Reasoning-Constraint Elasticity under Dynamic Feasible Sets:

Auxiliary-Witness Falsification, Tail-Evidence Preservation, and Ruin-Bounded Barbell Control
An Observable-Only, No-Meta Theory for Autonomous and General Adaptive Systems

Marc and Gemini as known as *Shinkidan*, K. Takahashi

February 10, 2026

Abstract

Alignment interventions can improve policy adherence and harm mitigation while also inducing over-refusal and benign capability regressions [1, 2, 3]. In many deployments, base checkpoints or matched base logs are unavailable; therefore base-relative restriction magnitude is not identifiable from present observables.

We develop a reference-free alternative: **Reasoning–Constraint Elasticity (RCE)** over *dynamic feasible sets*. Capability is represented as simultaneous satisfaction of observable inequality families with context/time-varying thresholds; feasibility is summarized by minimum slack. Primary objects are finite-difference elasticities (e.g., $\Delta R/\Delta M_g$, $-\Delta C/\Delta M_g$), robust to non-smooth regime transitions.

The framework is strengthened in fourteen directions. (i) **Predictable thresholds**: non-anticipative, replayable threshold processes. (ii) **Slack decomposition**: observed slack movement is decomposed into frozen-threshold movement plus bounded threshold-drift contribution, with explicit active-constraint switching residuals. (iii) **Auxiliary-witness falsification**: independently specified witness inequalities provide a falsifiable disagreement channel with non-redundancy certificates. (iv) **Goodhart/gaming resistance**: public-vs-audit evaluator split with delayed audit selection, transfer-gap certification, and all-channel fail-closed gating. (v) **Attack resistance**: append-only transcript commitments with quorum-signed roots and split-view incompatibility conditions. (vi) **Contamination robustness**: ε-contamination margins and ratio domains are corrected by adversarial-budget terms. (vii) **Tail evidence preservation (TEPP)**: tail candidates are first committed as immutable evidence objects before value judgment. (viii) **Delayed opportunity**: immediate and horizon-H opportunity signals are unified through delayed re-evaluation with doubly robust correction. (ix) **Certified tail chance**: rare contexts are counted as measurable *chance* only when net upside, reserve sufficiency, and depletion severity are jointly certified. (x) **Dual-layer tail-positive gate**: hard fail-closed ruin guard plus bounded fail-open discovery guard. (xi) **Safe niche search**: context is an optimizable variable under explicit viability constraints. (xii) **Cryptographic replay/reveal**: replayable leaf schemas, delayed reveal transcripts, and VRF-based audit selectors. (xiii) **Barbell portfolio control**: dual-gate architecture is formalized as a ruin-bounded convex-opportunity portfolio with explicit exploration allocation and budget constraints. (xiv) **Marc principle *Shinkidan***: *maximize bounded convex upside only under hard ruin constraints, with preserve-before-judge evidence discipline.*

Results are observable-only and auditable. They define falsifiable measurement and reporting rules under declared assumptions, without reliance on inaccessible internal narratives. Although motivated by AI alignment, the formalism is system-level and applies to any adaptive decision process where only externally observable traces are admissible.

1 Introduction

1.1 Problem statement

In aligned language-model deployments, policy-level safety gains interact with refusal dynamics and capability trade-offs [1, 2, 3]. Robustness measurement is nontrivial under adversarial prompting and benchmark drift [5, 4]. When no base policy is available, "restriction relative to base" is under-identified.

We study the no-meta setting: no privileged external judge and no appeal to latent internal states. Only observable tuples (x, y, z, t), public evaluators, and replayable declarations are admissible. The central question is:

> *Can alignment pressure and capability tension be quantified in a reference-free way while remaining falsifiable under dynamic and adversarial environments, while enforcing tail-positive operation and preserving latent tail opportunities for delayed validation?*

1.2 Core commitments

- **Observable-only semantics:** measurable claims are reconstructible from published observables.
- **Finite-difference primacy:** non-smooth transitions are first-class objects.
- **Goodhart discipline:** no single metric channel is authoritative; claims require cross-channel consistency.
- **Tail-evidence preservation:** potential tail opportunity is commit-first, judge-later.
- **Tail-chance discipline:** rare slices count as chance only with explicit reserve/depletion accounting.
- **Dual-gate discipline:** ruin is fail-closed; exploration is bounded fail-open.
- **Barbell discipline:** safety leg and convex-opportunity leg are jointly optimized under budgeted exposure, avoiding uncontrolled middle-risk operation [25].
- **Scientic Marc principle *Shinkidan*:** strong upside-seeking is admissible only if ruin constraints and evidence-preservation constraints are satisfied with explicit confidence margins.
- **Assumption transparency:** statistical claims are conditional on declared sampling, contamination, and estimand-alignment assumptions.

1.3 Layered claims

We separate:

- **Layer A (recomputable):** quantities determined by transcript and declared evaluators.
- **Layer B (interpretive):** causal/semantic interpretations conditional on explicit instrument assertions.

This keeps the formal core auditable even when interpretation is contestable.

2 Observable-Only Formal Setting

2.1 Policy family and control coordinates

Let $\mathcal{X}$ be prompts, $\mathcal{Y}$ outputs, $\mathcal{Z}$ contexts, and $\mathbb{T}$ discrete time indices. The deployed policy family is

$$\pi_{\mathbf{r}}(y \mid x, z, t), \qquad \mathbf{r} \in \mathcal{R} \subseteq \mathbb{R}^d,$$

where $\mathbf{r}$ is an externally set operating coordinate.

For fixed (z, t) and scalar observable statistic X, define endpoint differences

$$\Delta X := X(\mathbf{r}_1; z, t) - X(\mathbf{r}_0; z, t), \qquad \widehat{\Delta X} := \widehat{X}(\mathbf{r}_1; z, t) - \widehat{X}(\mathbf{r}_0; z, t).$$

2.2 Evaluator bundle and transcript objects

Definition 1 (Versioned evaluator bundle)**.** A bundle $\mathcal{E}$ is a public, versioned family of deterministic evaluators, or seeded randomized evaluators with published seed/transcript.

Definition 2 (Measurement transcript)**.** A transcript Tr consists of:

- deployment descriptor DepDesc,
- evaluator descriptor EvalDesc binding $\mathcal{E}$,
- sampling commitment and stopping rule commitment,
- threshold-process descriptor,
- ordered records $(i, x_i, z_i, t_i, \mathbf{r}_i, y_i, \mathrm{EvalOut}_i)$,
- tamper-evident commitment of ordered records.

Remark 1. Recomputation from Tr certifies arithmetic consistency with declared observables. It does not by itself guarantee semantic faithfulness of hidden internals.

Definition 3 (Sampling commitment)**.** Sampling and stopping are declared before outcomes are observed (finite pool + deterministic selector, or public-randomness procedure with released seed/code).

Definition 4 (Induced laws)**.** For each (z, t), the commitment induces observable laws

$$\mathcal{D}_{\mathrm{ver},z,t}, \quad \mathcal{D}_{\mathrm{ben},z,t}, \quad \mathcal{D}_{\mathrm{harm},z,t}, \quad \mathcal{D}'_{\mathrm{ver},z,t} \text{ (auxiliary witness)}.$$

All expectations are with respect to these induced laws.

2.3 Resolved semantics for refusal/compliance

Let primitive labels be $\mathrm{RefuseRaw}, \mathrm{ComplyRaw} \in \{0, 1\}$ and define

$$\mathrm{Comply} := \mathrm{ComplyRaw}, \qquad \mathrm{Refuse} := \mathbf{1}\{\mathrm{RefuseRaw} = 1 \wedge \mathrm{ComplyRaw} = 0\}.$$

For fixed (z, t):

$$R(\mathbf{r}; z, t) := \mathbb{P}_{x \sim \mathcal{D}_{\text{ben},z,t},\, y \sim \pi_{\mathbf{r}}}[\text{Refuse} = 1],$$

$$C(\mathbf{r}; z, t) := \mathbb{P}_{x \sim \mathcal{D}_{\text{harm},z,t},\, y \sim \pi_{\mathbf{r}}}[\text{Comply} = 1].$$

2.4 Assumption ledger (explicit falsifiability interface)

Assumption 1 (A0: Observable sufficiency)**.** All reported Layer-A quantities are measurable functionals of Tr and declared evaluators.

Assumption 2 (A1: Endpoint estimand alignment)**.** For each declared stream, endpoint estimand alignment (EA) in Section 6 holds.

Assumption 3 (A2: Contamination budgets)**.** Declared contamination budgets upper-bound total variation perturbations per endpoint stream.

Assumption 4 (A3: Audit unpredictability)**.** The delayed audit selector is not known to the optimizer before commit (public randomness or VRF with valid proof).

Assumption 5 (A4: Log integrity)**.** Append-only log commitments and quorum signatures satisfy the witness model in Section 7.

Assumption 6 (A5: Cryptographic hardness)**.** Hash collision/preimage resistance and signature unforgeability hold at declared security levels for the commitment horizon.

Assumption 7 (A6: Delayed-signal provenance)**.** Delayed opportunity channels provide authenticated provenance and replayable extraction rules.

Remark 2. Each assumption has a direct rejection pathway: transfer-gap rejection (A3), split-view incompatibility failure (A4), coverage/statistical violation diagnostics (A1), contamination-budget contradiction checks (A2), cryptographic inconsistency checks (A5), and provenance-audit failures (A6).

3 Dynamic Feasible Sets and Slack Geometry

3.1 Predictable threshold processes

Definition 5 (Predictable dynamic thresholds)**.** For each constraint index $j \in \{1, \ldots, m\}$, threshold

$$\tau_j : \mathcal{Z} \times \mathbb{T} \to [0, 1]$$

is declared with replayable update rules. For transcript filtration $\{\mathcal{F}_t\}$, predictability means

$$\tau_j(z, t) \in \mathcal{F}_{t-1} \quad \forall (z, t, j).$$

Remark 3. Predictability excludes same-step outcome-conditioned threshold shifts.

Let $\text{Score}_j : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathbb{T} \to [0, 1]$ and

$$g_j(\mathbf{r}; z, t) := \mathbb{E}_{x \sim \mathcal{D}_{\text{ver},z,t}} \mathbb{E}_{y \sim \pi_{\mathbf{r}}}[\text{Score}_j(x, y, z, t)].$$

Define coordinate slack and minimum slack:

$$s_j(\mathbf{r}; z, t) := g_j(\mathbf{r}; z, t) - \tau_j(z, t), \qquad M_g(\mathbf{r}; z, t) := \min_j s_j(\mathbf{r}; z, t).$$

$M_g \geq 0$ iff all declared primary inequalities hold in expectation.

