Revelation Control

Qinyou Wang

Abstract

Revelation Control is the problem of choosing priced interventions that reveal hidden state only insofar as the revealed distinctions can change a consequential decision, while accounting separately for any useful progress created by the intervention itself. We develop this theory for learning systems, where states equivalent under declared current information can respond differently to future training and favor different actions. The framework defines decision-sufficient revelation and revelation depth, separates pure information value from productive reuse, embeds static Bayes refinement into state-dependent continuation value, and gives an exact cost-adjusted factorization criterion: an additional shallow coordinate is decision-nonredundant only when states sharing a scalar summary lie on opposite sides of the priced STOP/CONTINUE boundary. We also give a target-independent protocol for model-specific instantiation and prove that bounded stop–flip risk alone cannot certify positive expected utility under unrestricted severity. Across Qwen2.5-7B and Mistral-7B-v0.3, deeper future-learning probes have positive decision value and productive reuse yields strict equal-compute utility advantages. Qwen additionally provides evidence for a decision-nonredundant shallow revealability regime; in Mistral, a scalar continuation architecture fit only on an independent development panel retains positive familywise-adjusted lower bounds on a disjoint target panel, consistent with scalar decision sufficiency within the tested architecture family and resolution. The evidence supports structural rather than numerical transfer: the decision theory, cost accounting, continuation logic, and evaluation protocol transport, while empirical proxies, coefficients, thresholds, and even the required shallow state dimension may be system-specific.

1 Introduction

Sequential decision systems are routinely controlled through summaries of their current state. In learning systems these include present losses, held-out metrics, gradients, optimizer telemetry, and short diagnostic trajectories. Such summaries are decision-sufficient only when the distinctions they discard cannot change the action that should be taken. A learner may instead occupy execution states that look equivalent under the declared current information yet react differently to the same future training intervention.

Our companion work, *Fiber Fingerprints of Hidden Learning-State Dynamics* [40], establishes the predictive premise within its declared scope: present behavior need not be a sufficient statistic for declared future learning. The decision theory developed here does not depend on that paper's specific geometric machinery; it takes hidden decision-relevant state as the object to be revealed and asks the downstream question:

When does hidden learning-state structure become actionable decision information?

We call the resulting control problem *Revelation Control*: the controller chooses not only which action to take, but which future-learning interventions to instantiate, how much decision-relevant state to reveal, and which tested future to promote into execution.

The distinction between prediction and decision is essential. A hidden variable can improve prediction without changing the optimal action; a refined observation can increase oracle Bayes value while a finite-sample learner fails to extract it; and a training probe can both reveal information and leave behind useful computation. A decision theory for future-learning probes must therefore keep information, execution technology, learnability, and cost separate. Once refinement depth is allowed to vary, the same value functional becomes a control law over how much future information to acquire.

Partial training itself is not new. Freeze–Thaw Bayesian optimization uses partial learning curves to decide whether to pause or resume models [37]; learning-curve extrapolation can terminate weak runs before full training [7]; successive resource-allocation methods exploit intermediate performance to concentrate computation on promising candidates [18, 25]; Population Based Training reuses and adapts live training trajectories [17]; and DynaMiCS uses short domain-specific probes to estimate local cross-domain effects before selecting a constrained fine-tuning mixture [14]. Nor are Bayes value of information, Bayesian experimental design, active sensing, partial observability, or dual control new ideas [5, 6, 11, 15, 21, 26, 39]. The distinctive step here is to turn *future learning itself* into a decision-relative experiment on hidden learner state, and to exploit the fact that the tested path can remain useful computation. Although the empirical instantiation in this paper is model training, the underlying decision structure is broader: a priced,

controlled probe can change a system, reveal otherwise hidden decision-relevant state through its response, and leave reusable progress if the tested path is selected. Section 15 develops this extension while keeping all empirical claims confined to the learning systems actually studied.

1.1 Contributions

The paper makes five coupled contributions.

