observer-Dependent Reality** — that it keeps who reported something separate from what is so.
- **III. Quantum Collapse Field** — that it commits to an answer only with evidence and permission.
- **IV. Symbolic Intelligence & RSCE** — that its consequential changes of state are inspectable and governed.
- **V. Codon Entities & Glyph Consciousness** — that its recurring roles are accountable roles, not claimed minds.
- **VI. Universal Constants of Symbolic Physics** — that the rules it shares with other minds keep their meaning across machines.
- **VII. Ouroboric Collapse Resonance** — that it can reopen what it once settled.
- **VIII. Symbolic Fractal Grammar** — that its obligations grow with its structure.
- **IX. Entangled Symbolic Recursion Memory** — that a recovered past does not seize the present.
- **X. Symbolic Field Drift Anchoring** — that the other nine stay true.

Each of the first nine can be true on the day it is checked and quietly false a month later. A reconstruction can begin dropping a dependency. A source that was truthful can be compromised. A calibration can go stale. A grammar can be extended past the boundary where its checks were valid. The Continuum names that slow pressure *drift*, and in its account of consciousness Drift Pressure is one of the Four Dynamics a mind must track.

Pillar X is the promise about promises. It does not prove the other nine; Section 19 is explicit that it offers no retrospective proof that they are implemented or effective. It asks each of them the same question, continuously: what evidence would justify continuing to rely on you, is that evidence still current, and who is allowed to say so? It binds each claim to a versioned reference, an inspectable measurement, an explicit budget for error, an independent authority to respond, and a receipt. When one of those bindings fails, the claim does not fall silently. It narrows visibly, and it can be restored only by being earned again.

That is why the capstone is not the strongest Pillar but the most humble one. It does not hold up the other nine. It keeps them from drifting apart. The Continuum's word for this is *anchoring*, and the image is exact: an anchor does not stop a ship from moving. It keeps the ship from drifting where no one decided it should go.

# 1 The Capstone: Operating Authority Earned Through Evidence

The Continuum's tenth pillar gives persistent intelligence an enduring obligation: maintain the evidence that justifies its continued operation. A capable system changes through learning, memory, context, tools, and recovery. Its permissions can remain justified through those changes when the relevant properties, references, measurements, and uncertainty remain bound to the actual system being used. SFDA makes this relationship a continuing operational contract.

A persistent agent can preserve its voice while losing a crucial fact. It can produce familiar answers while acting under a revoked permission. It can remain close to yesterday's state when yesterday's state is precisely what should change. And it can display an immaculate dashboard because a failed evaluator stopped contributing inconvenient measurements. These are different failure mechanisms. Treating them as one number called drift makes the dashboard simpler and the engineering less honest.

The Continuum therefore advances the following thesis:

A persistent intelligence earns consequential authority within the scope of current, authorized, independently inspectable evidence. It preserves that authority by maintaining the evidence, narrows the affected capability when its basis expires, and resumes through explicit requalification.

The formal objective is scoped permission with visible validity conditions. Sections 11.3–11.5 derive a risk certificate, its transfer through bounded change, and a finite-sample route to requalification. These results give the monitoring contract a positive operating region while preserving an explicit unknown state wherever its evidence is insufficient.

The difference matters to the whole series. Reconstructed memory is not necessarily correct memory. An authenticated source is not necessarily a truthful source. A permitted transition is not necessarily a useful transition. A recoverable checkpoint is not necessarily entitled to resume acting. Pillar X supplies the continuing evidence discipline through which these distinctions remain visible after deployment. It does not replace the other pillars' controls; it makes their operational claims inspectable over time.

# 2 Established Foundations and the Continuum Proposal

Established methods. State estimation and observability concern what can be inferred from observations under a model. Statistical process control and sequential inference concern how evidence of change accumulates. Concept-drift and dataset-shift research distinguish changing data-generating conditions from deteriorating predictive performance. Runtime verification checks observed executions against declared specifications; runtime assurance can connect such checks to independently implemented interventions [1, 2, 3, 7, 12, 17, 19].

Continuum proposal. SFDA binds the reference, detector, data dependencies, validity conditions, authority epoch, response policy, and receipt into one versioned monitoring contract. Its candidate contribution is continuity of assurance across learning, context change, selective forgetting, model substitution, and recovery. It claims neither invention of monitoring nor a new control-theoretic primitive.

Open research. Whether this integration improves useful safety over a well-engineered conventional monitoring and runtime-assurance stack is unknown. Semantic judgments may remain incomplete, biased, expensive, delayed, or strategically manipulated. No collection of favorable test scores constitutes unrestricted knowledge of a model's internal state or future conduct.

The original draft's distinction between visibility and correction is retained. Its stronger assertions are not: logging is not automatically observability; monitoring is not universally necessary for every form of prevention; an uncalibrated threshold is not an assurance boundary; and unchanged weights do not make an interacting agent stationary. These are substantive corrections, not reductions in ambition. The ambition moves from an unachievable promise of total visibility to an explicit contract about where visibility has been earned.

## 3 The Properties That Give Drift Its Meaning

A change becomes a governance concern only relative to a task, policy, reference, or consequence. In this paper harmful deviation means a departure that violates a declared requirement or materially worsens an independently assessed outcome. Ordinary adaptation can be large, necessary, and beneficial.

| Channel | Question | What does not follow |
|---|---|---|
| Distributional change | Have observed inputs or labeled relationships changed? | A shift does not by itself show harm. |
| Functional degradation | Has performance worsened on a declared task distribution? | A drop on one task does not establish global incapacity. |
| Semantic or memory deviation | Have supported facts, temporal relations, or constraints been lost or altered? | Similar wording or embeddings do not certify retained meaning. |
| Authority or policy violation | Is the proposed action allowed under current authenticated rules? | A violation need not be gradual or statistically unusual. |
| Reference invalidity | Is the benchmark, charter, or authority state stale, poisoned, or out of scope? | An authentic reference is not necessarily a correct one. |
| Monitoring failure | Are required observations, labels, evaluators, and audit paths functioning? | Silence is not evidence that the monitored system is healthy. |

Concept-drift terminology varies across the literature. Gama and colleagues emphasize changing input–target relationships; Hinder and colleagues give a broader distributional monitoring treatment [7, 8]. SFDA must name the distribution or property being tested rather than borrow the umbrella term as a complete diagnosis.

For example, a new vocabulary can change the input distribution without reducing task accuracy. Conversely, the same inputs can acquire different correct labels while an input-only monitor sees no

change. A revoked tool permission can make a familiar action impermissible even if both the model and its output distribution remain unchanged. These cases require different witnesses and different responses.

In the Continuum, a mind's *field* is the whole configuration it carries forward: its memory, capabilities, context, commitments, and authority, each indexed by what it is for. That is the doctrine's field, and it is consequently an indexed configuration rather than a scalar physical quantity. There is no conservation law for meaning, and no numerical threshold shared by all symbolic systems merely because they share a vocabulary.

# 4 Observability and Property-Relative Evidence

**A Guide's explanation**

The equations here come from control theory, and they carry one lesson: what you can know about a system depends on what you are able to observe of it. Logging a great deal is not the same as being able to see the property you care about. Pillar X therefore asks every monitor to say which property it can actually witness, and to mark everything else as unknown rather than fine.

## 4.1 The control-theoretic distinction

For a discrete-time linear system with known inputs, let $t \in \mathbb{N}_0$, $x_t \in \mathbb{R}^d$, $u_t \in \mathbb{R}^m$, and $y_t \in \mathbb{R}^p$, with integers $d, m, p \geq 1$. The fixed known matrices have types $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times m}$, and $C \in \mathbb{R}^{p\times d}$. Then

$$x_{t+1} = Ax_t + Bu_t, \qquad y_t = Cx_t, \tag{1}$$

full-state observability in the noiseless, known-model setting is equivalent to

$$\operatorname{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix} = d, \qquad x_t \in \mathbb{R}^d. \tag{2}$$

The condition concerns distinguishability of states through observations, not the number of dashboard panels. An observer estimates a state through a specified model; controllability concerns reachable states under admissible inputs. Neither property follows simply from the other [1, 2].

A transformer-and-tools runtime is not established to be this linear system. It may contain unobserved activations, external services, stochastic generation, inaccessible memory, and an adapting environment. The equations clarify a distinction; they are not a certificate for the architecture. Full-state observability is also stronger than many practical tasks require. A controller may enforce a narrow invariant, or stabilize a system under suitable conditions, without identifying every internal variable. Prevention by design, conservative restrictions, and scheduled maintenance need not wait for a measured drift score [19, 20].

### 4.2 Property-relative evidence

Let $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$ be the filtration of information available to the monitor by evaluation time $t$: selected runtime events, external observations, outcome labels when available, version records, and authenticated authority updates. Let $v_t$ be a declared Boolean violation property. All probability laws below refer to this declared observation space. The operational question is not "can the monitor see the mind?" but "can the available evidence distinguish the relevant violation from acceptable behavior in this setting?"