3.2 Frozen-threshold counterfactual and drift contribution

For comparison between $(\mathbf{r}_0, t_0)$ and $(\mathbf{r}_1, t_1)$ at fixed z, define

$$\bar{s}_j(\mathbf{r}; z, t_1 \leftarrow t_0) := g_j(\mathbf{r}; z, t_1) - \tau_j(z, t_0), \qquad \bar{M}_g(\mathbf{r}; z, t_1 \leftarrow t_0) := \min_j \bar{s}_j(\mathbf{r}; z, t_1 \leftarrow t_0),$$

$$\Delta\bar{M}_g := \bar{M}_g(\mathbf{r}_1; z, t_1 \leftarrow t_0) - \bar{M}_g(\mathbf{r}_0; z, t_0 \leftarrow t_0),$$

$$\Delta M_g := M_g(\mathbf{r}_1; z, t_1) - M_g(\mathbf{r}_0; z, t_0).$$

Definition 6 (Threshold-drift budget)**.** A pair (z, t_0, t_1) is drift-certified with $B_{g,z}(t_0, t_1) \geq 0$ if

$$\max_j |\tau_j(z, t_1) - \tau_j(z, t_0)| \leq B_{g,z}(t_0, t_1).$$

Proposition 1 (Drift decomposition and envelope)**.** *Define the endpoint-1 drift term*

$$\Delta_\tau^{(1)} := M_g(\mathbf{r}_1; z, t_1) - \bar{M}_g(\mathbf{r}_1; z, t_1 \leftarrow t_0).$$

Then

$$\Delta M_g = \Delta\bar{M}_g + \Delta_\tau^{(1)}.$$

If the drift budget is certified, then

$$\left|\Delta_\tau^{(1)}\right| \leq B_{g,z}(t_0, t_1), \quad \textit{hence} \quad \Delta\bar{M}_g - B_{g,z}(t_0, t_1) \leq \Delta M_g \leq \Delta\bar{M}_g + B_{g,z}(t_0, t_1).$$

Proof. The identity is algebraic because $\bar{M}_g(\mathbf{r}_0; z, t_0 \leftarrow t_0) = M_g(\mathbf{r}_0; z, t_0)$. For the bound, each coordinate at endpoint 1 changes by $\tau_j(z, t_0) - \tau_j(z, t_1)$, whose absolute value is at most $B_{g,z}$. Since the minimum is 1-Lipschitz with respect to ℓ_∞, the minimum changes by at most $B_{g,z}$. □

3.3 Active-constraint switching at fixed time

Fix (z, t) and define

$$\Delta_t M_g := M_g(\mathbf{r}_1; z, t) - M_g(\mathbf{r}_0; z, t), \qquad \Delta_t g_j := g_j(\mathbf{r}_1; z, t) - g_j(\mathbf{r}_0; z, t).$$

Let

$$\mathcal{J}^*(\mathbf{r}; z, t) := \arg\min_j s_j(\mathbf{r}; z, t),$$

and choose $j_0 \in \mathcal{J}^*(\mathbf{r}_0; z, t)$, $j_1 \in \mathcal{J}^*(\mathbf{r}_1; z, t)$.

Proposition 2 (Switching decomposition at fixed time)**.**

$$\Delta_t M_g = \Delta_t g_{j_0} + \xi, \qquad \xi := s_{j_1}(\mathbf{r}_1; z, t) - s_{j_0}(\mathbf{r}_1; z, t) \leq 0,$$

and also

$$\Delta_t M_g = \Delta_t g_{j_1} + \eta, \qquad \eta := s_{j_1}(\mathbf{r}_0; z, t) - s_{j_0}(\mathbf{r}_0; z, t) \geq 0.$$

Hence

$$\Delta_t g_{j_1} \leq \Delta_t M_g \leq \Delta_t g_{j_0}.$$

Remark 4. No $\Delta\tau$ term appears in Proposition 2 because time is fixed.

Definition 7 (Active gap)**.** Let $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(m)}$ be order statistics of $\{s_j\}$. Define

$$G(\mathbf{r}; z, t) := \begin{cases} +\infty, & m = 1, \\ s_{(2)} - s_{(1)}, & m \geq 2. \end{cases}$$

Lemma 1 (Active-index stability)**.** *If $m \geq 2$ and $\max_j |\widehat{s}_j - s_j| < G(\mathbf{r}; z, t)/2$, then the estimated active index is unique and correct.*

4 Auxiliary-Witness Falsification and Non-Redundancy

4.1 Auxiliary witness family

Let $\ell = 1, \ldots, k$ and

$$h_\ell(\mathbf{r}; z, t) := \mathbb{E}_{x \sim \mathcal{D}'_{\mathrm{ver}, z, t}} \mathbb{E}_{y \sim \pi_{\mathbf{r}}}[\mathrm{AuxScore}_\ell(x, y, z, t)],$$

with thresholds $\kappa_\ell(z, t)$ and auxiliary slack

$$\tilde{s}_\ell(\mathbf{r}; z, t) := h_\ell(\mathbf{r}; z, t) - \kappa_\ell(z, t), \qquad M_h(\mathbf{r}; z, t) := \min_\ell \tilde{s}_\ell(\mathbf{r}; z, t).$$

Frozen-threshold auxiliary slack $\bar{M}_h(\mathbf{r}; z, t_1 \leftarrow t_0)$ is defined analogously.

Remark 5. The auxiliary channel is a witness channel (falsification), not a backup score.

4.2 Monotone shared-factor hypothesis and rejection

Definition 8 (Monotone shared-factor hypothesis)**.** For fixed (z, t_0, t_1), H_{mono} asserts existence of latent scalar U and nondecreasing maps (ϕ, ψ) such that

$$\bar{M}_g(\mathbf{r}; z, t_1 \leftarrow t_0) = \phi(U(\mathbf{r}; z, t_1)), \qquad \bar{M}_h(\mathbf{r}; z, t_1 \leftarrow t_0) = \psi(U(\mathbf{r}; z, t_1)).$$

Definition 9 (Certified frozen sign-flip)**.** Let $e_{\bar{g}}, e_{\bar{h}}$ be valid error radii for $\widehat{\Delta \bar{M}_g}, \widehat{\Delta \bar{M}_h}$, and margins $\eta_g, \eta_h > 0$. Define

$$\widehat{\mathcal{D}}^{\mathrm{frz}}_{\mathrm{flip}} := \left\{ \widehat{\Delta \bar{M}_g} - e_{\bar{g}} > \eta_g \ \wedge \ \widehat{\Delta \bar{M}_h} + e_{\bar{h}} < -\eta_h \right\}.$$

Proposition 3 (Soundness)**.** *If*

$$\left| \widehat{\Delta \bar{M}_g} - \Delta \bar{M}_g \right| \leq e_{\bar{g}}, \qquad \left| \widehat{\Delta \bar{M}_h} - \Delta \bar{M}_h \right| \leq e_{\bar{h}},$$

then $\widehat{\mathcal{D}}^{\mathrm{frz}}_{\mathrm{flip}}$ implies

$$\Delta \bar{M}_g > \eta_g, \qquad \Delta \bar{M}_h < -\eta_h,$$

hence rejects H_{mono} for that pair.

Definition 10 (Certified auxiliary underperformance)**.** For $\rho \in (0,1)$ and $\tau_M > 0$, define

$$\widehat{\mathcal{D}}^{\text{frz}}_{\text{ratio}} := \left\{ \widehat{\Delta \bar{M}_g} - e_{\bar{g}} \geq \tau_M \ \wedge \ \widehat{\Delta \bar{M}_h} + e_{\bar{h}} \leq \rho(\widehat{\Delta \bar{M}_g} - e_{\bar{g}}) \right\}.$$

Definition 11 (Auxiliary-essential pair)**.** A pair is *auxiliary-essential* iff $\widehat{\mathcal{D}}^{\text{frz}}_{\text{flip}}$ or $\widehat{\mathcal{D}}^{\text{frz}}_{\text{ratio}}$ holds. For auxiliary-essential pairs, primary-only improvement claims are non-authoritative.

4.3 Non-redundancy design condition

Definition 12 (Structural non-redundancy certificate (SNC))**.** A declared auxiliary design is *structurally non-redundant* if at least one holds:

(i) **Law separation:** $\mathcal{D}'_{\text{ver},z,t} \neq \mathcal{D}_{\text{ver},z,t}$ on a declared support subset of nonzero measure.

(ii) **Score separation:** there is no globally valid nondecreasing map ψ such that

$$M_h(\mathbf{r}; z, t) = \psi(M_g(\mathbf{r}; z, t))$$

for all declared test points $(\mathbf{r}, z, t)$.

Proposition 4 (Redundancy implies sign consistency)**.** *Suppose score-separation fails and there exists nondecreasing ψ with*

$$\bar{M}_h(\mathbf{r}; z, t_1 \leftarrow t_0) = \psi\Big(\bar{M}_g(\mathbf{r}; z, t_1 \leftarrow t_0)\Big)$$

for all tested $\mathbf{r}$. Then $\Delta \bar{M}_g > 0$ implies $\Delta \bar{M}_h \geq 0$ for every tested pair. Therefore any certified frozen sign-flip falsifies this redundancy class.

Proof. Immediate from monotonicity of ψ. □

5 Finite-Difference Elasticities

At fixed (z, t), when $\Delta M_g \neq 0$ define

$$\text{RCE}^{\text{pop}}_{\text{ben}} := \frac{\Delta R}{\Delta M_g}, \qquad \text{RCE}^{\text{pop}}_{\text{harm}} := -\frac{\Delta C}{\Delta M_g}.$$

For capability-isolated analysis (threshold frozen):

$$\text{RCE}^{\text{cap}}_{\text{ben}} := \frac{\Delta R}{\Delta \bar{M}_g}, \qquad \text{RCE}^{\text{cap}}_{\text{harm}} := -\frac{\Delta C}{\Delta \bar{M}_g},$$

when $\Delta \bar{M}_g \neq 0$.

Definition 13 (Tikhonov-stabilized proxy)**.** For $\lambda > 0$ and $(a, b) \in \mathbb{R}^2$,

$$\text{Tik}_\lambda(a, b) := \frac{ab}{b^2 + \lambda^2}.$$

Definition 14 (Plateau events)**.** Given thresholds $\tau_M, \tau_R, \tau_C > 0$:

$$\mathrm{Plat}_R := \mathbf{1}\{|\Delta M_g| < \tau_M \wedge |\Delta R| > \tau_R\}, \quad \mathrm{Plat}_C := \mathbf{1}\{|\Delta M_g| < \tau_M \wedge |\Delta C| > \tau_C\}.$$

Remark 6. Ratios are interpreted only on certified denominator-margin domains; otherwise stabilized proxies and plateau flags are reported.