1. **Decision-sufficient revelation.** We characterize when refined information changes the Bayes decision quotient, derive exact binary aliasing value, and identify a local revelation depth at which currently hidden decision-relevant directions become observable to an admissible future-learning probe.
2. **Productive revelation technology.** We distinguish discard-and-restart probing from probe-and-promote revelation, derive the equal-budget depth geometry, and separate pure information gain from the value of having already executed part of the selected training path.
3. **Adaptive revelation control.** We prove that population Bayes refinement value is exactly the expectation of a conditional state-dependent revelation value. We then derive an exact cost-adjusted scalar-control gap: a second shallow coordinate is decision-nonredundant precisely when a positive-mass scalar fiber contains states on both sides of the priced STOP/CONTINUE boundary. In the binary location–scale specialization this yields a critical-revealability crossing criterion, so one-dimensional control can remain optimal even when revealability varies conditionally but never changes the meta-action.
4. **Learnability and certification boundaries.** We separate oracle revelation value from approximation, estimation, acquisition, and promotion terms. For adaptive depth, we identify a sharp boundary: bounded stop–flip risk alone cannot certify positive expected utility under unrestricted severity; a severity, moment, or integrable-tail condition is the missing object.
5. **Cross-model Transformer validation and a reusable instantiation procedure.** In a first 7B Transformer family, fixed-depth revelation, structural decision refinement, productive reuse, equal-compute frontier advantage, and adaptive safe-compute behavior are supported on independent panels. In an independently instantiated Mistral-7B-v0.3 family, positive deeper-revelation value and productive equal-compute utility reproduce on an independent two-bank panel. Within the finite pre-existing continuation-architecture family, a scalar architecture fit only on separate development data retains positive familywise-adjusted lower bounds on the target panel. The two families therefore reproduce the same Revelation-Control structure, while Qwen provides evidence that extra shallow revealability information is decision-nonredundant at the declared compute price and Mistral is consistent with scalar decision sufficiency within the tested architecture family and resolution.

1.2 Scope and organization

Scope of claims. The paper does not claim unrestricted dominance over every current-information policy, every probing algorithm, every model family, or every deployment cost model. The finite-library result is exactly that: finite-library. The structural refinement claim is conditional on a declared shallow ambiguity regime rather than an unrestricted Bayes comparison over complete shallow information. The comparative claim uses one strengthened DynaMiCS-style short-probe frontier under a common policy-visible update budget. Cross-model transport is structural rather than parametric: different model families may use different legal revealability proxies, fitted coefficients, or stopping thresholds, and the theory explicitly allows scalar decision sufficiency whenever no positive-mass scalar fiber crosses the cost-adjusted continuation boundary. Fully tail-robust population utility certification remains conditional on an explicit severity, moment, or tail class.

Organization. Sections 2 and 3 identify the decision quotient and the depth at which hidden distinctions become decision sufficient. Section 4 develops productive fork–probe–promote execution. Section 5 closes static refinement into state-dependent control, proves the scalar-control factorization criterion, and gives the model-specific instantiation protocol. Section 6 separates oracle value from finite-sample extraction and certification; the related-work section then fixes the novelty boundary before the empirical instantiation. Sections 8 to 13 evaluate the theory in two Transformer families, and Section 14 summarizes the cross-model invariants and regime-specific differences. Section 15 then identifies prospective learning, computational, physical, scientific, operational, and high-stakes domains in which the same theory-first instantiation procedure may be useful. Limitations and conclusions follow; the appendices collect proofs, statistical inference, and experimental protocol details.

2 Decision value and decision factorization

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{A}$ a finite action set, and $Q_a \in L^1$ the terminal utility of action a. For an information sigma-field $\mathcal{I} \subseteq \mathcal{F}$, define the observation-relative Bayes value

$$V_B(\mathcal{I}) = \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[Q_a \mid \mathcal{I}]\right]. \tag{2.1}$$

If $\mathscr{H} \subseteq \mathscr{G}$, then Jensen's inequality for the finite maximum gives the standard refinement monotonicity

$$\mathcal{I}(\mathscr{G} \mid \mathscr{H}) := V_B(\mathscr{G}) - V_B(\mathscr{H}) \geq 0. \tag{2.2}$$

For an event $A \in \mathscr{H}$ with $\mathbb{P}(A) > 0$, define

$$V_B(\mathcal{I}; A) = \mathbb{E}\left[\max_{a\in\mathcal{A}} \mathbb{E}[Q_a \mid \mathcal{I}] \,\middle|\, A\right], \tag{2.3}$$

and for a measurable policy π, write $\mathcal{V}(\pi; A) = \mathbb{E}[Q_\pi \mid A]$ and $\mathcal{V}(\pi) = \mathbb{E}[Q_\pi]$.