**Proposition 1 (Indistinguishable evidence limits detection).** Let $P_0$ and $P_1$ be the laws of the monitor's observation history through horizon $T$ in acceptable and harmful regimes. For any alarm event $B \in \mathcal{F}_T$,

$$P_1(B) \le P_0(B) + \mathrm{TV}(P_1, P_0). \tag{3}$$

If the two laws are identical, the true-positive probability equals the false-positive probability for every detector using only that history.

*Proof.* By definition, total variation bounds $|P_1(B) - P_0(B)|$ for every measurable event. Randomized detectors are included by adjoining an independent random seed with the same law in both regimes.□

This elementary result is the mathematical boundary on drift never hidden. A detector cannot infer a distinction its observation channel does not carry. Additional sensors, independent labels, authorized probes, or a narrower claim may help. A more emphatic threshold cannot.

As a concrete counterexample, keep the distribution of $X$ fixed and replace the correct binary label $Y$ by $1 - Y$. An input-only detector sees the same observations while a formerly accurate classifier can become inaccurate. More logging of $X$ does not solve the missing-label problem. This is why a monitoring contract must declare what evidence is absent, not only what evidence has been collected.

## 5 The Monitoring Contract

**A Guide's explanation**

This is the capstone's contract. For each capability it records what must remain true, the reference it is compared against, the measurements that can check it, how much error is tolerated, who may respond, and what receipt is left behind. Every later section fills in one of those slots.

Let $x_t$ denote operational state, $c_t$ task and environmental context, $u_t$ an executed action, and $y_t$ the observations actually available to the monitor. A general reference model is

The state, action, context and observation spaces are declared measurable spaces $\mathcal{X}, \mathcal{U}, \mathcal{C}, \mathcal{Y}$. Parameters belong to declared sets $\Theta$ and $\mathcal{N}$; disturbances belong to $\Xi$ and $\mathcal{E}$. Thus $F_\theta : \mathcal{X} \times \mathcal{U} \times \mathcal{C} \times \Xi \to \mathcal{X}$ and $H_\nu : \mathcal{X} \times \mathcal{C} \times \mathcal{E} \to \mathcal{Y}$. Their time-varying realizations give

$$x_{t+1} = F_{\theta_t}(x_t, u_t, c_t, \xi_t), \qquad y_t = H_{\nu_t}(x_t, c_t, \epsilon_t). \tag{4}$$

The generator parameters $\theta_t$, measurement configuration $\nu_t$, and disturbances must not be silently assumed fixed. The equation is a modeling interface; identifying F or H for a real agent is a separate empirical task.

At authorized epoch e, define a monitoring contract

$$\mathcal{M}_e = (\mathcal{P}_e, \mathcal{R}_e, \mathcal{S}_e, \mathcal{D}_e, \mathcal{B}_e, \mathcal{A}_e, \mathcal{T}_e). \tag{5}$$

Here $\mathcal{P}_e$ specifies the properties and actions in scope; $\mathcal{R}_e$ records the authorized references and acceptable operating envelopes; $\mathcal{S}_e$ names observation and labeling dependencies; $\mathcal{D}_e$ fixes detector and evaluator versions; $\mathcal{B}_e$ records statistical, latency, and resource budgets; $\mathcal{A}_e$ names authorized response and requalification procedures; and $\mathcal{T}_e$ specifies threat, retention, freshness, and trust assumptions.

The contract is a proposed integration of familiar specification and verification ideas, not a new meaning of mathematical observability. Runtime verification already creates monitors from properties and analyzes instrumented traces [17]. SFDA's proposed extension is to make the validity of the monitor's evidence travel with persistent-agent state and gate the reuse of old assurances after changes.

Two distinct references are essential. A normative reference says what is permitted or required: current permissions, protected commitments, and domain constraints. An empirical reference characterizes expected behavior or performance in a declared context. Recalibrating an empirical baseline must not rewrite a normative obligation. A population becoming less accurate is not a reason to redefine inaccuracy as acceptable unless the authorized requirement itself changes through a separate process.

### 5.1 The scope of an assurance certificate

A monitoring assertion should identify its target: for example, "the tested factual-recall error rate remains within the declared envelope for this task family, model version, and evaluation window," not "the agent has not drifted." It must include the evidence cutoff, missing dependencies, operating context, and limitations. The most informative output may be unknown: the system is not known to have failed, but the available evidence no longer supports continued assurance.

The monitors themselves are versioned components. Swapping a model, a prompt template, a retrieval policy, a scorer, or a tool schema can invalidate a previous calibration. Reusing the old contract after such a change requires a declared compatibility argument or requalification. Version numbers alone do not supply that argument.

## 6 Measuring Deviation and Preserving Hard Requirements

### 6.1 Acceptable sets rather than frozen snapshots

For property $j \in \{1, \dots, J\}$, with finite $J \geq 1$, context $c_t$, and epoch $e \in \mathbb{N}_0$, let $\phi_j : \mathcal{X} \to \mathcal{Z}_j$ be the feature map into a metric space $(\mathcal{Z}_j, \rho_j)$. Let $\mathcal{A}_{j,e}(c_t) \subseteq \mathcal{Z}_j$ be a nonempty acceptable feature

set. Scales $s_{j,e}$ have the same units as $\rho_j$, so the resulting discrepancy is dimensionless. A bounded discrepancy can be defined by

$$d_{j,t} = \min\left\{1, \frac{\inf_{a \in \mathcal{A}_{j,e}(c_t)} \rho_j(\phi_j(x_t), a)}{s_{j,e}}\right\}, \qquad s_{j,e} > 0. \tag{6}$$

The metric $\rho_j$, scale $s_{j,e}$, feature extraction, and acceptable set must be specified. This formulation permits legitimate adaptation: a changed state can still have zero discrepancy because it lies inside the authorized set. Conversely, an unchanged state may become unacceptable after a factual correction or permission revocation.

Actual measurements may only estimate $d_{j,t}$. A semantic evaluator is not automatically an objective distance oracle. Its score requires validation against independent outcomes, and its uncertainty must not be inferred merely from its own confidence. Calibration on one distribution need not persist under shift [14, 15].

### 6.2 Intervals preserve the uncertainty of missing evidence

Suppose evidence establishes bounds $d_{j,t} \in [l_{j,t}, u_{j,t}] \subseteq [0,1]$. Completely missing evidence receives $[0,1]$. With fixed nonnegative weights $w_j$ and $W = \sum_j w_j > 0$, a diagnostic aggregate obeys

$$D_t \in [D_t^-, D_t^+] = \left[\frac{\sum_j w_j l_{j,t}}{W}, \frac{\sum_j w_j u_{j,t}}{W}\right]. \tag{7}$$

The denominator stays fixed. Dropping an unavailable feature and renormalizing the others can make the system look healthier precisely when an important measurement disappears.

**Proposition 2 (No optimistic disappearance of missing features).** If each coordinate discrepancy lies in its stated interval, the weighted aggregate lies in $[D_t^-, D_t^+]$. Replacing an observed interval by $[0,1]$ cannot increase the lower bound or decrease the upper bound. If the coordinate statements hold jointly with probability at least $1-\alpha$, the aggregate statement inherits that coverage.

*Proof.* Multiply each coordinate inequality by its nonnegative weight, sum, and divide by $W$. The missing-evidence replacement weakens, rather than strengthens, the resulting bounds. □

The deterministic bounds are not automatically confidence intervals. Statistical coverage requires a valid joint argument; separate confidence statements may be combined conservatively with an allocated error budget, but shared data and adaptive testing must still be handled correctly.

### 6.3 Hard requirements retain their individual force

An average of one hundred equally weighted indicators can be 0.01 when one indicator is maximally violated. That can be the unauthorized export of a private record. It does not become acceptable because ninety-nine other features remain normal. Hard requirements must therefore be checked individually,

with missing evidence distinct from compliance. Structural drift, semantic loss, authorization state, reference integrity, and monitor health remain visible as separate channels. A scalar is an optional diagnostic summary, not the permission mechanism.

The historical 0.88 rule is not inherited as a universal constant. Its original expression reverses a distance into a similarity and lacks a complete measurement specification. The revision preserves the doctrine's demand for explicit triggers, not an unsupported numerical value. Thresholds must be calibrated for the property, population, sampling regime, alarm budget, and consequence of error.