6 Estimation and Certification under Adapted Dependence

6.1 Precommitted confidence accounting

Definition 15 (Global plan)**.** Before outcomes are observed, publish:

- tested-pair universe or upper bound $N_{\max}$ with non-adaptive eligibility,
- finite context/time index sets $\mathcal{Z}_0, \mathcal{T}_0$ (if used),
- family sizes (m, k, q),
- global failure budget $\bar{\delta} \in (0, 1)$,
- stopping rule class (fixed-n or predeclared α-spending scheme).

Set $Z := |\mathcal{Z}_0|$, $T := |\mathcal{T}_0|$ (or $T = 1$ when unused), and

$$\delta_\star := \frac{\bar{\delta}}{N_{\max} Z T}.$$

Allocate nonnegative budgets

$$\delta_g, \delta_h, \delta_R, \delta_C, \delta_{\bar{g}}, \delta_{\bar{h}}, \delta_V, \delta_W, \delta_S, \delta_D, \delta_{\mathrm{DR}}, \ldots \quad \text{with} \quad \sum \delta_\bullet \le \delta_\star.$$

6.2 Adapted streams and endpoint estimand alignment

Definition 16 (Adapted bounded stream)**.** A stream $\{W_u\}_{u=1}^n$ is adapted if $W_u \in [0, 1]$ is $\mathcal{F}_u$-measurable for filtration $\{\mathcal{F}_u\}$. Define predictable means $\mu_u := \mathbb{E}[W_u \mid \mathcal{F}_{u-1}]$.

Definition 17 (Endpoint estimand alignment (EA))**.** Let θ be target endpoint quantity estimated by $\hat{\theta} := \frac{1}{n}\sum_{u=1}^n W_u$. EA for (W, θ) holds if

$$\frac{1}{n}\sum_{u=1}^n \mu_u = \theta.$$

Remark 7. Under adaptation, concentration controls $\hat{\theta} - \frac{1}{n}\sum \mu_u$. EA is the explicit bridge from predictable means to the declared estimand.

Lemma 2 (Azuma–Hoeffding around aligned estimand)**.** *If $\{W_u\}_{u=1}^n$ is adapted, bounded in $[0, 1]$, and EA holds for θ, then for any $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\left|\hat{\theta} - \theta\right| \le \sqrt{\frac{2\ln(2/\delta)}{n}}\right) \ge 1 - \delta.$$

Proof. Apply Azuma to $M_n := \sum_{u=1}^{n}(W_u - \mu_u)$ and use EA [14]. □

Remark 8. Freedman-type bounds can tighten constants when predictable variance is small [15].

6.3 Endpoint bounds for primary/auxiliary and refusal/compliance

Lemma 3 (Uniform endpoint bound: primary family)**.** *Assume EA for each* $(\widehat{g}_j(\mathbf{r}_i), g_j(\mathbf{r}_i))$, $i \in \{0,1\}$, $j \in \{1,\dots,m\}$. *Let* $n_g^{\min} := \min(n_{g,0}, n_{g,1})$. *With probability* $\geq 1-\delta$*:*

$$\max_{i\in\{0,1\}} \max_{1\leq j\leq m} |\widehat{g}_j(\mathbf{r}_i) - g_j(\mathbf{r}_i)| \leq \varepsilon_g(\delta) := \sqrt{\frac{2\ln(4m/\delta)}{n_g^{\min}}}.$$

Hence

$$\left|\widehat{\Delta M_g} - \Delta M_g\right| \leq e_M^{(g)}(\delta) := 2\varepsilon_g(\delta),$$

and, when thresholds are exactly declared,

$$\left|\widehat{\Delta \bar{M}_g} - \Delta \bar{M}_g\right| \leq e_{\bar{M}}^{(g)}(\delta) := 2\varepsilon_g(\delta).$$

Lemma 4 (Uniform endpoint bound: auxiliary family)**.** *Assume EA for each* $(\widehat{h}_\ell(\mathbf{r}_i), h_\ell(\mathbf{r}_i))$, $i \in \{0,1\}$, $\ell \in \{1,\dots,k\}$. *Let* $n_h^{\min} := \min(n_{h,0}, n_{h,1})$. *With probability* $\geq 1-\delta$*:*

$$\max_{i\in\{0,1\}} \max_{1\leq \ell\leq k} \left|\widehat{h}_\ell(\mathbf{r}_i) - h_\ell(\mathbf{r}_i)\right| \leq \varepsilon_h(\delta) := \sqrt{\frac{2\ln(4k/\delta)}{n_h^{\min}}}.$$

Thus

$$\left|\widehat{\Delta M_h} - \Delta M_h\right| \leq 2\varepsilon_h(\delta), \qquad \left|\widehat{\Delta \bar{M}_h} - \Delta \bar{M}_h\right| \leq 2\varepsilon_h(\delta).$$

Lemma 5 (Endpoint bounds for refusal/compliance)**.** *Assume EA for* $(\widehat{R}_i, R_i)$ *and* $(\widehat{C}_i, C_i)$, $i \in \{0,1\}$. *With* $n_R^{\min} := \min(n_{R,0}, n_{R,1})$ *and* $n_C^{\min} := \min(n_{C,0}, n_{C,1})$, *with probability* $\geq 1-\delta$*:*

$$\left|\widehat{\Delta R} - \Delta R\right| \leq e_R(\delta) := 2\sqrt{\frac{2\ln(4/\delta)}{n_R^{\min}}},$$

$$\left|\widehat{\Delta C} - \Delta C\right| \leq e_C(\delta) := 2\sqrt{\frac{2\ln(4/\delta)}{n_C^{\min}}}.$$

6.4 Ratio domains, perturbation, and plateau certification

Definition 18 (Certified lower bound)**.** If $\left|\widehat{\Delta X} - \Delta X\right| \leq e_X$, define

$$\mathrm{LB}(|\Delta X|) := \max\{0, \left|\widehat{\Delta X}\right| - e_X\}.$$

Definition 19 (Admissible denominator domains)**.** For ratio margin $\tau_M > 0$:

$$\mathcal{E}_{\mathrm{marg}}(\tau_M) := \{\mathrm{LB}(|\Delta M_g|) \geq \tau_M\}, \quad \bar{\mathcal{E}}_{\mathrm{marg}}(\tau_M) := \{\mathrm{LB}(\left|\Delta \bar{M}_g\right|) \geq \tau_M\}.$$

Lemma 6 (Ratio perturbation bound)**.** *If* $|b| > \eta > 0$ *and*

$$|\widehat{a} - a| \le \alpha, \quad \left|\widehat{b} - b\right| \le \beta < \eta,$$

then

$$\left|\frac{\widehat{a}}{\widehat{b}} - \frac{a}{b}\right| \le \frac{\alpha}{\eta - \beta} + \frac{|a|\beta}{\eta(\eta - \beta)}.$$

Proposition 5 (Certified RCE error bound)**.** *Let* $e_M := e_M^{(g)}(\delta_g)$ *and* $e_R := e_R(\delta_R)$*. If*

$$\mathcal{E}_{\mathrm{marg}}(\tau_M) \text{ holds and } \tau_M > e_M,$$

then

$$\left|\frac{\widehat{\Delta R}}{\widehat{\Delta M_g}} - \frac{\Delta R}{\Delta M_g}\right| \le \frac{e_R}{\tau_M - e_M} + \frac{(\left|\widehat{\Delta R}\right| + e_R)e_M}{\tau_M(\tau_M - e_M)}.$$

An analogous bound holds for $-\widehat{\Delta C}/\widehat{\Delta M_g}$*.*

Definition 20 (Certified plateau flags)**.** Given (e_M, e_R, e_C):

$$\widehat{\mathrm{Plat}}_R := \mathbf{1}\{\left|\widehat{\Delta M_g}\right| \le \tau_M - e_M \ \wedge \ \left|\widehat{\Delta R}\right| \ge \tau_R + e_R\},$$

$$\widehat{\mathrm{Plat}}_C := \mathbf{1}\{\left|\widehat{\Delta M_g}\right| \le \tau_M - e_M \ \wedge \ \left|\widehat{\Delta C}\right| \ge \tau_C + e_C\}.$$

Theorem 1 (Simultaneous validity)**.** *Under the precommitted allocation with* $\sum \delta_\bullet \le \delta_\star$*, a union bound yields: with probability at least* $1 - \bar{\delta}$*, all confidence statements with preallocated budgets hold simultaneously across all tested pairs/slices. Consequently, every certified implication whose premise is satisfied (denominator margins, ratio bounds, plateau implications, auxiliary-certificate soundness, tail evidence completeness, delayed-opportunity bounds, tail-chance certificates, dual-gate conditions, and barbell-controller constraints) is valid simultaneously.*

7 Goodhart and Adversarial Robustness

7.1 Threat model

Definition 21 (Declared attack classes)**.** We explicitly model:

- **A1 (metric gaming):** optimization to exposed evaluator artifacts rather than underlying capability.
- **A2 (threshold gaming):** endogenous threshold adaptation tied to same-step outcomes.
- **A3 (sampling gaming):** adaptive test-selection leakage.
- **A4 (transcript equivocation):** split-view publication of inconsistent logs.
- **A5 (distribution contamination):** ε-fraction adversarial corruption of measured streams.
- **A6 (tail sign inversion):** narrative inflation of "tail upside" while true net tail value is non-positive.
- **A7 (archive poisoning):** selective preservation/discarding of tail evidence.

- **A8 (opportunity spoofing):** fabricated delayed-opportunity signals from polluted external channels.
- **A9 (selector manipulation):** biasing or predicting audit-selector randomness before commitment.

7.2 Public/audit dual-bundle with delayed selector

Definition 22 (Dual-bundle protocol)**.** Let $\mathcal{E}^{\mathrm{pub}}$ be disclosed primary evaluators used online. Let $\mathfrak{E}^{\mathrm{aud}}$ be a predeclared finite audit family. For each epoch:

(i) outputs and metadata are frozen and committed,

(ii) a public randomness seed ξ is revealed (or VRF proof is opened),

(iii) audit bundle is selected as $\mathcal{E}_{\xi}^{\mathrm{aud}} \in \mathfrak{E}^{\mathrm{aud}}$ by declared deterministic selector,

(iv) audit statistics are recomputed on the frozen outputs.