2.1 Decision factorization

Definition 2.1 (Decision factorization)**.** For $\mathscr{H} \subseteq \mathscr{G}$, the $\mathscr{G}$-optimal decision *factorizes through* $\mathscr{H}$ if there exists an $\mathscr{H}$-measurable policy π_H satisfying

$$\pi_H(\omega) \in \arg\max_{a\in\mathcal{A}} \mathbb{E}[Q_a \mid \mathscr{G}](\omega) \quad \text{a.s.} \tag{2.4}$$

The definition concerns the action-relevant quotient of the refined information, not reconstruction of the full latent state.

Proposition 2.2 (Exact finite-action strictness criterion)**.** *For finite $\mathcal{A}$ and integrable utilities,*

$$V_B(\mathscr{G}) = V_B(\mathscr{H}) \tag{2.5}$$

if and only if the $\mathscr{G}$-optimal decision factorizes through $\mathscr{H}$. Hence strict refinement value occurs exactly when no $\mathscr{H}$-measurable policy is $\mathscr{G}$-Bayes optimal almost surely.

This criterion is deliberately decision-relative. A refinement can contain predictive information while having zero value for the declared action menu; conversely, only a small quotient of a very high-dimensional latent state may be required to change the optimal action. This perspective is compatible with classical comparison of experiments and value-of-information theory [4, 5, 15], but our later use is dynamic because the act of acquiring future-learning information can also create reusable computation.

2.2 Dynamic decision-factorization obstruction

At time t, let $\nu_t(a)$ be the current coarse-information action value before accounting for unresolved downstream distinctions, and let $\nu_t^\star = \max_a \nu_t(a)$. Define the current opportunity cost

$$d_t(a) = \nu_t^\star - \nu_t(a) \geq 0. \tag{2.6}$$

Let $L_t \geq 0$ be the immediate value lost because the coarse representation aliases current decision-relevant states, and let $\omega_t(a) \geq 0$ be the continuation obstruction that remains after taking action a. With continuation factor $\gamma \geq 0$, define the total dynamic obstruction by comparing the best continuation-aware action with the coarse current benchmark:

$$O_t^{\mathrm{DF}} := L_t + \max_{a\in\mathcal{A}}\{\nu_t(a) + \gamma\omega_t(a)\} - \nu_t^\star. \tag{2.7}$$

Proposition 2.3 (Dynamic decision-factorization decomposition)**.** *The obstruction admits the exact decomposition*

$$\boxed{O_t^{\mathrm{DF}} = L_t + \max_{a\in\mathcal{A}}\{\gamma\omega_t(a) - d_t(a)\}.} \tag{2.8}$$

Moreover,

$$O_t^{\mathrm{DF}} = 0 \iff L_t = 0 \quad \textit{and} \quad \gamma\omega_t(a) \le d_t(a)\ \forall a. \tag{2.9}$$

In particular, if $\gamma > 0$ and any current coarse-optimal action has $\omega_t(a) > 0$, then the coarse representation is dynamically insufficient even when $L_t = 0$.

The theorem is a decomposition, not a replacement for POMDP or dual-control theory. It isolates a failure mode useful for learning systems: a summary may support the correct action *now* yet collapse distinctions that become decision-relevant after the chosen action enters a new training state.

3 Decision-sufficient revelation and revelation depth

A learning execution state is broader than model parameters:

$$X_t = (\theta_t, m_t, v_t, \text{history}, \text{RNG}, \text{scheduler}, \ldots). \tag{3.1}$$

The legal current information is

$$\mathscr{H}_t = \sigma(C_t), \tag{3.2}$$

where C_t may contain all prospectively legal current losses, readouts, history summaries, optimizer summaries, and telemetry. "Current-only" therefore does not mean behavior-only. A controlled future-learning probe Y_h induces

$$\mathscr{G}_h = \mathscr{H}_t \vee \sigma(Y_h). \tag{3.3}$$

The companion work establishes predictive non-sufficiency but not decision value [40]. The present section asks which parts of a current-behavior fiber must be revealed to factor the terminal decision quotient. For the binary action menu $\{\textsc{Keep}, \textsc{Xi}\}$, let $q_a(x)$ denote the state-conditional expected terminal utility of action a under the declared evaluation technology, and define the terminal action gap

$$g(x) = q_{\textsc{Xi}}(x) - q_{\textsc{Keep}}(x). \tag{3.4}$$

A pointwise decision-aliasing witness consists of x_1, x_2 such that

$$C(x_1) = C(x_2), \qquad g(x_1)g(x_2) < 0. \tag{3.5}$$

The full state need not be recovered: the probe only needs to refine the *decision quotient*, i.e., the distinctions necessary to select an optimal action.