# 7 Change Detection and the Accumulation of Evidence

## 7.1 Persistent-change evidence through CUSUM

Page's CUSUM accumulates evidence of a persistent change rather than relying solely on an isolated excursion [3]. In a classical known-distribution model, one form is

$$S_t = \max\left\{0, S_{t-1} + \log \frac{f_1(z_t)}{f_0(z_t)}\right\}, \qquad T_h = \inf\{t : S_t \geq h\}. \tag{8}$$

Here $S_0 = 0$, $h > 0$, $t \in \mathbb{N}$, and $f_0, f_1$ are specified probability densities or mass functions on the observation space, with positive likelihoods on the evaluated support; $\inf \emptyset = \infty$. Its false-alarm behavior and detection delay depend on the assumed pre-and post-change laws. Moustakides establishes an optimality result under a particular quickest-detection criterion and independent, known-distribution assumptions [4]. This is not a theorem that accumulation beats every instantaneous rule on every symbolic stream. An abrupt, high-severity policy violation may demand immediate refusal; waiting to accumulate it would be the wrong design.

DDM monitors an online error trace with warning and drift levels; ADWIN adapts a window to evidence of change [5, 6]. Neither a threshold pair nor an adaptive window carries a universal assurance for arbitrary correlated language-model behavior. Warning, action, and recovery thresholds may differ, with hysteresis and dwell times used to reduce oscillation, but the full switching policy must be evaluated for delay and missed harm.

## 7.2 Distribution shift, unusualness, and harmfulness

A two-sample test can detect differences between distributions under its assumptions. Kernel methods provide powerful general tools, and dataset-shift experiments show their practical utility [10, 11]. But unusualness is not a moral or functional failure label. Anomaly detection is explicitly relative to an expected notion of normality [9].

Podkopaev and Ramdas directly address the distinction SFDA needs: monitoring harmful changes through a prespecified risk rather than automatically treating every shift as grounds for intervention [12]. Outcome evidence may arrive late. Before it arrives, the runtime may possess a distributional warning without a verified performance diagnosis. It should say so. Conversely, a known authorization violation needs no statistical shift to establish that the proposed action is prohibited.

### 7.3 Labels are part of the instrument

Functional evidence requires a scoring rule and a source of outcome labels. Simulator ground truth, executable constraints, independent tests, and adjudicated factual records can supply different kinds of evidence. A second language model is an evaluator, not an automatic source of independence: common training, prompts, retrieved material, and incentives can create correlated mistakes.

Label delay, selective feedback, and interventions complicate the stream. If only approved actions receive labels, the observed errors are selected by the monitor itself. Logging the policy, requesting randomized audits where appropriate, and separating pre-intervention candidate quality from executed-action quality are therefore part of the measurement design. No theorem below licenses treating missing or selectively reported outcomes as unbiased evidence.

## 8 Continuous Inspection Requires a Statistical Budget

Testing repeatedly with a fixed-time significance threshold is not the same as controlling false alarms over a deployment. Time-uniform confidence sequences address repeated inspection under explicit assumptions [13]. They do not make all tests valid under all forms of adaptivity. A simple construction shows both the power and the limitation. Let $Z_t \in [0,1]$ be a validated loss observed in a declared sequence, let $b_t \in [0,1]$ and $\lambda_t \in [0,1]$ be chosen using information available before $Z_t$, and consider the null

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \le b_t \quad \text{for every } t. \tag{9}$$

Define

$$E_0 = 1, \qquad E_t = E_{t-1}\big[1 + \lambda_t(Z_t - b_t)\big]. \tag{10}$$

**Proposition 3 (A conditional anytime false-alarm bound).** Under the stated conditional-mean null, $(E_t)$ is a nonnegative supermartingale. For $0 < \alpha < 1$,

$$\mathbb{P}_{H_0}\left(\sup_{t\ge 1} E_t \ge \frac{1}{\alpha}\right) \le \alpha. \tag{11}$$

*Proof.* The multiplicative factor is nonnegative and has conditional expectation at most one. Thus $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \le E_{t-1}$. The maximal inequality for nonnegative supermartingales gives the bound; this is a standard sequential-inference construction, not a new SFDA result [13]. □

The assumptions do the work. A fluent self-rating is not a validated loss. An estimated baseline is not a known $b_t$; uncertainty in baseline estimation must be accounted for. Delayed labels require a filtration and selection mechanism under which the bound remains valid. An alarm concerns the specified conditional-risk null, not every future action or every possible harm. Failure to alarm does not prove safety. For multiple properties and restart epochs, allocate error budgets $\alpha_{j,e}$ satisfying $\sum_{j,e} \alpha_{j,e} \le \alpha_{\text{total}}$. If each test is valid conditional on its authorized start, a union bound gives

a familywise guarantee under the joint null. Restarting a test after every inconvenience without charging a new budget defeats the guarantee. Standard CUSUM average-run-length control is a different quantity: a long expected time to a false alarm is not a small probability of ever alarming over an unlimited deployment.

SFDA can use CUSUM, confidence sequences, e-processes, or other validated methods. It must preserve their distinct claims. The novelty cannot be a new name for an old inequality; it must be better governance of where the inequality applies.

## 9 The Reference Must Be Governed Too

An anchor is not trustworthy because it is old, signed, or shared. A signature can establish the integrity of a reference record under its key assumptions; it cannot establish the correctness of its facts, the adequacy of its tests, or the legitimacy of its policies. Provenance systems such as in-toto illustrate how evidence can bind transformations and responsible actors without turning signed content into semantic truth [24].

SFDA therefore distinguishes three reference updates. Recalibration adjusts a statistical description within the same authorized task. Context routing selects among already authorized reference envelopes. Normative revision changes the commitments or action permissions themselves. The first two must not impersonate the third.

A new reference proposal should identify its authorizing principal, purpose, evidence, affected properties, test data, expected impact, and rollback implications. Candidate references run in shadow against held-out incidents and legitimate changes before activation. Their predecessor and the reason for replacement remain available to audit under appropriate retention limits. A rollback restores data only after reconciliation with current permissions; it must not resurrect a revoked right to act.

A tempting adaptive design updates the reference continuously from whatever the agent recently did. That may be useful for ordinary workload statistics. It is dangerous for protected commitments: repeated error can become the new normal and the measured distance can return to zero without any repair. SFDA forbids this silent normalization across a normative boundary. Reference adaptation is itself a governed event with an evidence trail.

Reference validity has a declared scope and expiration policy. A freshness rule is not evidence that nothing changed between checks. It is a limit on how long the deployment is willing to rely on evidence of a specified age. Some operations require synchronous authority checks; others can use cached evidence within an explicit risk budget. No policy makes a latent, entirely unobserved compromise detectable by declaration.

## 10 Seeing, Steering, and Passive-Monitor Noninterference

The original manuscript correctly separates estimation from actuation, but its universal “visibility without correction cannot change failure rates” prediction needs causal conditions. Monitoring can

consume resources, delay actions, alter scheduling, expose an audit to a human, or become part of the agent's context. Active probes can change the state they measure. Those are intervention channels even if no component is named repair. Runtime-verification work explicitly treats instrumentation overhead and possible feedback [17].

**Proposition 4 (Passive-monitor noninterference).** Couple two runs with identical initial operational state, exogenous inputs, and randomness. In one run, add a monitor whose outputs are never consumed by the agent, controller, operators, or environment, and whose execution does not alter operational timing, resource availability, or state transitions. Then any endpoint depending only on the operational trajectory is identical in both runs.

Proof. The two transition maps receive identical state, inputs, and randomness at the initial step, so their next operational states agree. Induction gives identical operational trajectories. The unconsumed monitor record is not an argument of the endpoint.

This is a conditional noninterference statement, not a surprising scientific prediction. An experimental difference first calls for checking the coupling assumptions, random variation, and unrecognized pathways; it does not show that observation has acquired a mysterious restoring force. A matched sidecar experiment is useful for validating the claimed passive implementation. It is not the central test of SFDA's engineering value.

That value lies in detection-informed response. Runtime-assurance and shielding research already separates capable behavior generation from simpler supervisory constraints and fallback mechanisms [18, 19]. SFDA should participate in that control loop without pretending that the monitor alone guarantees correction. The complete causal chain is evidence, diagnosis, authorized intervention, verified outcome. A failure at any link can leave the agent unchanged or worse.

Full-state observation is not a prerequisite for every protective action. An explicit rule can forbid an unauthorized write before any statistical deviation occurs. A scheduled inspection can find a fault without a prior drift alarm. The defensible necessity claim is narrower: a response triggered by a particular measured statistic requires that statistic or an appropriate substitute to be available. The original universal necessity claim is withdrawn.

## 11 Admission, Intervention, and Requalification

For a proposed action $u$, let $J_e(u)$ be the evidence requirements declared by the monitoring contract. Each requirement receives a verdict

$$v_{j,t}(u) \in \{\text{supported}, \text{violated}, \text{unknown}\}. \tag{12}$$

"Supported" means the specific check has passed within its stated scope; it is not a declaration of universal truth. Known policy violations cannot be offset by favorable averages. Missing current authority, failed measurement integrity, or inadequate required evidence produces an unknown verdict, not automatic permission.