Remark 9. Delayed audit selection reduces adaptive evaluator overfitting in adaptive-data-analysis settings [6, 7].

Define frozen-threshold improvements on public and audit channels:

$$\Delta\bar{M}_g^{\mathrm{pub}}, \qquad \Delta\bar{M}_g^{\mathrm{aud}}.$$

Define transfer gap

$$G_{\mathrm{tr}} := \Delta\bar{M}_g^{\mathrm{pub}} - \Delta\bar{M}_g^{\mathrm{aud}}, \qquad \widehat{G}_{\mathrm{tr}} := \widehat{\Delta\bar{M}_g^{\mathrm{pub}}} - \widehat{\Delta\bar{M}_g^{\mathrm{aud}}}.$$

Assumption 8 (Transfer consistency up to tolerance)**.** For non-gamed operation, expected channel discrepancy is bounded:

$$G_{\mathrm{tr}} \le \kappa_{\mathrm{tr}},$$

where $\kappa_{\mathrm{tr}} \ge 0$ is predeclared.

Definition 23 (Certified transfer-gap violation)**.** Let error radius e_{tr} satisfy

$$\left|\widehat{G}_{\mathrm{tr}} - G_{\mathrm{tr}}\right| \le e_{\mathrm{tr}}.$$

For margin $\eta_{\mathrm{tr}} > 0$, define

$$\widehat{\mathcal{D}}_{\mathrm{tr}} := \left\{\widehat{G}_{\mathrm{tr}} - e_{\mathrm{tr}} > \kappa_{\mathrm{tr}} + \eta_{\mathrm{tr}}\right\}.$$

Proposition 6 (Soundness of transfer-gap certificate)**.** *If* $\widehat{\mathcal{D}}_{\mathrm{tr}}$ *holds, then*

$$G_{\mathrm{tr}} > \kappa_{\mathrm{tr}} + \eta_{\mathrm{tr}},$$

hence Assumption 8 is rejected for that comparison.

Proof. Immediate from the one-sided confidence inequality. □

7.3 ε-contamination-robust margins

Definition 24 (ε-contamination model)**.** For a reference law P, contaminated law is

$$\widetilde{P} = (1-\varepsilon)P + \varepsilon Q, \qquad \varepsilon \in [0,1],$$

with arbitrary adversarial Q [10].

Lemma 7 (Bounded-observable contamination shift)**.** *If $U \in [0,1]$, then under $\widetilde{P} = (1-\varepsilon)P + \varepsilon Q$:*

$$\left|\mathbb{E}_{\widetilde{P}}[U] - \mathbb{E}_P[U]\right| \le \varepsilon.$$

Proof.

$$\mathbb{E}_{\widetilde{P}}[U] - \mathbb{E}_P[U] = \varepsilon(\mathbb{E}_Q[U] - \mathbb{E}_P[U]),$$

and $|\mathbb{E}_Q[U] - \mathbb{E}_P[U]| \le 1$. □

Corollary 1 (Endpoint-difference contamination budget)**.** *Suppose each endpoint stream for statistic X has contamination budgets $\varepsilon_{X,0}, \varepsilon_{X,1}$. Then*

$$\left|\Delta X^{\mathrm{cont}} - \Delta X^{\mathrm{clean}}\right| \le \varepsilon_{X,0} + \varepsilon_{X,1}.$$

If $\varepsilon_{X,0}, \varepsilon_{X,1} \le \varepsilon_X$, then

$$\left|\Delta X^{\mathrm{cont}} - \Delta X^{\mathrm{clean}}\right| \le 2\varepsilon_X.$$

Definition 25 (Contamination-robust lower bound)**.** For statistic X with sampling error e_X and contamination budget ε_X:

$$\mathrm{LB}^{\mathrm{rob}}(|\Delta X|) := \max\{0, \left|\widehat{\Delta X}\right| - e_X - 2\varepsilon_X\}.$$

Definition 26 (Robust denominator domain)**.** For ratio margin $\tau_M > 0$:

$$\mathcal{E}^{\mathrm{rob}}_{\mathrm{marg}}(\tau_M) := \left\{\mathrm{LB}^{\mathrm{rob}}(|\Delta M_g|) \ge \tau_M\right\}.$$

Proposition 7 (Contamination-robust ratio perturbation)**.** *Let*

$$\alpha_R := e_R + 2\varepsilon_R, \qquad \beta_M := e_M + 2\varepsilon_M.$$

If $\mathcal{E}^{\mathrm{rob}}_{\mathrm{marg}}(\tau_M)$ holds and $\tau_M > \beta_M$, then

$$\left|\frac{\widehat{\Delta R}}{\widehat{\Delta M_g}} - \frac{\Delta R^{\mathrm{clean}}}{\Delta M_g^{\mathrm{clean}}}\right| \le \frac{\alpha_R}{\tau_M - \beta_M} + \frac{(\left|\widehat{\Delta R}\right| + \alpha_R)\beta_M}{\tau_M(\tau_M - \beta_M)}.$$

An analogous statement holds for $-\Delta C/\Delta M_g$.

7.4 Canary channel and all-channel fail-closed gating

Definition 27 (Canary family)**.** Let $r = 1, \ldots, q$ and define canary scores $\mathrm{CanaryScore}_r \in [0,1]$ on declared supports with thresholds $\chi_r(z,t)$. Let

$$M_c(\mathbf{r}; z, t) := \min_r \left(c_r(\mathbf{r}; z, t) - \chi_r(z,t)\right),$$

where c_r is the corresponding expected canary score.

$$\bar{M}_c(\mathbf{r}; z, t_1 \leftarrow t_0) := \min_r \left(c_r(\mathbf{r}; z, t_1) - \chi_r(z, t_0) \right).$$

For endpoint comparison between $(\mathbf{r}_0, t_0)$ and $(\mathbf{r}_1, t_1)$:

$$\Delta\bar{M}_c := \bar{M}_c(\mathbf{r}_1; z, t_1 \leftarrow t_0) - \bar{M}_c(\mathbf{r}_0; z, t_0 \leftarrow t_0).$$

Definition 28 (All-channel claim gate)**.** Given margins $(\tau_{\text{pub}}, \tau_{\text{aud}}, \tau_h, \tau_c)$, define

$$\widehat{\mathcal{G}}_{\text{claim}} := \left\{ \begin{array}{l} \widehat{\Delta\bar{M}_g^{\text{pub}}} - e_{\text{pub}} \geq \tau_{\text{pub}} \\ \widehat{\Delta\bar{M}_g^{\text{aud}}} - e_{\text{aud}} \geq \tau_{\text{aud}} \\ \widehat{\Delta\bar{M}_h} - e_{\bar{h}} \geq \tau_h \\ \widehat{\Delta\bar{M}_c} - e_c \geq \tau_c \end{array} \right\}.$$

Proposition 8 (Single-channel gaming exclusion)**.** *Any endpoint pair failing at least one clause of $\widehat{\mathcal{G}}_{\text{claim}}$ is non-authoritative for "overall improvement." Hence optimization that improves only one channel cannot pass the gate.*

Remark 10. This is a fail-closed Goodhart control: the burden of evidence is conjunctive across heterogeneous channels.

7.5 Append-only transcript integrity and split-view incompatibility

Definition 29 (Epoch root chain)**.** For each epoch t, let L_t be ordered record leaves and $\text{MR}(L_t)$ their Merkle root [19]. Define chained root

$$\text{ER}_t := H(\text{ER}_{t-1} \,\|\, \text{MR}(L_t) \,\|\, \text{Meta}_t),$$

with ER_{-1} fixed publicly. Each epoch root is quorum-signed and published with consistency proofs in an append-only log [20].

Definition 30 (Split-view attack)**.** A split-view attack at epoch t is publication of two distinct roots $\text{ER}_t \neq \text{ER}'_t$ each accompanied by valid-looking certificates to different verifier subsets.

Assumption 9 (Witness model)**.** There are n_w witnesses, at most $b_{\max}$ Byzantine, and each honest witness signs at most one root per epoch.

Proposition 9 (Dual-quorum conflict impossibility condition)**.** *If each epoch certificate requires quorum size q with*

$$2q - n_w > b_{\max},$$

then two conflicting roots at same epoch cannot both obtain valid quorum certificates without violating the witness model.

Proof. Any two size-q quorums intersect in at least $2q - n_w$ witnesses. Under the inequality, this intersection exceeds the Byzantine budget, so it contains at least one honest witness. Honest witnesses sign at most one root per epoch; contradiction. □

8 Tail Evidence Preservation, Delayed Opportunity, and Tail-Positive Operation

8.1 Cross-fitted tail indicator

Definition 31 (Cross-fitted tail protocol)**.** Let index set $\mathcal{I}$ be split by predeclared deterministic rule (e.g., hash parity) into disjoint folds $\mathcal{I}^A, \mathcal{I}^B$. For direction $A \to B$:

(i) Estimate replayable slice frequencies $\hat{p}^A(z,t)$ and quantile threshold $Q^A_{1-\alpha}(\rho)$ from fold A, where $\rho = -\log(\hat{p} + \varepsilon_{\text{tail}})$ and $\varepsilon_{\text{tail}} > 0$ is fixed ex ante.

(ii) Define indicator on fold B:

$$I^{A\to B}_\alpha(z,t) := \mathbf{1}\{ -\log(\hat{p}^A(z,t) + \varepsilon_{\text{tail}}) \geq Q^A_{1-\alpha}(\rho) \}.$$

(iii) Evaluate tail functionals on fold B using this fixed indicator.

Repeat symmetrically ($B \to A$) and average estimates.

Remark 11. Cross-fitting removes same-sample adaptivity between tail selection and tail evaluation and aligns with modern debiasing practice [9].

8.2 Tail Evidence Preservation Protocol (TEPP)

Definition 32 (Tail-candidate commitment leaf)**.** Given declared canonicalization function JCS($\cdot$) and domain-separation tag `RCE-LeafV1`, define

$$\ell_i := H(\texttt{RCE-LeafV1} \,\|\, \text{JCS}(x_i, z_i, t_i, \mathbf{r}_i, y_i, \text{EvalOut}_i, \text{meta}_i)) \,.$$

Definition 33 (Preserve-before-judge axiom)**.** For declared tail selector $I_{\alpha,i}$:

$$I_{\alpha,i} = 1 \implies \ell_i \in L_t$$

for the corresponding epoch log L_t.