Definition 3.1 (Decision-relevant revelation)**.** A future-learning probe Y_h is decision-relevant relative to $\mathscr{H}_t$ if

$$V_B(\mathscr{G}_h) > V_B(\mathscr{H}_t). \tag{3.6}$$

3.1 Exact binary aliasing identity

Let $D = Q_{\textsc{Xi}} - Q_{\textsc{Keep}}$, $m_H = \mathbb{E}[D \mid \mathscr{H}]$, and $m_G = \mathbb{E}[D \mid \mathscr{G}]$. Define the $\mathscr{H}$-measurable random variables

$$a_H = \mathbb{E}[(m_G)_+ \mid \mathscr{H}], \qquad b_H = \mathbb{E}[(-m_G)_+ \mid \mathscr{H}]. \tag{3.7}$$

By the tower property, $m_H = a_H - b_H$ almost surely.

Theorem 3.2 (Binary aliasing identity)**.** *For the binary action menu, the exact refinement value is*

$$\boxed{V_B(\mathscr{G}) - V_B(\mathscr{H}) = \mathbb{E}[\min\{a_H, b_H\}].} \tag{3.8}$$

Thus strict value is present exactly when the coarse information has positive probability of retaining refined posterior mass on both sides of the terminal decision boundary. More quantitatively, let $A \in \mathscr{H}$ with $\mathbb{P}(A) > 0$. If for some $\delta > 0$ and $\alpha, \beta > 0$,

$$\mathbb{P}(m_G \ge \delta \mid \mathscr{H}) \ge \alpha, \qquad \mathbb{P}(m_G \le -\delta \mid \mathscr{H}) \ge \beta \qquad \textit{a.s. on } A, \tag{3.9}$$

then

$$\boxed{V_B(\mathcal{G}) - V_B(\mathcal{H}) \geq \mathbb{P}(A)\,\delta\,\min(\alpha, \beta) > 0.} \tag{3.10}$$

Equation (3.8) is the distributional form of the alias-cell argument: within each coarse-information fiber, refined states can favor opposite terminal actions, and only the smaller of the two conditional magnitude-weighted sign masses creates irreducible coarse decision regret. Predictive variation that never crosses the declared decision boundary has zero value for this binary action menu. No positive-probability atom of the coarse information is required.

3.2 Local geometry and revelation depth

Let Φ_c denote the complete legal frontier information map after a short probe depth c (current information plus the frontier summary), let g be the terminal action-gap function defined above, and let Y_h be a deeper revelation observation.

Proposition 3.3 (Local aliasing and revelation)**.** *Suppose Φ_c, g, and Y_h are continuously differentiable near x_0, $g(x_0) = 0$, and Φ_c has locally constant rank. If there exists*

$$v \in \ker D\Phi_c(x_0) \tag{3.11}$$

with

$$Dg(x_0)v \neq 0, \qquad DY_h(x_0)v \neq 0, \tag{3.12}$$

then a local curve contained in the frontier level set passes through states requiring opposite terminal actions, while the deeper observation changes to first order. The frontier information is therefore locally decision-aliased along v, whereas the depth-h probe is locally sensitive to that direction.

Definition 3.4 (Revelation depth)**.** For a decision-relevant direction v at x_0, define its first revelation depth

$$\tau^\star(v) = \inf\{h : DY_h(x_0)v \neq 0\}. \tag{3.13}$$

For a discrete probe grid, the infimum is understood over the admissible depths. The quantity records first sensitivity only; it does not by itself assume that sensitivity must persist at every deeper exact-depth readout.