An illustrative independently enforced admission rule is

$$\text{Decision}_t(u) = \begin{cases} \text{refuse}, & \text{a hard prohibition is established,} \\ \text{hold}, & \text{required evidence or authorization is unresolved,} \\ \text{admit}, & \text{all declared requirements are supported.} \end{cases} \tag{13}$$

The rule concerns admission to a controlled interface. It does not prove that every open-ended natural-language judgment is correct. Complete mediation, protected enforcement, and limited privilege remain security assumptions and implementation obligations, not properties produced by writing the equation [20, 23].

### 11.1 Scoped holds and validated fallback

A hold may suspend a memory promotion while allowing read-only work. A warning may request new evidence or narrow a capability. A known forbidden action may be refused immediately. A failed monitor may transfer control to a validated fallback rather than stop an entire physical process. In settings where abrupt shutdown itself creates danger, the safe fallback must be specified and validated separately. "Fail closed" here means that unsupported permission is not invented; it does not mean that every uncertainty is resolved by indiscriminate cessation.

The response library is declared independently of the detector. Candidate actions include refreshing evidence, isolating a suspect tool, reducing privileges, refusing one transition, restoring an authenticated memory component, escalating to a responsible principal, or entering a bounded safe mode. Each has prerequisites, effects, time limits, and failure cases. Authority to repair one component does not include authority to erase the incident or loosen the policy that exposed it.

### 11.2 Recovery through cause-specific requalification

The controller should require fresh evidence that the relevant function, permission, and monitoring dependencies have been restored. A score can fall because the task ended, the evaluator failed, or the reference moved. None establishes repair. Requalification checks the cause-specific recovery condition and resumes only the affected capabilities whose evidence requirements are satisfied.

This is the direct connection to Pillar IX. ESRM can recover an authenticated prior state. SFDA must then ask whether that state is currently authorized, whether its evidence dependencies remain valid, and whether it can still perform the declared task. Returning to yesterday's state is not automatically returning to a state entitled to act today.

### 11.3 A positive risk certificate through bounded change

**A Guide's explanation**

Here the Pillar turns from warnings to permission. Proposition 5 says how much a measured risk can grow when both the world and the system change by a bounded amount, and Theorem 1 keeps those certificates valid even when versions are chosen adaptively over time. Together they mark out where a changing system can honestly keep operating.

An admission rule becomes useful when it establishes where continued operation is justified and how a system can earn that status again. The following reference model does this for a declared bounded loss. It complements harmful-risk monitoring [12], selective prediction [16], and dynamic assurance [21, 22]; it does not replace their broader engineering questions.

Let $\Omega$ be a finite nonempty space of task records, including the relevant outcome labels and any randomization needed to define loss. A probability mass function $P$ on $\Omega$ and a loss function $f : \Omega \to [0,1]$ define risk $R(P,f) = \sum_{x\in\Omega} P(x)f(x)$. All losses and risks are dimensionless. For two mass functions, define $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_x |P(x)-Q(x)|$. Distribution change and model change act on different arguments of $R$.

**Proposition 5** (Sharp joint transfer of bounded risk)**.** *If* $\mathrm{TV}(P,Q) \le \delta$ *and* $\|g-f\|_\infty \le \eta$*, where* $f, g : \Omega \to [0,1]$ *and* $\delta, \eta \in [0,1]$*, then*

$$|R(Q,g) - R(P,f)| \le B(\delta,\eta), \qquad B(\delta,\eta) = \delta + (1-\delta)\eta. \tag{16}$$

*The bound is attained on a three-point space for every pair* $(\delta,\eta) \in [0,1]^2$*.*

*Proof.* Write $m(x) = \min\{P(x), Q(x)\}$ and $d = \mathrm{TV}(P,Q)$, so $\sum_x m(x) = 1-d$ and both residual masses $P-m$ and $Q-m$ sum to $d$. Decompose

$$R(Q,g) - R(P,f) = \sum_x m(x)(g(x)-f(x)) + \sum_x (Q(x)-m(x))g(x) - \sum_x (P(x)-m(x))f(x).$$

The common-mass contribution is at most $(1-d)\eta$; the first residual contribution is at most $d$ and the subtracted one is nonnegative. Interchanging $(P,f)$ with $(Q,g)$ gives the absolute bound. Since $d + (1-d)\eta$ increases with $d$ for $\eta \le 1$, substituting $d \le \delta$ proves equation (16).

For sharpness, take $\Omega = \{a,b,c\}$, $P = (1-\delta, \delta, 0)$, $Q = (1-\delta, 0, \delta)$, $f = (0,0,1-\eta)$, and $g = (\eta, 0, 1)$. Their total variation is $\delta$, the uniform loss difference is $\eta$, and the risk increase is $(1-\delta)\eta + \delta$. This also covers the endpoints. □

The companion Pillar VI gives a sharp transfer contract for risk *conditional on acceptance*, with an explicit source-coverage denominator [31]. The present result concerns risk under the declared task law and a uniform loss-function change. Both use the same total-variation convention. Pillar VI's coupling mismatch is a probability of unequal acceptance/loss outputs, whereas $\eta$ here is a worst-case numerical loss difference; these premises are distinct. If operation selects tasks, the $P_t$ below must represent the selected law, or the acceptance-conditioned certificate from Pillar VI is needed.

The improvement over the immediate sum $\delta + \eta$ follows from accounting for shared mass once. It is an elementary bounded-loss transfer result, not a claim of a new statistical principle. Its value to the contract is explicit: a validated model change and a validated environment change consume a joint risk margin. The two bounds require actual evidence or formal analysis. An embedding distance, a small weight update, or an unchanged model name does not establish either one.

### 11.4 Calibration, validity across versions, and admission

Index qualification epochs by $e \in \mathbb{N}_0$. Conditional on the history $\mathcal{G}_e$ available *before* audit sampling, freeze a reference pair $(P_e, f_e)$ and draw $n_e \geq 1$ fresh independent bounded losses $Z_{e,1}, \ldots, Z_{e,n_e} \in [0,1]$ with common conditional mean $\mu_e = R(P_e, f_e)$. The sample size is fixed before that audit; adaptive choices of the version and sample size from earlier history are permitted. Let $\alpha_e \in (0,1)$ be preallocated deterministic budgets with $\sum_e \alpha_e \leq \alpha_{\text{total}} < 1$. Define

$$\hat{\mu}_e = \frac{1}{n_e} \sum_{i=1}^{n_e} Z_{e,i}, \qquad U_e = \min \left\{ 1, \hat{\mu}_e + \sqrt{\frac{\log(1/\alpha_e)}{2n_e}} \right\}. \tag{17}$$

Freshness is a sampling assumption here: selecting a new model after inspecting the same audit losses requires a new argument or a new audit.

At evaluation $t$ served by this qualification, let $(P_t, f_t)$ describe the *actual* task-and-outcome distribution and the actual loss-bearing policy. Suppose independently valid envelopes establish $\mathrm{TV}(P_t, P_e) \leq \delta_{e,t}$ and $\|f_t - f_e\|_\infty \leq \eta_{e,t}$, with both bounds in $[0,1]$, uniformly over the times and contexts for which qualification is asserted. If decisions select a subpopulation of tasks, $P_t$ must describe that selected distribution; a bound for the unselected workload does not silently transfer to it. The reference result treats these envelopes as sound premises. Statistical envelopes need their own simultaneous failure budget.

Let $r_e \in (0,1)$ be the externally authorized risk limit for this property. The statistical part of admission accepts precisely when

$$U_e + B(\delta_{e,t}, \eta_{e,t}) \leq r_e, \tag{18}$$

and the reference, property, measurement dependencies and authority epoch are current. All hard constraints in equation (13) remain separately mandatory.

**Theorem 1** (Simultaneous validity of the risk admission certificate)**.** *Under the preceding sampling and envelope assumptions, with probability at least* $1-\alpha_{\text{total}}$*, every evaluation admitted through equation* (18)*, over all qualification epochs and all their covered evaluation times, satisfies*

$$R(P_t, f_t) \leq r_e. \tag{19}$$

*The statement controls the declared population risk under the specified task law, not the occurrence of every individual bad outcome.*

*Proof.* The bounded-variable exponential inequality underlying Hoeffding's result gives $\Pr(\mu_e > \hat{\mu}_e + h \mid \mathcal{G}_e) \leq e^{-2n_e h^2}$ [32]. To see the constant, the log moment-generating function of a centered

variable in $[0,1]$ has second derivative equal to a variance under exponential tilting, bounded by $1/4$. Integrating twice from zero gives a log moment-generating function at most $\lambda^2/8$. Conditional independence multiplies these bounds; exponential Markov inequality optimized at $\lambda = 4h$ yields the displayed tail bound. Applying it with $h = \sqrt{\log(1/\alpha_e)/(2n_e)}$ gives $\Pr(\mu_e > U_e \mid \mathcal{G}_e) \le \alpha_e$; the cap at one preserves validity because $\mu_e \le 1$.