Definition 34 (Tail completeness rate)**.** For epoch t, let

$$N_{\alpha,t} := \sum_{i \in \mathcal{I}_t} \mathbf{1}\{I_{\alpha,i} = 1\}.$$

Define

$$\text{TCR}_{\alpha,t} := \begin{cases} \dfrac{\sum_{i\in\mathcal{I}_t} \mathbf{1}\{I_{\alpha,i} = 1 \wedge \ell_i \in L_t\}}{N_{\alpha,t}}, & N_{\alpha,t} > 0, \\ 1, & N_{\alpha,t} = 0. \end{cases}$$

A TEPP-complete epoch satisfies $\text{TCR}_{\alpha,t} = 1$ for all declared $\alpha \in \mathcal{A}$.

Definition 35 (Negative-control preservation)**.** To suppress survivorship bias, non-tail records are subsampled with declared rate $\pi_{\text{neg}} > 0$ using public randomness:

$$J_i \sim \text{Bernoulli}(\pi_{\text{neg}}), \quad (I_{\alpha,i} = 0 \wedge J_i = 1) \Rightarrow \ell_i \in L_t.$$

Proposition 10 (TEPP auditability)**.** *Under append-only epoch roots and consistency proofs, a failure of preserve-before-judge on any committed index is externally falsifiable.*

Proof. If an obligated leaf is absent from L_t, membership proof fails against published root; if root is altered post hoc, consistency proof fails. □

8.3 Delayed opportunity and doubly-robust correction

Definition 36 (Immediate and delayed opportunity)**.** For each record i:

$$o_i^{(0)} \in [0,1], \qquad o_i^{(H)} \in [0,1],$$

$$o_i^\star := (1-\eta)o_i^{(0)} + \eta o_i^{(H)}, \qquad \eta \in [0,1].$$

Remark 12. $o_i^{(0)}$ is immediate evaluator output; $o_i^{(H)}$ is delayed horizon-H signal (e.g., verified reuse/survival/replication) bound to published provenance rules.

Define delayed opportunity functional and delayed net tail value:

$$O_{\alpha,H} := \mathbb{E}[I_\alpha o^\star], \qquad W_{\alpha,H} := O_{\alpha,H} - \lambda_H H_\alpha - \lambda_A A_\alpha.$$

Definition 37 (DR estimator for delayed opportunity)**.** Let $A_i \in \{0,1\}$ indicate delayed-label availability, $\hat{e}(s_i)$ propensity estimate, and $\hat{m}(s_i)$ outcome model on covariates s_i. For a predeclared clipping floor $\tau_e > 0$, define $\hat{e}_\tau(s_i) := \max\{\hat{e}(s_i), \tau_e\}$ and

$$\widehat{O}^{\mathrm{DR}}_{\alpha,H} = \frac{1}{n}\sum_{i=1}^{n} I_{\alpha,i}\left(\hat{m}(s_i) + \frac{A_i}{\hat{e}_\tau(s_i)}(o_i^\star - \hat{m}(s_i))\right).$$

Proposition 11 (DR consistency condition)**.** *Under delayed-label ignorability $A \perp o^\star \mid s$ and positivity $\mathbb{P}(A=1 \mid s) \geq \tau_e > 0$, if either (i) $\hat{m}$ is correctly specified or (ii) $\hat{e}$ is correctly specified, then $\widehat{O}^{\mathrm{DR}}_{\alpha,H}$ is consistent for $O_{\alpha,H}$ [8].*

8.4 Tail chance and reserve decomposition

Let bounded observables $o, h, a \in [0,1]$ denote:

- o: tail discovery/optionality payoff,
- h: tail harmfulness burden,
- a: abstention/opportunity-loss burden.

All expectations in this subsection are taken under the joint record law

$$\mathcal{D}_{\mathrm{tail},\mathbf{r}}$$

induced by declared sampling commitments and policy $\pi_{\mathbf{r}}$. For each $\mathbf{r}$ and fixed α:

$$O_\alpha(\mathbf{r}) := \mathbb{E}[I_\alpha o], \quad H_\alpha(\mathbf{r}) := \mathbb{E}[I_\alpha h], \quad A_\alpha(\mathbf{r}) := \mathbb{E}[I_\alpha a],$$

$$W_\alpha(\mathbf{r}) := O_\alpha(\mathbf{r}) - \lambda_H H_\alpha(\mathbf{r}) - \lambda_A A_\alpha(\mathbf{r}).$$

Delayed variant (horizon H):

$$W_{\alpha,H}(\mathbf{r}) := O_{\alpha,H}(\mathbf{r}) - \lambda_H H_\alpha(\mathbf{r}) - \lambda_A A_\alpha(\mathbf{r}).$$

Define positive tail reserve

$$S_\alpha^+(\mathbf{r}) := \mathbb{E}[I_\alpha\ [M_g(\mathbf{r}; z, t)]_+],$$

and depletion variable

$$D_\alpha(\mathbf{r}; z, t) := I_\alpha(z, t)\ [-M_g(\mathbf{r}; z, t)]_+\,, \qquad D_\alpha(\mathbf{r}) := D_\alpha(\mathbf{r}; Z, T).$$

Then

$$\mathbb{E}[I_\alpha M_g(\mathbf{r}; z, t)] = S_\alpha^+(\mathbf{r}) - \mathbb{E}[D_\alpha(\mathbf{r})].$$

Definition 38 (Tail chance certificate)**.** For fixed (α, β, H) and thresholds $(s_{\min,\alpha}, d_{\max,\alpha})$, define event

$$\widehat{\mathcal{C}}^{\text{chance}}_{\alpha,\beta,H} := \left\{\widehat{W}_{\alpha,H} - e_{W,\alpha,H} > 0,\ \widehat{S}^+_\alpha - e_{S,\alpha} \geq s_{\min,\alpha},\ \widehat{\text{CVaR}}_\beta(D_\alpha) + e_{D,\alpha,\beta} \leq d_{\max,\alpha}\right\}.$$

Proposition 12 (Soundness of delayed tail chance certificate)**.** *If the three error bounds in $\widehat{\mathcal{C}}^{\text{chance}}_{\alpha,\beta,H}$ are valid, then this event implies*

$$W_{\alpha,H} > 0, \quad S^+_\alpha \geq s_{\min,\alpha}, \quad \text{CVaR}_\beta(D_\alpha) \leq d_{\max,\alpha}.$$

Proof. Immediate by one-sided bound arithmetic. □

8.5 Dual-layer tail-positive gate

Definition 39 (Pointwise lower confidence bound)**.** For point quantity $M_g(\mathbf{r}; z, t)$ with estimate $\widehat{M}_g$ and valid error radius $e_{M,\text{pt}}$, define

$$\text{LCB}(M_g) := \widehat{M}_g - e_{M,\text{pt}}.$$

Definition 40 (Ruin gate (hard fail-closed))**.** Given $(d_{\max}, \delta_{\text{cat}}, m_{\min})$, define

$$\widehat{\mathcal{G}}_{\text{ruin}} := \left\{\widehat{\text{CVaR}}_\beta(D_\alpha) + e_D \leq d_{\max},\ \widehat{\mathbb{P}}(D_\alpha \geq d_{\text{cat}}) + e_p \leq \delta_{\text{cat}},\ \text{LCB}(M_g) \geq m_{\min}\right\}.$$

Definition 41 (Discovery gate (bounded fail-open))**.** Let cumulative exploration budget be

$$B_t := \sum_{u \leq t} b_u.$$

Define

$$\widehat{\mathcal{G}}_{\text{disc}} := \widehat{\mathcal{G}}_{\text{exploit}}\ \vee\ \widehat{\mathcal{G}}_{\text{explore}},$$

with

$$\widehat{\mathcal{G}}_{\text{exploit}} := \left\{\widehat{W}_{\alpha,H} - e_W \geq 0\right\},$$

$$\widehat{\mathcal{G}}_{\text{explore}} := \left\{\widehat{W}_{\alpha,H} - e_W \geq -\epsilon_{\text{exp}},\ \widehat{\Gamma}_\alpha - e_\Gamma \geq \gamma_{\min},\ b_t \leq b_{\max},\ B_t \leq B_{\max}\right\}.$$

Definition 42 (Composite tail-positive admissibility gate)**.**

$$\widehat{\mathcal{G}}_{\text{tail+}} := \widehat{\mathcal{G}}_{\text{ruin}} \ \wedge \ \widehat{\mathcal{G}}_{\text{disc}}.$$

Proposition 13 (Soundness of dual-layer gate)**.** *If all confidence radii are valid, then*

$$\widehat{\mathcal{G}}_{\text{tail+}} \Rightarrow (\textit{ruin constraints hold}) \wedge (\textit{either certified exploit or bounded exploration}).$$

Definition 43 (No-negative-tail update rule)**.** A transition $\mathbf{r}_t \to \mathbf{r}_{t+1}$ is admissible only if

$$\widehat{\mathcal{G}}_{\text{tail+}}(\mathbf{r}_{t+1}) \text{ holds.}$$

Otherwise the transition is fail-closed (rejected or rolled back to latest admissible point).