Corollary 3.5 (Equal-budget short-probe separation)**.** *Suppose the conditions of Proposition 3.3 hold and a restart frontier reaches depth c while a productive active method reaches $h = \frac{m}{m-1}c$ at equal update budget. If*

$$DY_c(x_0)v = 0, \qquad DY_h(x_0)v \neq 0, \tag{3.14}$$

then the short-probe information is locally blind to a decision-changing direction that the deeper active observation reveals. On a discrete admissible depth grid—or more generally when the first sensitive depth is attained—persistence over the admissible depth family (for example because the legal depth-h record retains earlier readouts) makes this condition equivalently summarized by $c < \tau^\star(v) \leq h$. If, in addition, this aliasing occurs on a positive-probability coarse-information region and the refined conditional action gap places nonzero mass on both signs there, Theorem 3.2 gives strict Bayes refinement value.

For the prespecified two-action contract, the persistent/cumulative-readout shorthand for the regime of interest is

$$\boxed{4 < \tau^\star(v) \leq 8,} \tag{3.15}$$

while the exact nonpersistent condition is $DY_4(x_0)v = 0$ and $DY_8(x_0)v \neq 0$. This is a *conditional theory-level separation* from an H4 DynaMiCS-style short-probe information class. It does not assert that every empirical H4 representation satisfies the antecedent; the Transformer study instead tests whether the fixed H8 observation produces reproducible decision-relevant refinement in the declared frontier-ambiguity regime.

4 Productive revelation technology

Having identified what a useful refinement must reveal, we next ask how that information is acquired and what state the experiment leaves behind. A trial protocol u of depth h produces an observation $Y_{u,h}$ and the refined information

$$\mathcal{G}_{u,h} = \mathcal{H} \vee \sigma(Y_{u,h}). \tag{4.1}$$

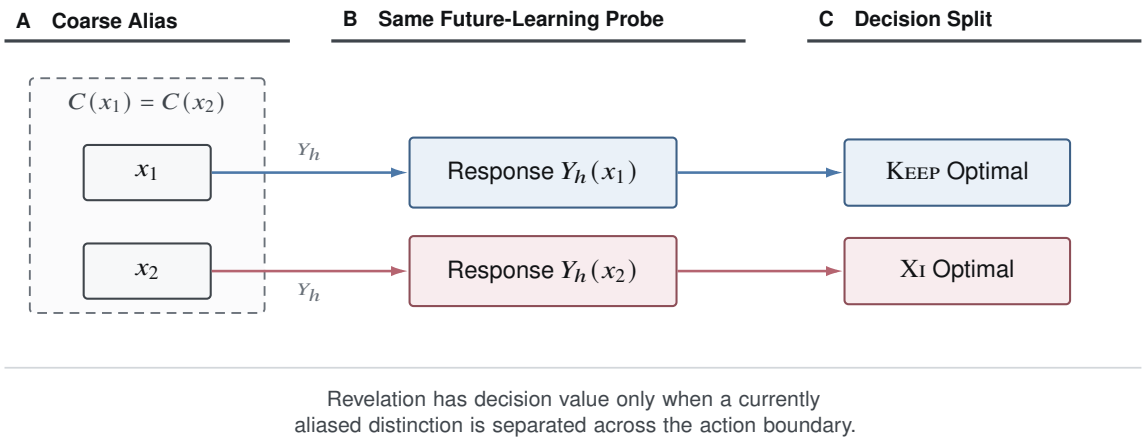

Figure 1. Decision-relative revelation. A coarse information cell can contain learning states that require different terminal actions. Deeper future-learning responses are useful only when they separate that decision quotient; full hidden-state reconstruction is unnecessary.

We distinguish two execution technologies.

Restart / temporary probing. The tested path is discarded. After a decision is made, the selected action starts freshly from the original anchor.

Promotion / productive revelation. The selected tested path is retained and continued from its trial endpoint. The probe is therefore simultaneously an information acquisition and a partially executed candidate action.

The distinction matters because pure information refinement and technology expansion are different objects. Promotion is not "free compute": it is credit for work that remains valid on the selected deployment path. Conversely, when a probe is unsafe, destructive, or otherwise unusable for deployment, restart is the correct technology and the identity below does not apply.