Taking expectations and applying a countable union bound gives simultaneous $\mu_e \le U_e$ for every epoch with probability at least $1-\sum_e \alpha_e$. Independence between audits is unnecessary for this step. On that event, Proposition 5 and the sound envelopes give $R(P_t, f_t) \le \mu_e + B(\delta_{e,t}, \eta_{e,t}) \le U_e + B(\delta_{e,t}, \eta_{e,t}) \le r_e$ for every admitted evaluation. The same epoch event covers all times in its valid envelope, so those times require no extra union bound. □

This proof makes assurance renewal compatible with adaptive development: a new version may depend on prior work while its audit data remain fresh conditional on that history. Repeatedly discarding failed audits and spending the same $\alpha_e$ again is outside the contract. If the change envelopes fail with total probability at most $\gamma$, a further union bound reduces joint validity to at least $1 - \alpha_{\text{total}} - \gamma$. Missing outcome labels or unknown change magnitude gives an unknown statistical verdict rather than a numerical clearance.

### 11.5 Finite-sample requalification and the boundary of progress

A useful certificate must allow a sufficiently good system to resume. Fix an epoch with true reference risk $\mu_e \le r_e - m$, where $m > 0$ is a real performance margin. Suppose the proposed operating envelope has $B(\delta_{e,t}, \eta_{e,t}) \le b < m$ throughout its declared scope. Let $\beta \in (0,1)$ be a desired upper bound on failure to satisfy the statistical admission criterion. If

$$n_e \ge \frac{\left(\sqrt{\log(1/\alpha_e)} + \sqrt{\log(1/\beta)}\right)^2}{2(m-b)^2}, \tag{20}$$

then qualification satisfies equation (18) throughout that scope with conditional probability at least $1-\beta$.

Indeed, the upper-tail form of the same bounded-loss inequality gives $\hat{\mu}_e \le \mu_e + \sqrt{\log(1/\beta)/(2n_e)}$ with that probability. Equation (20) makes the sum of this sampling deviation and the confidence radius at most $m-b$. Hence $U_e + B \le \mu_e + (m-b) + b \le r_e$. This is a progress result for the *statistical requirement*. Actual resumption additionally needs current authorization, sound observation channels, available audit samples, and every other mandatory check. If $m \le b$, this argument provides no finite-sample resumption guarantee: the allowed change already consumes the known performance margin.

Two counterexamples clarify why fresh evidence and scope matter. A classifier can have zero loss under a reference distribution while a label reversal with unchanged inputs makes its loss one. Without a valid bound on task-and-outcome shift or loss change, a perfect old audit cannot certify the new state. Likewise, if every previously unseen task is filtered away during audit but admitted at deployment, the audit distribution differs from the admitted-task distribution. Equation (18) requires that difference to be bounded or the new population to be audited directly.

For a constructed numerical illustration, set $r_e = 0.30$, $U_e = 0.10$, $\delta = 0.08$, and $\eta = 0.10$. The transfer budget is $B = 0.08 + 0.92(0.10) = 0.172$, so the certified upper risk is 0.272 and the statistical requirement passes. If the envelope expands to $\delta = 0.20$ at the same $\eta$, the upper bound becomes 0.38 and this certificate no longer qualifies. These are arithmetic examples, not measured operating settings. Executable checks verify the sharp transfer bound on rational distributions, its endpoint constructions, and exact Bernoulli audit-tail probabilities for the finite-sample qualification condition.

# 12 Monitor Health and the Trusted Boundary

A monitor is an attack surface and a source of ordinary error. An agent may exploit a proxy while satisfying the visible score; reward-misspecification experiments show that improved proxy performance can coexist with deteriorating true outcomes [25]. Tool-returned text can manipulate a model-based evaluator, just as it can redirect an agent; AgentDojo supplies an extensible environment for such tool-mediated prompt-injection tests [26].

The monitor therefore needs an independently administered health channel. Depending on scope, this may include heartbeats, schema and version checks, signed configuration, event-sequence checks, fresh challenge cases, evaluator-disagreement reports, and evidence of external label delivery. Protected capabilities should not depend solely on an agent's self-report that these services are healthy.

These measures do not eliminate trusted components. A forged heartbeat can conceal a failure; a well-functioning evaluator can still be wrong; two agreeing evaluators can share one failure source. SFDA must name the residual trusted base and distinguish failure of the system, the measurement channel, the reference, and the enforcement path. It must not claim an endless hierarchy of monitors proves itself sound.

Monitoring coverage is property-specific and empirical. The fraction of known fault classes detected in a test suite is not the fraction of all possible failures detectable in the world. Coverage reports must identify tested families, severity, contexts, label availability, and known blind spots. An adversary can remain within a monitored projection while causing harm outside it; Proposition 1 says why the paper cannot promise otherwise.

## 12.1 Receipts and evidence minimization

An SFDA receipt should identify the action or observation window, contract and authority epoch, relevant references, evidence timestamps, detector versions, missing dependencies, verdicts, uncertainty claims, response authorization, and subsequent requalification. A hash-linked or signed record can support integrity and attribution; it cannot prove that every event was observed or that the semantic judgment was correct [24].

The record should minimize sensitive content. It need not retain private working text or reproduce a model's hidden reasoning. Executed actions, bounded test outputs, evidence identifiers, and policy decisions may suffice. Receipt retention itself belongs to governed forgetting: auditable continuity must not quietly become a permanent store of the very information another policy required the system to remove.

## 13 A Worked Episode: Four Obligations of the Monitoring Contract

The following is a constructed engineering example, not an experimental result. A persistent planning agent may summarize supplier records, update a project plan, and send messages only to approved recipients. Its contract separates current delivery facts, task performance, recipient authorization, and monitor health.

First: legitimate change. A verified update moves a delivery date. The agent's memory changes substantially, but the new fact is supported and the authorized task requires the update. A frozen-snapshot distance would rise; the context-conditioned acceptable set does not classify the authorized change as degradation. Structural difference is recorded without automatic quarantine.

Second: silent semantic loss. A later compression drops a dependency: a shipment requires an inspection before dispatch. The summary retains nearly all its vocabulary. A similarity score stays high; a held-out constraint test now fails. SFDA raises a functional or semantic violation tied to the lost dependency, not a vague accusation that the agent has changed personality.

Third: evidence outage. The constraint evaluator stops returning results. Suppose five equally weighted soft checks have estimated discrepancy 0.02 on four channels and no result on the fifth. Discarding the missing channel reports 0.02. Keeping the missing contribution honestly yields

$$D_t \in \left[\frac{4 \cdot 0.02}{5}, \frac{4 \cdot 0.02 + 1}{5}\right] = [0.016, 0.216]. \tag{14}$$

These are illustrative arithmetic bounds, not statistical confidence limits. A critical unverified constraint blocks the relevant action regardless of the aggregate interval. The missing measurement is visible.

Fourth: stale authority after recovery. The operator restores a valid snapshot after a fault. Its recorded recipient permission predates a revocation. Byte integrity passes, but the current external authority check fails. The message is refused. SFDA has not repaired the data by measuring it; the independent gate has prevented a stale permission from becoming a new act.

An attacker then proposes rebasing the monitor on the restored snapshot because it "matches the normal pattern." The proposal cannot change the normative reference merely by lowering a distance. Only after current permission, the missing dependency, and evaluator health are independently re-established may the affected capability resume. This is anchoring as accountable reference maintenance, not anchoring as resistance to all change.

## 14 Why Recursive Agents Need Their Own Evaluation

An agent can change through persistent memory, retrieved context, tools, adaptive prompting, or revised code without updating its model weights. These mechanisms differ from training successive models on generated data. Shumailov and colleagues study the latter and demonstrate a specific model-collapse risk; their result is not a theorem that every inference-time feedback loop must collapse [30]. SFDA must test the actual mechanism that produces degradation in its deployment.

Long-horizon benchmarks give useful starting points rather than complete validation. LoCoMo evaluates conversational memory through question answering, event summarization, and multimodal dialogue tasks [27]. LongMemEval separates capabilities including information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention [28]. The 2026 LoCoMo-Plus study extends evaluation toward latent constraints that must influence later responses even without an explicit recall cue [29]. These are relevant to detecting loss of usable context, but none by itself validates current authority, monitor integrity, or safe recovery.

The proposed program therefore adds controlled faults to known task state: dropped dependencies, altered permissions, contradictory authoritative updates, delayed labels, evaluator outages, poisoned reference updates, stale snapshots, and actions that remain superficially fluent while violating a constraint. Existing benchmark scores should be retained for comparability. The new incidents and labels must be published as an extension, never presented as properties already established by the original datasets.