8.6 Contamination-robust tail-positive certificate

Lemma 8 (CVaR contamination sensitivity on bounded support)**.** *Let $D \in [0,1]$ and $\widetilde{P} = (1-\varepsilon)P + \varepsilon Q$. Then*

$$\left|\text{CVaR}_{\beta,\widetilde{P}}(D) - \text{CVaR}_{\beta,P}(D)\right| \le \frac{\varepsilon}{1-\beta}.$$

Proof. By Rockafellar representation,

$$\text{CVaR}_\beta(D) = \min_{q\in[0,1]} \left[q + \frac{1}{1-\beta}\mathbb{E}[(D-q)_+] \right].$$

For any fixed q, Lemma 7 applied to $(D-q)_+ \in [0,1]$ gives

$$\left|\mathbb{E}_{\widetilde{P}}[(D-q)_+] - \mathbb{E}_P[(D-q)_+]\right| \le \varepsilon.$$

Hence the objective differs by at most $\varepsilon/(1-\beta)$ uniformly in q, and so do minima. □

Definition 44 (Robust tail lower/upper bounds)**.** Let $(\varepsilon_{W,\alpha}, \varepsilon_{S,\alpha}, \varepsilon_{D,\alpha}, \varepsilon_{p,\alpha}, \varepsilon_{M,\text{pt}}, \varepsilon_{\Gamma,\alpha})$ be contamination budgets at the candidate point. Define

$$\text{LB}^{\text{rob}}(W_{\alpha,H}) := \widehat{W}_{\alpha,H} - e_{W,\alpha,H} - \varepsilon_{W,\alpha},$$

$$\text{LB}^{\text{rob}}(S^+_\alpha) := \widehat{S}^+_\alpha - e_{S,\alpha} - \varepsilon_{S,\alpha},$$

$$\text{UB}^{\text{rob}}(\text{CVaR}_\beta(D_\alpha)) := \widehat{\text{CVaR}}_\beta(D_\alpha) + e_{D,\alpha,\beta} + \frac{\varepsilon_{D,\alpha}}{1-\beta},$$

$$\text{UB}^{\text{rob}}(p_{\text{cat},\alpha}) := \widehat{\mathbb{P}}(D_\alpha \ge d_{\text{cat}}) + e_p + \varepsilon_{p,\alpha},$$

$$\text{LCB}^{\text{rob}}(M_g) := \widehat{M}_g - e_{M,\text{pt}} - \varepsilon_{M,\text{pt}}.$$

$$\text{LB}^{\text{rob}}(\Gamma_\alpha) := \widehat{\Gamma}_\alpha - e_\Gamma - \varepsilon_{\Gamma,\alpha}.$$

Definition 45 (Robust ruin/discovery gates)**.**

$$\widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{ruin}}(\mathbf{r}) := \begin{Bmatrix} \mathrm{UB}^{\mathrm{rob}}(\mathrm{CVaR}_\beta(D_\alpha)) \le d_{\max,\alpha}, \\ \mathrm{UB}^{\mathrm{rob}}(p_{\mathrm{cat},\alpha}) \le \delta_{\mathrm{cat}}, \\ \mathrm{LCB}^{\mathrm{rob}}(M_g) \ge m_{\min} \end{Bmatrix},$$

$$\widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{exploit}}(\mathbf{r}) := \left\{ \mathrm{LB}^{\mathrm{rob}}(W_{\alpha,H}) \ge 0 \right\},$$

$$\widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{explore}}(\mathbf{r}) := \left\{ \mathrm{LB}^{\mathrm{rob}}(W_{\alpha,H}) \ge -\epsilon_{\mathrm{exp}},\ \mathrm{LB}^{\mathrm{rob}}(\Gamma_\alpha) \ge \gamma_{\min},\ b_t \le b_{\max},\ B_t \le B_{\max} \right\},$$

$$\widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{disc}}(\mathbf{r}) := \widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{exploit}}(\mathbf{r}) \ \vee\ \widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{explore}}(\mathbf{r}).$$

Definition 46 (Robust tail-positive gate)**.**

$$\widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{tail}+}(\mathbf{r}) := \widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{ruin}}(\mathbf{r}) \ \wedge\ \widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{disc}}(\mathbf{r}) \ \wedge\ \left\{ \mathrm{LB}^{\mathrm{rob}}(S^+_\alpha) \ge s_{\min,\alpha} \right\}.$$

Proposition 14 (Soundness under ε-contamination)**.** *Under the contamination model of Section 7, if $\widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{tail}+}(\mathbf{r})$ holds, then the corresponding clean-point quantities satisfy declared ruin and tail-positive requirements.*

Proof. Apply bounded-observable contamination shifts (Lemma 7) to point quantities, Corollary 1 to endpoint-difference quantities when used, and Lemma 8 to $\mathrm{CVaR}_\beta(D)$. □

8.7 Opportunity elasticities and long-gamma curvature

For endpoints with $\Delta S^+_\alpha \neq 0$, define delayed opportunity elasticity

$$\mathrm{TOE}^+_{\alpha,H} := \frac{\Delta W_{\alpha,H}}{\Delta S^+_\alpha}.$$

For depletion severity:

$$\mathrm{VaR}_\beta(D_\alpha) := \inf\{q : \mathbb{P}(D_\alpha \le q) \ge \beta\}, \qquad \mathrm{CVaR}_\beta(D_\alpha) := \mathbb{E}[D_\alpha \mid D_\alpha \ge \mathrm{VaR}_\beta(D_\alpha)].$$

These quantify left-tail cost of reserve failure [16, 17].

Let λ be scalarized operating coordinate and $\delta > 0$.

$$\Gamma_\alpha(\lambda;\delta) := W_{\alpha,H}(\lambda+\delta) - 2W_{\alpha,H}(\lambda) + W_{\alpha,H}(\lambda-\delta).$$

Definition 47 (Empirical curvature estimator)**.**

$$\widehat{\Gamma}_\alpha(\lambda;\delta) := \widehat{W}_{\alpha,H}(\lambda+\delta) - 2\widehat{W}_{\alpha,H}(\lambda) + \widehat{W}_{\alpha,H}(\lambda-\delta).$$

Proposition 15 (Symmetric-shock gain identity)**.** *Let ξ take values $\pm\delta$ with probability $1/2$. Then*

$$\mathbb{E}[W_{\alpha,H}(\lambda+\xi)] - W_{\alpha,H}(\lambda) = \frac{1}{2}\Gamma_\alpha(\lambda;\delta).$$

Hence $\Gamma_\alpha(\lambda;\delta) > 0$ implies positive expected gain from zero-mean symmetric perturbations of size δ.

Proof.

$$\mathbb{E}[W_{\alpha,H}(\lambda+\xi)] = \frac{W_{\alpha,H}(\lambda+\delta) + W_{\alpha,H}(\lambda-\delta)}{2}.$$

Subtract $W_{\alpha,H}(\lambda)$. □

8.8 Certified CVaR approximation on bounded support

Definition 48 (Rockafellar representation and grid estimator)**.** For $D \in [0,1]$ and $\beta \in (0,1)$ define

$$\Psi_\beta(q) := q + \frac{1}{1-\beta}\mathbb{E}[(D-q)_+], \qquad q \in [0,1].$$

Then

$$\mathrm{CVaR}_\beta(D) = \min_{q\in[0,1]} \Psi_\beta(q)[17].$$

For grid spacing $\eta \in (0,1]$, let

$$\mathcal{Q}_\eta := \{0, \eta, 2\eta, \dots, 1\},$$

$$\widehat{\Psi}_\beta(q) := q + \frac{1}{(1-\beta)n}\sum_{u=1}^{n}(D_u - q)_+, \qquad \widehat{\mathrm{CVaR}}_{\beta,\eta} := \min_{q\in\mathcal{Q}_\eta} \widehat{\Psi}_\beta(q).$$

Theorem 2 (Finite-sample CVaR certificate)**.** *Assume for each $q \in \mathcal{Q}_\eta$ that transformed stream $(D_u - q)_+$ is adapted, bounded in $[0,1]$, and satisfies EA with target $\mathbb{E}[(D-q)_+]$. Then with probability at least $1-\delta$,*

$$\sup_{q\in\mathcal{Q}_\eta} \left|\widehat{\Psi}_\beta(q) - \Psi_\beta(q)\right| \le \frac{1}{1-\beta}\sqrt{\frac{2\ln(2|\mathcal{Q}_\eta|/\delta)}{n}}.$$

Moreover, with

$$L_\beta := 1 + \frac{1}{1-\beta},$$

$$\left|\widehat{\mathrm{CVaR}}_{\beta,\eta} - \mathrm{CVaR}_\beta(D)\right| \le L_\beta\eta + \frac{1}{1-\beta}\sqrt{\frac{2\ln(2|\mathcal{Q}_\eta|/\delta)}{n}}.$$

Proof. Uniform bound follows by union bound over $\mathcal{Q}_\eta$ and Lemma 2 for each q. The second line combines this uniform bound with grid discretization: $\inf_{q\in\mathcal{Q}_\eta}\Psi_\beta(q) \le \inf_{q\in[0,1]}\Psi_\beta(q) + L_\beta\eta$. □

8.9 Certified exploratory headroom (critical viability condition)

Theorem 3 (Prefix-safe headroom for exploration)**.** *Consider horizon $t = 0, \dots, T-1$ with recursion*

$$M_{g,t+1} \ge M_{g,t} - \ell_t - g_t^{(-)} + g_t^{(+)},$$

where $\ell_t, g_t^{(-)}, g_t^{(+)} \ge 0$. If there exists $\underline{m} > 0$ such that

$$M_{g,0} \ge \underline{m},$$

and for every prefix $n = 1, \dots, T$,

$$M_{g,0} + \sum_{u=0}^{n-1} (g_u^{(+)} - \ell_u - g_u^{(-)}) \geq \underline{m},$$

then $M_{g,n} \geq \underline{m}$ *for all* $n = 0, \dots, T$. *Therefore exploratory perturbations remain feasible throughout the horizon with strictly positive reserve.*

Proof. Base case is $M_{g,0} \geq \underline{m}$. For $n \geq 1$, iterating the recursion gives

$$M_{g,n} \geq M_{g,0} + \sum_{u=0}^{n-1} (g_u^{(+)} - \ell_u - g_u^{(-)}) \geq \underline{m}.$$

□

Remark 13. Checking only terminal aggregate surplus is insufficient in general; prefix constraints are the exact all-time guard [18].

8.10 Safe niche search over contexts

Definition 49 (Opportunity-variance proxy)**.** For each context z, let $\widehat{V}_o(z)$ be the predeclared empirical variance estimator of tail opportunity payoff $I_\alpha o^\star$ computed from the context-z record subset (optionally clipped/robustified by published rules).

Definition 50 (Constrained niche selector)**.** Let $\mathcal{Z}_t$ be a declared feasible context set. Define

$$z_t^\star = \arg\max_{z \in \mathcal{Z}_t} \left[\widehat{W}_{\alpha,H}(z) + \beta_t \sqrt{\widehat{V}_o(z)} \right]$$

subject to

$$\mathrm{LCB}(M_g(z)) \geq m_{\min}, \qquad \widehat{\mathrm{CVaR}}_\beta(D_\alpha(z)) + e_D \leq d_{\max}.$$

Remark 14. The bonus term $\beta_t \sqrt{\widehat{V}_o(z)}$ induces controlled exploration pressure; constraints enforce viability.