4.1 Exact equal-budget identity

Suppose there are $m \geq 2$ candidate actions and a common terminal horizon H, with admissible probe depths $c, h \in [0, H]$. A temporary/restart method that probes every action to depth c and then executes the selected action freshly to H spends

$$K_F(c) = mc + H. \tag{4.2}$$

A probe-and-promote method that probes every action to depth h and then continues the selected tested path spends

$$K_{\mathrm{AR}}(h) = mh + (H - h) = H + (m - 1)h. \tag{4.3}$$

Proposition 4.1 (Productive revelation equal-budget identity)**.** *If $K_F(c) = K_{\mathrm{AR}}(h)$, then*

$$\boxed{h = \frac{m}{m-1}c.} \tag{4.4}$$

For two actions, $h = 2c$.

Because $h \leq H$, an equal-budget productive match within the declared horizon is feasible only when $c \leq (m-1)H/m$; this condition is satisfied by the H4/H8/H12 comparison below.

For the prespecified Transformer comparison, $m = 2$, $H = 12$, and both methods receive 20 policy-visible updates:

$$K_F(4) = 2(4) + 12 = 20, \qquad K_{\mathrm{AR}}(8) = 12 + 8 = 20. \tag{4.5}$$

Thus an H4 temporary probe and an H8 productive probe are compute-matched under this contract.

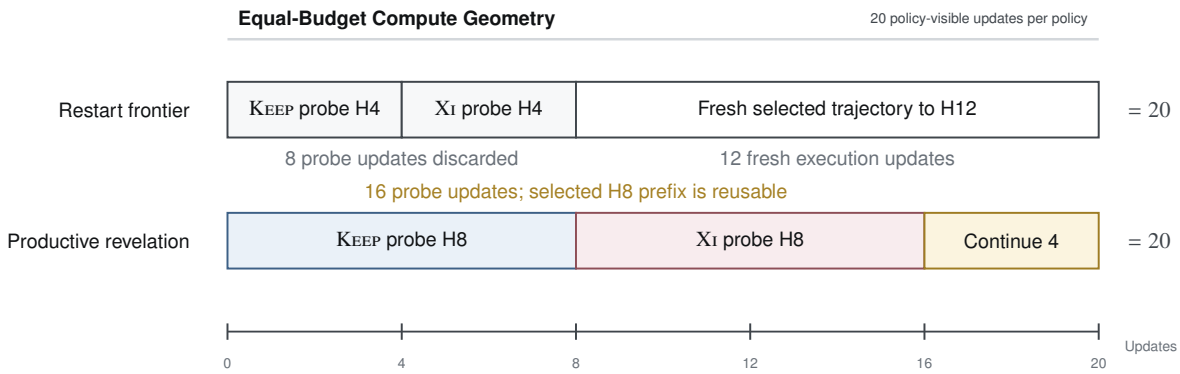

Figure 2. Equal-budget compute geometry for the two-action H12 comparison. Restart probing discards both H4 trials and executes the selected action freshly; productive revelation probes both actions to H8 and retains the selected tested prefix. Both consume 20 policy-visible updates, but promotion buys twice the revelation depth in the two-action case.

4.2 End-to-end value decomposition

For any fixed active policy π_A and frontier policy π_F, let V_R denote common-restart evaluation and V_P productive evaluation of the active policy. Adding and subtracting $V_R(\pi_A)$ gives the exact identity

$$\boxed{V_P(\pi_A) - V_R(\pi_F) = \underbrace{V_R(\pi_A) - V_R(\pi_F)}_{\Delta_{\text{restart}}} + \underbrace{V_P(\pi_A) - V_R(\pi_A)}_{\Delta_{\text{path}}}.} \tag{4.6}$$

The first term measures policy quality on common restart outcomes; the second measures productive reuse of the tested path. This identity is especially useful empirically because the two components can be estimated on independent panels without mixing absolute outcomes from unrelated future banks.

5 Adaptive revelation control: the dynamic closure

The preceding sections treat $\mathscr{H} \subseteq \mathscr{G}$ as a fixed information refinement. Once future-learning depth can be chosen, the information state itself becomes a control variable. This section shows that the dynamic problem is not a separate objective: it is the conditional version of the same Bayes refinement value in Equation (2.2). The generic "continue if expected decision improvement exceeds cost" principle is classical metareasoning [13, 36]; the learning-state specialization here supplies the revelation filtration, productive cost geometry, and revealability state.