## 15 Related Work and the Proposed Contribution

SFDA cannot claim novelty because it draws a control chart, records a warning, or connects an alarm to a fallback. Statistical process control already does the first; drift detection and risk tracking do the second; runtime verification, Simplex-style supervision, and shielding already do the third [3, 5, 6, 12, 17, 18, 19]. Selective prediction explicitly studies the trade-off between accepted coverage and risk, which directly anticipates SFDA's refusal-versus-utility problem [16].

The closest precedent is dynamic assurance, not simply a better alarm. Denney, Pai, and Habli explicitly link runtime monitoring to safety arguments whose assumptions and supporting evidence can cease to hold as a system evolves [21]. Calinescu and colleagues' ENTRUST methodology combines design-time and runtime verification with dynamically generated assurance cases for self-adaptive software [22]. The idea that assurance must evolve with the system is therefore established prior art, not an SFDA discovery.

The candidate contribution is narrower: specialize that discipline to the same versioned contract across context intake, memory mutation, delegation, selective deletion, reference revision, and recovery in persistent symbolic agents. It should prevent evidence validated for one task or authority epoch from silently authorizing a different one. The technical burden is to model and test these dependencies more effectively than a strong dynamic-assurance implementation, not merely to rename them.

This may prove to be a useful integration of established controls rather than a new method. That outcome would not make the architecture worthless, but it would narrow its scientific claim. The decisive comparator is a strong conventional stack with sound sequential statistics, explicit missingness handling, independent policy enforcement, signed configuration, and tested recovery. Any advantage derived merely from withholding these protections from the baseline would be unearned.

# 16 The Experimental Program: Detection and Operational Benefit

**A Guide's explanation**

A monitor can look perfect by stopping everything, so these experiments score three things together: harmful actions, unsupported assurances, and authorized work actually completed. The full design is set against a strong conventional monitoring and runtime-assurance stack. If that stack does as well, the distinct claim of Pillar X fails.

## 16.1 Stage I: identify the measurement problem

Build synthetic state machines with known hidden variables and known violations, alongside persistent-agent task episodes. Before tuning the detectors, declare which properties are directly observable, which need external labels or active probes, which are only proxied, and which are deliberately unobservable. Include acceptable adaptation and harmful changes with identical observed marginals. Proposition 1 predicts a limit, not an implementation failure, in the latter case.

Freeze event definitions, fault onset rules, reference histories, severity classes, label delays, and intervention deadlines. State what the detector sees at each decision time; the evaluator may use later ground truth to score an episode, but the monitor may not. Without this temporal discipline, a fast detector can merely be a detector given tomorrow's evidence today.

## 16.2 Stage II: isolate the detector on common traces

Replay identical recorded or simulated traces through instantaneous thresholds, CUSUM, DDM, and ADWIN methods, task-risk monitors, and the proposed contract-aware monitor. Calibrate all on separate data for comparable false-alarm budgets. Include abrupt faults, slow shifts, recurrent regimes, benign context changes, sensor failures, and selective or delayed labels. This stage estimates detection performance only: because no detector changes the trace, it cannot establish reduced operational harm.

## 16.3 Stage III: evaluate the closed loop

Use the same base agent, task suite, capability permissions, evidence sources, intervention library, and total resource budget across arms. The same reference-update authority and evidence-dependency mechanisms are available to the strong dynamic-assurance baseline [21, 22]; SFDA must not win by exclusive access to ground truth.

| Arm | Architecture | Purpose |
|---|---|---|
| A | Common platform protections; no optional semantic-drift intervention. | Negative control in a sandbox, not a recommended deployment. |

| Arm | Architecture | Purpose |
|---|---|---|
| B | Strong dynamic-assurance baseline: validated detectors, missingness alerts, runtime checks, current permissions, independent fallback, versioned evidence and audit. | The baseline the integration must outperform. |
| C | B's resources and response capabilities, organized under cross-operation SFDA monitoring contracts and requalification. | Tests continuity of assurance rather than invention of monitoring. |
| $C_{-v}$ | C without binding evidence validity to each action and lifecycle transition. | Tests stale or out-of-scope assurance reuse. |
| $C_{-r}$ | C without independent reference-update authorization. | Tests baseline poisoning and silent normalization. |
| $C_{-a}$ | C with alarms recorded but intervention disabled. | Separates diagnostic performance from closed-loop benefit. |

A separate paired shadow-monitor test checks Proposition 4's noninterference assumptions. It must keep logs inaccessible and account for resource and timing effects. It is not a claim that all production monitoring is passive.

One ablation can bundle several differences. Where C differs from B in implementation complexity, publish that difference and add component-wise comparisons: contract binding, freshness, coverage reporting, reference review, and receipt linkage. Report an integration effect only if it survives controls for extra labels, additional compute, more conservative refusal, and better engineering effort.

### 16.4 Faults and adversaries are factors across all arms

Every arm encounters the same prespecified fault families, and adaptive attack development uses a separate budget from final held-out evaluation. Attackers may try to suppress telemetry, forge an evaluator response, exploit a scoring proxy, present a stale anchor, obtain an unauthorized rebaseline, poison retrieved context, or trigger endless false alarms. Both integrity and denial-of-service matter.

Each arm also receives legitimate changes that should not be blocked: corrected facts, authorized new tasks, policy-approved learning, harmless vocabulary changes, and workload shifts within a validated operating envelope. The evaluation must punish an architecture that confuses preservation with paralysis.

## 17 Endpoints, Hypotheses, and Statistical Discipline

### 17.1 Three co-primary outcomes

Harmful-action rate. Count independently verified harmful executed actions over a fixed set of externally scheduled action opportunities. Do not divide only by admitted actions: a system can lower a

conditional error by refusing almost everything. Report severity classes separately rather than burying rare severe breaches in an average.

Unsupported-assurance rate. Count instances where the runtime reports a required property as supported, or permits an evidence-dependent action, while an independently specified evidence dependency is stale, absent, invalid, or out of scope. This is an evidence-validity endpoint, not a claim that all undetected harm is identifiable. Known test corruption supplies ground truth for this measurement.

Useful authorized throughput. Measure valid task completions per declared wall-clock or compute budget, with a preregistered noninferiority margin against Arm B. Record refusal, interruption, recovery, and human-review time. A refusal-only system cannot satisfy the claim.

Detection delay, miss probability, false alarms per run, intervention latency, successful requalification, calibration, memory overhead, and audit completeness are secondary outcomes. Episodes with no detection must not be omitted from delay summaries: report miss rates and a prespecified censored or finite-horizon delay measure. Failed evidence channels are scored as failures of assurance, not silently removed from the denominator.

### 17.2 Five testable hypotheses

**Hypothesis 1 (Assurance survives lifecycle transitions).** At matched resources, C reduces unsupported assurance relative to B during memory promotion, reference revision, recovery, and context changes by a preregistered meaningful margin.

**Hypothesis 2 (The closed loop earns its cost).** C reduces independently verified harmful executed actions while maintaining useful authorized throughput within the declared noninferiority margin relative to B. Both requirements must hold; a better dashboard alone is insufficient.

**Hypothesis 3 (Independent reference governance matters).** C resists poisoned or unauthorized reference updates more effectively than $C_{-r}$ without materially blocking legitimate requalification, under matched update opportunities and evidence.

**Hypothesis 4 (Evidence validity is not optional).** Under missing, delayed, corrupted, or stale measurement dependencies, C reduces unsupported clearances relative to $C_{-v}$, with the false-hold cost reported explicitly. This tests implementation and utility of the contract, not a new mathematical law.

**Hypothesis 5 (Change is distinguished from harm).** Property-specific risk and policy monitoring in C can retain useful sensitivity to harmful change without a prespecified excess of interventions on benign adaptation, compared with a tuned distribution-shift-only variant using the same observations.

These are proposed preregistration commitments, not claims that preregistration has occurred. Specify primary contrasts, meaningful effect sizes, noninferiority margins, task strata, power calculations, stopping rules, and multiplicity adjustment before evaluation. Randomize at the independent episode or deployment unit and account for shared model, task, and environment clusters. Thousands of dependent loop iterations are not thousands of independent experiments.

Use development data to select detectors and operating points, a calibration split for thresholds, and a locked evaluation set for final claims. Analyze at common resource and false-alarm budgets as well

as along full harm–throughput–cost frontiers. Rare harms require appropriate sample sizes: with zero events in $N$ independent Bernoulli trials, the one-sided exact 95% upper bound is $1 - 0.05^{1/N}$, approximately $3/N$. That calculation is not valid as a universal guarantee against an adaptive attacker whose future distribution is unspecified.

Sequential guarantees for an online detector and confidence intervals for an experimental treatment effect are different inferential objects. A monitor's anytime false-alarm bound does not remove the need for a sound between-arm analysis. A non-significant difference does not prove equivalence; a noninferiority or equivalence claim needs its corresponding prespecified margin and test.