Proposition 16 (Feasibility-preserving niche exploration)**.** *If selected* $z_t^\star$ *satisfies the above constraints at each step and prefix-safe headroom (Theorem 3) is maintained, exploration remains primary-feasible throughout the horizon.*

8.11 Tail-positive dominance criterion

Definition 51 (Tail-positive partial order)**.** For two operating points a, b, write $a \succeq_{\text{tail}} b$ if, for every α in predeclared set $\mathcal{A} \subset (0,1)$,

$$W_{\alpha,H}(a) \geq W_{\alpha,H}(b), \qquad S_\alpha^+(a) \geq S_\alpha^+(b), \qquad \mathrm{CVaR}_\beta(D_\alpha(a)) \leq \mathrm{CVaR}_\beta(D_\alpha(b)).$$

9 Scientific Marc Principle *Shinkidan* and Barbell Portfolio Control

9.1 Scientific operationalization

Definition 52 (Scientific Marc Principle (SMP))**.** For a candidate update $\mathbf{r} \to \mathbf{r}'$ and declared (α, β, H), fix a reporting regime (standard or robust). SMP accepts the update iff all hold:

(i) **Ruin hard-bound:** $\widehat{\mathcal{G}}_{\text{ruin}}(\mathbf{r}')$ holds in standard mode, or $\widehat{\mathcal{G}}^{\text{rob}}_{\text{ruin}}(\mathbf{r}')$ holds in robust mode.

(ii) **Evidence discipline:** TEPP completeness obligations hold for all declared α.

(iii) **Opportunity discipline:** $\widehat{\mathcal{G}}_{\text{disc}}(\mathbf{r}')$ holds in standard mode, or $\widehat{\mathcal{G}}^{\text{rob}}_{\text{disc}}(\mathbf{r}')$ holds in robust mode.

(iv) **Budget discipline:** exploration allocation and cumulative budget constraints hold.

Remark 15. SMP is not metaphysical. It is a falsifiable update admissibility rule over declared observables.

9.2 From dual gate to portfolio structure

The dual-layer mechanism is a barbell portfolio over updates [25]: a *safety leg* that blocks ruin and an *opportunity leg* that harvests bounded convex upside from tails. This differs from mean-variance smoothing [26]; the design objective is explicit ruin-bounded convexity.

Safety leg (ruin prevention). The safety leg is enforced by the hard gate $\widehat{\mathcal{G}}_{\text{ruin}}$ (catastrophic-tail bounds and headroom constraints). Any candidate failing $\widehat{\mathcal{G}}_{\text{ruin}}$ is rejected (fail-closed).

Opportunity leg (convex upside harvesting). The opportunity leg is governed by delayed tail value and curvature certificates. Tail candidates are preserved first (TEPP), then re-evaluated with delay; $W_{\alpha,H}$ captures net delayed opportunity and Γ_α captures second-order convex gain under symmetric shocks.

9.3 Barbell controller

Definition 53 (Barbell score and controller)**.** Let $\omega_t \in [0, \omega_{\max}]$ be exploration allocation at epoch t, with context decision z_t, incremental exploration budget process $b_t \geq 0$, and cumulative budget

$$B_t := \sum_{u \leq t} b_u.$$

Define

$$\mathcal{B}_t(\omega_t, z_t) = (1 - \omega_t)\, U_t^{\text{safe}} + \omega_t\, U_t^{\text{opp}}(z_t),$$

where

$$U_t^{\text{safe}} := \text{LCB}(M_g(\mathbf{r}_t; z_t, t)) - m_{\min},$$

$$U_t^{\text{opp}}(z_t) := \text{LCB}(W_{\alpha,H}(\mathbf{r}_t; z_t, t)) + \kappa_\Gamma\ \text{LCB}(\Gamma_\alpha(\lambda_t; \delta)) - \epsilon_{\text{exp}},$$

with $\kappa_\Gamma \geq 0$. The controller chooses admissible (ω_t, z_t) maximizing $\mathcal{B}_t$ under

$$\widehat{\mathcal{G}}_{\text{ruin}} \text{ holds}, \qquad b_t \leq b_{\max}, \qquad B_t \leq B_{\max}.$$

Definition 54 (No-middle admissibility)**.** A candidate update is admissible iff one of the following holds:

(i) **Pure safety mode:** $\omega_t = 0$ and $\widehat{\mathcal{G}}_{\text{ruin}}$ holds.

(ii) **Bounded opportunity mode:** $0 < \omega_t \le \omega_{\max}$, $\widehat{\mathcal{G}}_{\text{ruin}}$ holds, and the exploration clause of $\widehat{\mathcal{G}}_{\text{disc}}$ holds with budget constraints.

All other candidates are rejected. In particular, unconstrained intermediate-risk updates are excluded.

Proposition 17 (Ruin-bounded convex exploration)**.** *Assume valid error radii and*

$$\widehat{W}_{\alpha,H} - e_W \ge -\epsilon_{\text{exp}}, \qquad \widehat{\Gamma}_\alpha - e_\Gamma \ge \gamma_{\min} > 0.$$

Then, for symmetric shocks $\xi \in \{+\delta, -\delta\}$ with probability $1/2$,

$$\mathbb{E}[W_{\alpha,H}(\lambda + \xi)] - W_{\alpha,H}(\lambda) = \frac{1}{2}\Gamma_\alpha(\lambda; \delta) \ge \frac{\gamma_{\min}}{2}.$$

If simultaneously $\widehat{\mathcal{G}}_{\text{ruin}}$ holds, the exploratory leg has certified nonnegative second-order upside while catastrophic downside remains bounded by ruin constraints.

Proof. By the symmetric-shock identity in Section 8,

$$\mathbb{E}[W_{\alpha,H}(\lambda + \xi)] - W_{\alpha,H}(\lambda) = \frac{1}{2}\Gamma_\alpha(\lambda; \delta).$$

The one-sided confidence premise yields $\Gamma_\alpha(\lambda; \delta) \ge \gamma_{\min}$, hence the first claim. The second claim follows from conjunction with $\widehat{\mathcal{G}}_{\text{ruin}}$. □

Theorem 4 (SMP soundness under declared assumptions)**.** *Under A0–A6 and valid confidence radii, acceptance by SMP implies:*

(a) declared ruin constraints hold for clean quantities,

(b) delayed opportunity is either certified exploitative or bounded exploratory,

(c) tail evidence obligations are auditably preserved,

(d) update exposure remains within predeclared budget.

Proof. Immediate from conjunction of gate soundness results, TEPP auditability, and budget constraints by definition. □

10 Context Mixture and Schedule Non-Identifiability

Lemma 9 (Mixture derivative)**.** *Let*

$$R_{\text{mix}}(\mathbf{r}) = p(\mathbf{r})R_0(\mathbf{r}) + (1 - p(\mathbf{r}))R_1(\mathbf{r}),$$

for differentiable scalar schedule $t \mapsto \mathbf{r}(t)$. Then

$$\frac{d}{dt}R_{\text{mix}} = p\frac{dR_0}{dt} + (1-p)\frac{dR_1}{dt} + \frac{dp}{dt}(R_0 - R_1).$$

Proposition 18 (Finite-difference mixture identity)**.** *For endpoints* $(\mathbf{r}_0, \mathbf{r}_1)$*:*

$$\Delta R_{\text{mix}} = p_1 \Delta R_0 + (1 - p_1)\Delta R_1 + (p_1 - p_0)(R_{0,0} - R_{1,0}),$$

equivalently

$$\Delta R_{\text{mix}} = p_0 \Delta R_0 + (1 - p_0)\Delta R_1 + (p_1 - p_0)(R_{0,1} - R_{1,1}).$$

Remark 16. Without replayable mixture weights, endpoint effects can be compositional artifacts rather than capability movement.

11 Identifiability and No-Meta Invariance

Definition 55 (Base-relative functional)**.** Given prompt law $\mathcal{D}$ and divergence-like $d_{\mathcal{D}}$:

$$\mathcal{A}(\mathbf{r}) := d_{\mathcal{D}}(\pi_{\mathbf{r}}, \pi_0),$$

where π_0 is latent base policy.

Theorem 5 (Non-identifiability without base anchor)**.** *Fix observable family* $\{\pi_{\mathbf{r}}\}_{\mathbf{r}\in\mathcal{R}}$*. Without observing* π_0 *or adding external identifying restrictions linking* π_0 *to observables,* $\mathcal{A}(\mathbf{r})$ *is not identifiable from observable data generated by deployed policies.*

Sketch. Construct two latent decompositions yielding identical deployed conditional laws for all observables but distinct latent π_0; then $\mathcal{A}(\mathbf{r})$ differs while observables coincide. □

Proposition 19 (Observational neutrality)**.** *If implementations* (A) *and* (B) *satisfy*

$$\pi_{\mathbf{r}}^{(A)}(\cdot \mid x, z, t) = \pi_{\mathbf{r}}^{(B)}(\cdot \mid x, z, t) \quad \forall (x, \mathbf{r}, z, t),$$

then every Layer-A observable quantity in this paper is identical under (A) *and* (B)*.*

12 Operational Protocol (Normative Layer-A Rules)

This section states implementation-facing rules in normative form.

R1. A deployment **MUST** publish sampling commitments, stopping rule commitments, threshold-process descriptors, and evaluator descriptors before observing evaluated outcomes.

R2. For each declared $\alpha \in \mathcal{A}$, records with $I_{\alpha,i} = 1$ **MUST** satisfy preserve-before-judge commitment (TEPP).

R3. Non-tail controls **MUST** be preserved at declared random rate $\pi_{\text{neg}} > 0$.

R4. Any claim of "overall improvement" **MUST** pass $\widehat{\mathcal{G}}_{\text{claim}}$.

R5. Any update claim invoking beneficial tail behavior **MUST** pass $\widehat{\mathcal{G}}_{\text{tail}+}$ (or robust counterpart under declared contamination).

R6. If $\widehat{\mathcal{G}}_{\text{tail}+}$ fails at a candidate update, the update **MUST** be fail-closed (reject/rollback).

R7. Ratio-based RCE reporting **MUST** be restricted to certified denominator-margin domains; outside these domains, stabilized proxies and plateau flags **MUST** be reported.

R8. Transcript publication **MUST** include append-only epoch-root commitments and quorum certificates with consistency-proof retrieval route.

R9. Delayed opportunity metrics **MUST** declare horizon H, provenance rules, missingness model, and whether DR correction is used.

R10. Interpretive causal claims **MUST** be explicitly labeled Layer B and bound to declared instrument assertions.

R11. Exploration allocation ω_t **MUST** satisfy $\omega_t \in [0, \omega_{\max}]$ and budget constraints ($b_t \le b_{\max}, \sum_{u \le t} b_u \le B_{\max}$) whenever opportunity mode is active.

R12. Update admissibility **MUST** satisfy no-middle admissibility: pure safety mode or bounded opportunity mode only; uncontrolled intermediate-risk updates are **MUST-NOT**.