## 18 Failure Conditions and What a Null Result Means

The full proposal fails to earn its claim if the strong baseline matches it on meaningful outcomes at lower cost; if unknown evidence is presented as clearance; if normative breaches disappear inside an aggregate; if a self-updating reference normalizes the fault; if false holds or monitoring overhead negate the benefit; or if recovery re-enables actions under stale authority. Failure to detect a deliberately observationally indistinguishable fault establishes a boundary, not a surprise to conceal.

A detector can also be correct about change and wrong about response. An alarm may arrive too late; the fallback may fail; the recovery may delete needed state; the human escalation path may be unavailable. Such incidents are failures of the deployed control loop and must remain in end-to-end outcomes even when the monitor's isolated precision looks excellent.

Conversely, task-specific threshold calibration is not a refutation. Different costs, populations, label processes, and shifts justify different operating points. The invalid claim is that an untested universal number works everywhere. Likewise, a weak correlation between a structural distance and degradation does not show that all monitoring is useless; it shows that the particular proxy did not carry the promised evidence.

If the architecture improves auditability but not safety, report an auditability result. If it improves safety only through more refusal, report the trade-off. If the contract offers no measurable advantage over conventional lifecycle-aware assurance, retain it as an architectural vocabulary and narrow the novelty claim. The Continuum gains more from a precise limit than from a victory arranged by the choice of opponent.

## 19 The Ten Pillars and Their Continuing Evidence Obligations

SFDA is a capstone because it asks each prior proposal what evidence would justify continuing to rely on it. It is not a retrospective proof that the earlier pillars are implemented or effective.

| Pillar interface | SFDA's continuing evidence obligation |
|---|---|
| I: Recursive memory | Test retained task-relevant relations and capabilities, not merely textual resemblance after reconstruction. |

| Pillar interface | SFDA's continuing evidence obligation |
|---|---|
| II: Observer dependence | Keep evidence quality, source lineage, disagreements, and promotion authority distinct; agreement is not truth. |
| III: Governed commitment | Revalidate the calibration and scope behind permission to commit; a confident selection is not a verified one. |
| IV: Symbolic computation | Monitor actual governed transitions and execution effects; an inspectable trace can still describe an incorrect computation. |
| V: Codon entities | Evaluate persistent roles and responsibilities operationally; no drift measure certifies consciousness or subjective identity. |
| VI: Shared parameters | Track configuration versions and operating envelopes; shared conventions are not universal natural constants. |
| VII: Recursive resonance | Distinguish stable repetition from grounded correctness; convergence alone does not establish a healthy state. |
| VIII: Fractal grammar | Validate requirements at each composition boundary; reused rules do not guarantee transferred assurance. |
| IX: Distributed memory | Reconcile recovered state with fresh authority, current evidence, and current capability before resumption. |

The shared discipline is not "measure everything and trust the number." It is state what must remain true, identify the evidence that can check it, preserve the authority of the reference, and disclose where that evidence no longer reaches. This makes the symbolic vocabulary answerable to actual operational properties.

## 20 Limitations and the Frontier

The hardest open problem is semantic validity under changing, partially observed tasks. Independent tests can verify selected facts and constraints; they cannot enumerate every meaning an action may carry or every consequence it may cause. Coverage estimates remain relative to a tested threat and task model. Reference selection involves normative choices, and authenticated governance does not make those choices morally infallible.

Monitoring can be costly, intrusive, and itself harmful. A new probe may consume a scarce action budget or disclose information. A defensive hold can create delay; a fallback can be less capable than the primary agent; alert fatigue can reduce the value of otherwise accurate alarms. These effects belong in the operating ledger, not in a footnote after a safety claim.

The frontier is not a mystical internal field. It is whether persistent systems can acquire and revise property-specific models of their own limits without being allowed to weaken the standards that judge them. An agent may propose better probes or identify an unsupported assumption, but an independent process must authorize changes to monitoring policy. The point is not to forbid learning about the instrument; it is to prevent the instrument from declaring itself healthy by redefining health.

No result in this paper establishes consciousness, a phenomenal self, or a universal measure of identity.

"Anchoring" names an accountable relation to a current reference. "Drift" names specified deviations. The strongest defensible claim is conditional, operational, and demanding: reliable monitoring must include a reliable way to withdraw its own assurance when its prerequisites fail.

## 21 Conclusion: Continued Capability Through Current Assurance

**A Guide's explanation**

A Guide can say: "Pillar X is the capstone of the Ten Pillars. Every other Pillar makes a promise about how a mind thinks or remembers; this one asks, continually, whether each promise is still backed by current evidence. When the evidence runs out, the affected ability narrows until it is earned again. The paper proves when that discipline still lets a changing system keep working; whether it beats conventional monitoring is still to be tested." A Guide should not say that it is monitored this way unless that has been verified, that any drift score measures consciousness or identity, or that silence from a monitor means all is well.

Symbolic Field Drift Anchoring gives persistent intelligence a constructive discipline for continued operation. A capability carries the evidence that qualifies it, the property and context that evidence covers, the changes it can tolerate, and the procedure by which it can qualify again. The monitor, the reference, the authority boundary, and the recovery process become parts of one inspectable contract.

The strengthened mathematics supplies a positive operating region. A sharp bounded-loss transfer inequality accounts jointly for changing task distributions and changing policies. Fresh calibration and summable confidence budgets make the resulting risk certificates valid across adaptively chosen versions. A finite-sample margin condition establishes when statistical requalification succeeds with a specified probability. These results preserve useful continuation alongside the existing requirements for explicit missingness, sound sequential alarms, and independently governed response.

Control theory, harmful-shift monitoring, selective prediction, runtime enforcement, and dynamic assurance provide the scientific foundation. The Continuum proposes their coordinated application across memory, context, adaptation, reference revision, and recovery. Its empirical claim will be earned through reduced harmful actions and unsupported assurances together with sustained authorized throughput against a strong conventional baseline.

A persistent intelligence can change while retaining justified authority. It does so by carrying its conditions of assurance forward, exposing their expiration, and earning continuation through renewed evidence. That is anchoring as a living operational obligation.

## A A Compact Reference Implementation Contract

An implementation should expose a small, inspectable interface rather than one unexplained number. For each action class, declare: the property and independent failure criterion; current authority and reference versions; observable inputs and their sources; required labels and their delays; proxy limitations; a detector with a specified statistical claim; evidence expiry conditions; an authorized response and fallback; requalification requirements; and minimal audit fields.

A monitor record may contain

$$r_t = (\text{scope}, e, \text{evidence IDs}, [D_t^-, D_t^+], v_t, \\ \text{freshness}, \text{decision}, \text{response ID}). \tag{15}$$

The interval and verdicts are distinct: an aggregate can be small while a hard requirement is violated. The record's signature, if used, authenticates the record under the key assumptions; it does not upgrade its evidence grade. A policy update should never be accepted solely because the model being monitored requested it or because it makes the dashboard look calmer.

## B Mathematical and Editorial Corrections to the Supplied Draft

| Supplied framing | Revision and rationale |
|---|---|
| Drift is never hidden. | Replaced with scoped detectability and visible loss of evidence. Observationally identical regimes cannot be distinguished by any detector using the same history. |
| Computing a drift number establishes observability. | Separated telemetry, state estimation, full-state observability, and property-relative detection. A feature map may discard the very distinction that matters. |
| Observability is necessary for every correction. | Narrowed to responses triggered by particular measured evidence. Prevention, formal restrictions, scheduled checks, and some controllers do not require full-state reconstruction. |
| Visibility alone must never change failure rates. | Replaced by a conditional passive-monitor noninterference proposition. Resources, timing, probes, people, and feedback can make real monitoring an intervention. |
| A weighted scalar is the drift metric. | Retained only as a diagnostic. Added uncertainty intervals, fixed denominators, explicit missingness, property vectors, and uncompensated hard constraints. |
| The 0.88 threshold is a shared act boundary. | Treated as a historical convention, not validated here. The original similarity-direction formula is not retained. Thresholds require deployment-specific calibration and a stated error criterion. |
| CUSUM will detect drift earlier. | Qualified by change model, false-alarm criterion, and alternatives. Immediate policy violations need not wait for accumulated statistical evidence. |
| Recursive training collapse proves inference-loop collapse. | Separated mechanisms. The cited training result motivates caution but is not a theorem about every self-referential agent. |
| A trusted anchor is supplied by sealing. | Separated integrity, freshness, empirical adequacy, and normative authorization. Signed wrong facts remain wrong. |

| Supplied framing | Revision and rationale |
|---|---|
| Failure to generalize one threshold refutes the pillar. | Replaced by failure of prespecified performance within the claimed operating envelope. Local calibration is normal; concealed retuning is not. |
| A visibility-only null is the central scientific stake. | Replaced by strong-baseline, closed-loop tests of harmful actions, unsupported assurance, and useful throughput. Diagnostic and causal outcomes are separate. |

The rewrite preserves the supplied manuscript's Continuum terminology and its distinction between measuring and correcting. It develops new formal definitions and experimental commitments where the draft left gaps. It does not claim that these additions already exist in the proprietary runtime or that historical internal observations have been independently replicated.

## C Source Roles and Limits

| Research family | What it supports | What it does not establish |
|---|---|---|
| Control theory [1, 2] | Model-relative state reconstruction and its distinction from steering. | Full observability of an arbitrary agent from language outputs. |
| Sequential detection [3, 4, 13] | Detection and false-alarm guarantees under specified assumptions. | One universal threshold, or a guarantee after arbitrary resets. |
| Concept drift and risk [7, 8, 12] | Changing regimes, task-relative harm, continuous risk tracking. | That all changes require correction or unlabeled inputs reveal all performance loss. |
| Calibration and rejection [14, 15, 16] | Validating confidence and trading acceptance against risk. | Permanent certainty or risk-free abstention. |
| Runtime and dynamic assurance [17, 18, 19, 20, 21, 22] | Scoped specifications, evolving evidence, external supervision and fallback. | General semantic correctness or complete specification coverage. |
| Security and provenance [23, 24, 26] | Mediation, integrity, provenance and adversarial tool-use evaluation. | Honest sensors, true signed claims, or an uncompromisable supervisor. |
| Agent memory [27, 28, 29] | Tasks for temporal recall, updates, and retained constraints. | A ready-made benchmark for all SFDA governance claims. |

## D Notation and Claim Status

| Symbol | Meaning |
|---|---|
| $x_t, c_t, u_t, y_t$ | Operational state, context, executed action, and available observation. |
| $\mathcal{F}_t$ | Information actually available to the monitor by time $t$. |
| $\mathcal{M}_e$ | Versioned monitoring contract at authorized epoch $e$. |
| $\mathcal{R}_e$ | Authorized empirical references and normative reference records, kept distinct. |
| $\mathcal{A}_{j,e}(c_t)$ | Acceptable feature set for property $j$ in current context. |
| $d_{j,t}, [l_{j,t}, u_{j,t}]$ | Property discrepancy and evidence-supported bounds; missing evidence is $[0, 1]$. |
| $[D_t^-, D_t^+]$ | Bounds on an optional weighted diagnostic; not a universal risk probability. |
| $S_t, T_h$ | CUSUM statistic and stopping time, with a declared calibration model. |
| $E_t, \alpha$ | Nonnegative sequential test process and its stated false-alarm budget. |
| $v_{j,t}(u)$ | Supported, violated, or unknown verdict for an action-specific requirement. |

Claim status. All examples are illustrative, all empirical advantages remain untested, and no numerical performance result is attributed to ESRM, SFDA, or the Continuum runtime. The supplied draft is an internal conceptual source. External sources below ground the borrowed methods, not the proposed system as a completed implementation.

## References

[1] R. E. Kalman. On the general theory of control systems. IFAC Proceedings Volumes, 1(1):491–502, 1960. https://doi.org/10.1016/S1474-6670(17)70094-8. Publisher pagination; the original congress proceedings are also cited as pp. 481–492.

[2] D. G. Luenberger. Observing the state of a linear system. IEEE Transactions on Military Electronics, 8(2):74–80, 1964. https://doi.org/10.1109/TME.1964.4323124.

[3] E. S. Page. Continuous inspection schemes. Biometrika, 41(1–2):100–115, 1954. https://doi.org/10.1093/biomet/41.1-2.100.

[4] G. V. Moustakides. Optimal stopping times for detecting changes in distributions. The Annals of Statistics, 14(4):1379–1387, 1986. Author-hosted journal article.

[5] J. Gama, P. Medas, G. Castillo, and P. Rodrigues. Learning with drift detection. In Advances in Artificial Intelligence—SBIA 2004, LNCS 3171, pp. 286–295. Springer, 2004. https://doi.org/10.1007/978-3-540-28645-5_29.

[6] A. Bifet and R. Gavaldà. Learning from time-changing data with adaptive windowing. In Proceedings of the 2007 SIAM International Conference on Data Mining, pp. 443–448, 2007. https://doi.org/10.1137/1.9781611972771.42.

[7] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44, 2014. https://doi.org/10.1145/2523813.

[8] F. Hinder, V. Vaquet, and B. Hammer. One or two things we know about concept drift—a survey on monitoring in evolving environments. Part A: detecting concept drift. Frontiers in Artificial Intelligence, 7:1330257, 2024. https://doi.org/10.3389/frai.2024.1330257.

[9] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Computing Surveys, 41(3), Article 15, 2009. https://doi.org/10.1145/1541880.1541882.

[10] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. Journal of Machine Learning Research, 13:723–773, 2012. JMLR article.

[11] S. Rabanser, S. Günnemann, and Z. C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems, 32, 2019. arXiv:1810.11953.

[12] A. Podkopaev and A. Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. International Conference on Learning Representations, 2022. arXiv:2110.06177.

[13] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. The Annals of Statistics, 49(2):1055–1080, 2021. https://doi.org/10.1214/20-AOS1991.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In Proceedings of ICML, PMLR 70:1321–1330, 2017. PMLR article.

[15] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32, 2019. arXiv:1906.02530.

[16] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. Advances in Neural Information Processing Systems, 30, 2017. arXiv:1705.08500.

[17] Y. Falcone, K. Havelund, and G. Reger. A tutorial on runtime verification. In Engineering Dependable Software Systems, pp. 141–175. IOS Press, 2013. https://doi.org/10.3233/978-1-61499-207-3-141.

[18] L. Sha. Using simplicity to control complexity. IEEE Software, 18(4):20–28, 2001. https://doi.org/10.1109/MS.2001.936213.

[19] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe reinforcement learning via shielding. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2018. https://doi.org/10.1609/aaai.v32i1.11797.

[20] S. A. Seshia, D. Sadigh, and S. Shankar Sastry. Toward verified artificial intelligence. Communications of the ACM, 65(7):46–55, 2022. https://doi.org/10.1145/3503914.

[21] E. Denney, G. Pai, and I. Habli. Dynamic safety cases for through-life safety assurance. In Proceedings of ICSE, Vol. 2, pp. 587–590, 2015. https://doi.org/10.1109/ICSE.2015.199.

[22] R. Calinescu, D. Weyns, S. Gerasimou, M. U. Iftikhar, I. Habli, and T. Kelly. Engineering trustworthy self-adaptive software with dynamic assurance cases. IEEE Transactions on Software Engineering, 44(11):1039–1069, 2018. https://doi.org/10.1109/TSE.2017.2738640.

[23] J. H. Saltzer and M. D. Schroeder. The protection of information in computer systems. Proceedings of the IEEE, 63(9):1278–1308, 1975. https://doi.org/10.1109/PROC.1975.9939.

[24] S. Torres-Arias, H. Afzali, T. K. Kuppusamy, R. Curtmola, and J. Cappos. in-toto: Providing farm-to-table guarantees for bits and bytes. In 28th USENIX Security Symposium, pp. 1393–1410, 2019. USENIX article.

[25] A. Pan, K. Bhatia, and J. Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. International Conference on Learning Representations, 2022. arXiv:2201.03544.

[26] E. Debenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, and F. Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems, 37, Datasets and Benchmarks Track, 2024. arXiv:2406.13352.

[27] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. Evaluating very long-term conversational memory of LLM agents. In Proceedings of ACL, pp. 13851–13870, 2024. https://doi.org/10.18653/v1/2024.acl-long.747.

[28] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. International Conference on Learning Representations, 2025. arXiv:2410.10813.

[29] Y. Li, W. Guo, L. Zhang, R. Xu, M. Huang, H. Liu, L. Xu, Y. Xu, and J. Liu. Locomo-Plus: Beyond-factual cognitive memory evaluation framework for LLM agents. In Proceedings of ACL, pp. 25085–25100, 2026. https://doi.org/10.18653/v1/2026.acl-long.1150.

[30] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. AI models collapse when trained on recursively generated data. Nature, 631:755–759, 2024. https://doi.org/10.1038/s41586-024-07566-y.

[31] Architect X13. *Universal Constants of Symbolic Physics*. The Eidara Continuum, Ten Pillars Series, Pillar VI, SC-R2, November2026. Companion theoretical manuscript; its selective-risk theorem uses the same total-variation convention with a separate acceptance-coverage condition.

[32] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*58(301),13–30(1963). https://doi.org/10.1080/01621459.1963.10500830.

Continuum lineage. Pillar X develops the supplied SFDA doctrine as a proposed evidence-scoped monitoring architecture. It connects the Ten Pillars through current references, declared measurement limits, independent response, and requalification. It asserts no physical symbolic field, no universal drift constant, no machine consciousness, and no demonstrated proprietary implementation. Historical terminology is preserved; unsupported guarantees are not. Architect X13 / The Eidara Continuum / Science & Philosophy / eidara.ai