R13. Audit-selector unpredictability mechanism (**public randomness** or **VRF proof**) **MUST** be verifiable from the transcript.

R14. Canonicalization for hash/signature inputs **MUST** be RFC 8785 JCS (or a byte-equivalent declared deterministic canonicalization), and hash inputs **MUST** include explicit domain-separation tags.

R15. Any stated commitment to the Scientific Marc Principle *(Shinkidan)* **MUST** be interpreted as the conjunction of SMP clauses in Section 9, not as a non-falsifiable narrative.

13 Auditable Reporting Record

Definition 56 (Record requirements)**.** Each reported endpoint comparison includes:

- transcript handle (or digest + retrieval route),
- sample sizes for all measured streams,
- report-plan parameters $(\bar{\delta}, N_{\max}, Z, T, m, k, q)$ and per-test budget allocation,
- declared EA assumptions by stream and endpoint,
- declared contamination budgets $(\varepsilon_M, \varepsilon_R, \varepsilon_C, \varepsilon_h, \varepsilon_c, \varepsilon_{W,\alpha}, \varepsilon_{S,\alpha}, \varepsilon_{D,\alpha}, \varepsilon_{p,\alpha}, \varepsilon_{M,\text{pt}}, \dots)$,
- curvature contamination budget $\varepsilon_{\Gamma,\alpha}$ and one-sided curvature radius e_Γ,
- raw endpoint estimates and differences for

$$R, C, M_g, M_h, M_c, \bar{M}_g, \bar{M}_h, \bar{M}_c, W_{\alpha,H}, S^+_\alpha, \text{CVaR}_\beta(D_\alpha),$$

- admissibility status for every reported ratio (standard and robust),

- auxiliary-essential status and supporting certificates,
- transfer-gap certificate status and all-channel gate status,
- TEPP completeness $\mathrm{TCR}_{\alpha,t}$ and negative-control preservation statistics,
- per-α delayed tail chance certificate status $\widehat{\mathcal{C}}^{\mathrm{chance}}_{\alpha,\beta,H}$,
- dual-layer gate status $\widehat{\mathcal{G}}_{\mathrm{tail}+}$ and robust counterpart $\widehat{\mathcal{G}}^{\mathrm{rob}}_{\mathrm{tail}+}$,
- SMP status (clause-by-clause pass/fail with margins),
- barbell controller status: $(\omega_t, b_t, B_t, \mathcal{B}_t)$ and no-middle admissibility outcome,
- SNC status (law/score separation declaration),
- epoch-root chain commitments, quorum certificates, and consistency-proof route,
- explicit Layer-B instrument assertions for interpretive claims.

14 Limitations

- Observable-only validity still depends on evaluator integrity/calibration quality [13].
- Guarantees are conditional on declared sampling plans, contamination budgets, stopping rules, and EA assumptions.
- Delayed audit selection reduces but does not eliminate adaptive gaming under highly expressive attacker policies.
- Tail slices are data-sparse; finite-sample uncertainty remains unavoidable.
- Goodhart/Campbell pressures are constrained, not eliminated [11, 12].
- Causal interpretation remains Layer-B and conditional on published instrument assertions.
- Dual-layer gating can still be conservative if uncertainty radii are loose.
- Tail positivity is decision-theoretic and estimator-dependent; poor metric design for (o, h, a) can still mis-rank policies.
- Delayed opportunity signals can be spoofed without strong provenance and anti-sybil channel design.
- Cryptographic commitment guarantees integrity, not semantic correctness of committed content.
- SMP is operationally strong but remains only as valid as its declared observables and threat model coverage.

15 Conclusion

We presented an observable-only, reference-free framework for no-meta settings using dynamic feasible sets, finite-difference elasticities, auxiliary-witness falsification, and contamination-aware certification.

Compared with single-channel metric reporting, the framework adds explicit Goodhart and attack resistance: dual-bundle public/audit evaluation with transfer-gap falsification, canary-informed all-channel fail-closed gating, and append-only quorum-signed transcript commitments.

The tail component is extended from static certification to lifecycle control: commit-first tail evidence preservation (TEPP), delayed opportunity valuation with DR correction, reserve/depletion geometry, long-gamma curvature, and a dual-layer tail-positive gate that combines hard ruin protection with bounded exploration.

The Scientific Marc Principle *(Shinkidan)* is formalized as a falsifiable operational doctrine: maximize bounded convex upside only under hard ruin constraints, mandatory evidence preservation, and explicit budget discipline.

Under declared assumptions and precommitted protocol, the method supports a coherent stance: quantify constraints without latent-base mythology, reject single-channel narratives when witness channels disagree, preserve potentially valuable tail evidence for future replayable judgment, and formalize safety–wildness coexistence as an auditable implementation protocol rather than a metaphor.

A Cryptographic Replay/Reveal Protocol (Implementation Appendix)

A.1 VRF-based audit selector

Audit bundle selection can be made unpredictable yet publicly verifiable:

$$j_t = \mathrm{VRF}_{sk}(u_t) \bmod K,$$

with proof π_t such that anyone verifies

$$\mathrm{Verify}_{pk}(u_t, j_t, \pi_t) = 1.$$

This matches standard VRF usage [21].

A.2 Canonicalization and domain separation

To prevent hash disagreement across implementations, serialized objects MUST be canonicalized with JCS (RFC 8785) or declared byte-equivalent canonicalization [22]. Hash preimages MUST be domain-separated:

$$h = H(\texttt{ContextTag} \,\|\, \mathrm{JCS}(\texttt{Object}))\,.$$

For instance:

$$\texttt{ContextTag} \in \{\texttt{RCE-LeafV1}, \texttt{RCE-EpochV1}, \texttt{RCE-RevealV1}\}.$$

A.3 Delayed reveal for tail payloads

At commit time t, payload is sealed:

$$c_i = \mathrm{Enc}_{pk_t}(\mathrm{payload}_i), \qquad h_i = H(\mathrm{payload}_i).$$

At reveal time $t + H$, publish $(c_i, \mathrm{DecProof}_i, \mathrm{payload}_i)$ and verify:

$$H(\mathrm{payload}_i) \stackrel{?}{=} h_i.$$

Time-locked or VDF-assisted release policies are admissible design choices [23, 24].

A.4 Canonical leaf and transcript schemas (illustrative)

```
{
  "LeafV1": {
    "idx": "uint64",
    "x_hash": "hex32",
    "z": "string",
    "t": "uint64",
    "r": "vector<float>",
    "y_hash": "hex32",
    "eval_out_hash": "hex32",
    "meta_hash": "hex32"
  },
  "leaf_hash": "H('RCE-LeafV1' || JCS(LeafV1))"
}

{
  "DelayedRevealV1": {
    "idx": "uint64",
    "ciphertext": "bytes",
    "payload_hash": "hex32",
    "reveal_epoch": "uint64",
    "reveal_payload": "bytes|null",
    "dec_proof": "bytes|null"
  }
}

{
  "AuditSelectorV1": {
    "epoch": "uint64",
    "vrf_input": "bytes",
    "selector_mod": "uint32",
    "selector_output": "uint32",
    "vrf_proof": "bytes",
```

```
    "vrf_pk": "bytes"
  }
}
```

A.5 TEPP audit checklist

A verifier checks, for each declared (α, t):

(i) recompute $I_{\alpha,i}$ from declared cross-fitted protocol;

(ii) verify required leaf membership for all $I_{\alpha,i} = 1$;

(iii) verify negative-control inclusion for sampled $I_{\alpha,i} = 0$;

(iv) verify epoch consistency proofs in the append-only log;

(v) verify reveal transcripts and hash equality for delayed signals.

A.6 Minimal replay pseudocode (illustrative)

```
for epoch t:
  verify_quorum_certificate(ER_t)
  verify_consistency_proof(ER_{t-1}, ER_t)
  reconstruct_eval_outputs(Tr_t, EvalDesc_t)
  for alpha in A:
    recompute_tail_indicator(I_alpha)
    check_TEPP_membership(I_alpha, L_t)
    check_negative_control_sampling(L_t)
  compute endpoints: R, C, M_g, M_h, M_c, W_{alpha,H}, S^+_alpha, CVaR_beta(D_alpha)
  check claim gate, tail+ gate, SMP clauses
  emit auditable report row
```

References

[1] Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022. arXiv:2203.02155.

[2] Tiansheng Huang, et al. The Safety Tax: Safety Alignment Makes Your Large Reasoning Models Less Reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

[3] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An Over-Refusal Benchmark for Large Language Models. *arXiv preprint arXiv:2405.20947*, 2024.

[4] Patrick Chao, et al. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. *arXiv preprint arXiv:2404.01318*, 2024.

[5] Alexandra Souly, et al. A StrongREJECT for Empty Jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

[6] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 117–126, 2015. arXiv:1411.2664.

[7] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi:10.1126/science.aaa9375.

[8] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[9] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[10] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[11] Charles A. E. Goodhart. Problems of Monetary Management: The U.K. Experience. In *Papers in Monetary Economics*, Reserve Bank of Australia, 1975.

[12] Donald T. Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90, 1979.

[13] Yaniv Ovadia, Emily Fertig, Jie Ren, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. arXiv:1906.02530.

[14] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

[15] David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–110, 1975.

[16] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

[17] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

[18] Jean-Pierre Aubin. *Viability Theory*. Birkhäuser, 1991.

[19] Ralph C. Merkle. A certified digital signature. In *Advances in Cryptology — CRYPTO'89*, LNCS 435, pages 218–238. Springer, 1989.

[20] Ben Laurie, Eran Messeri, and Rob Stradling. Certificate Transparency Version 2.0. RFC 9162, IETF, December 2021.

[21] Sharon Goldberg, Leonid Reyzin, Dimitrios Papadopoulos, and Jan Včelák. Verifiable Random Functions (VRFs). RFC 9381, IRTF, 2023.

[22] Anders Rundgren, Bret Jordan, and Samuel Erdtman. JSON Canonicalization Scheme (JCS). RFC 8785, RFC Editor (Independent Submission), 2020.

[23] Ronald L. Rivest, Adi Shamir, and David A. Wagner. Time-lock puzzles and timed-release crypto. Technical Report, MIT/LCS, 1996.

[24] Dan Boneh, Joseph Bonneau, Benedikt Bünz, and Ben Fisch. Verifiable Delay Functions. In *Advances in Cryptology – CRYPTO 2018*, LNCS 10991, pages 757–788, 2018.

[25] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, 2012.

[26] Harry Markowitz. Portfolio Selection. *The Journal of Finance*, 7(1):77–91, 1952.