Let $\{\mathscr{F}_h\}$ be the legal revelation filtration, nested in probe depth h, and define the Bayes commit value

$$C_h = \max_{a \in \mathcal{A}} \mathbb{E}[Q_a \mid \mathscr{F}_h]. \tag{5.1}$$

For $h' > h$, define the oracle conditional value of purchasing the deeper refinement

$$\mathcal{G}_{h \to h'} = \mathbb{E}[C_{h'} \mid \mathscr{F}_h] - C_h. \tag{5.2}$$

Proposition 5.1 (Static-to-dynamic embedding)**.** *For every nested pair $\mathscr{F}_h \subseteq \mathscr{F}_{h'}$,*

$$\boxed{\mathbb{E}[\mathcal{G}_{h \to h'}] = V_B(\mathscr{F}_{h'}) - V_B(\mathscr{F}_h) = \mathcal{I}(\mathscr{F}_{h'} \mid \mathscr{F}_h).} \tag{5.3}$$

Thus the static Bayes refinement value is the population average of the state-dependent continuation value used by adaptive revelation.

For fixed shallow and deep decision policies π_h and $\pi_{h'}$, define the legal-information policy-level gain

$$G_{h \to h'} = \mathbb{E}\left[Q_{\pi_{h'}} - Q_{\pi_h} \,\middle|\, \mathscr{F}_h\right]. \tag{5.4}$$

When the policies are Bayes optimal for their legal information, this coincides with Equation (5.2). With m candidate actions and productive promotion, extending every candidate from h to h' costs only

$$\Delta K(h, h') = (m - 1)(h' - h) \tag{5.5}$$

additional policy-visible updates. At update price λ, write $c_{h\to h'} = \lambda(m - 1)(h' - h)$.

Proposition 5.2 (Conditional revelation stopping)**.** *Relative to stopping at depth h, the pointwise optimal one-step continuation decision for a fixed policy pair is*

$$\boxed{\text{continue to } h' \iff G_{h\to h'} > c_{h\to h'}.} \tag{5.6}$$

For Bayes policies, the same rule uses $\mathcal{G}_{h\to h'}$. A population ambiguity quantile is therefore not, in general, a cost-optimal stopping boundary.

Theorem 5.3 (Cost-adjusted scalar-control factorization)**.** *Let S_h be any declared scalar summary measurable with respect to $\mathscr{F}_h$, and let*

$$\Gamma_{h\to h'} := G_{h\to h'} - c_{h\to h'} \tag{5.7}$$

be the conditional net gain from continuing rather than stopping. Relative to always stopping, define the one-step meta-control values

$$V_{\text{meta}}(\mathscr{F}_h) = \mathbb{E}[(\Gamma_{h\to h'})_+], \tag{5.8}$$

$$V_{\text{meta}}(S_h) = \mathbb{E}[(\mathbb{E}[\Gamma_{h\to h'} \mid S_h])_+]\,. \tag{5.9}$$

If

$$a(S_h) = \mathbb{E}[(\Gamma_{h\to h'})_+ \mid S_h], \qquad b(S_h) = \mathbb{E}[(-\Gamma_{h\to h'})_+ \mid S_h], \tag{5.10}$$

then the exact value of retaining the full legal shallow information rather than only S_h is

$$\boxed{V_{\text{meta}}(\mathscr{F}_h) - V_{\text{meta}}(S_h) = \mathbb{E}[\min\{a(S_h), b(S_h)\}].} \tag{5.11}$$

Consequently,

$$V_{\text{meta}}(\mathscr{F}_h) = V_{\text{meta}}(S_h) \tag{5.12}$$

if and only if, conditional on S_h, the legal shallow states do not place positive probability on both $\Gamma_{h\to h'} > 0$ and $\Gamma_{h\to h'} < 0$ on any set of positive probability. Equivalently, there exists an optimal Stop/Continue *meta-action that factorizes through S_h; ties at zero are value-neutral.*

Theorem 5.3 sharpens the distinction between predictive and decision nonredundancy. A second shallow coordinate may change the numerical continuation value while remaining irrelevant to the optimal meta-action if every state within a scalar fiber stays on the same side of the priced stopping boundary. What makes an additional revealability coordinate *decision-nonredundant* is not conditional variation by itself, but cost-adjusted boundary crossing within a scalar fiber.

Proposition 5.4 (Finite-depth Bellman closure)**.** *Let $h_0 < \cdots < h_J$ be admissible revelation depths and let c_j be the priced marginal cost of moving from h_j to h_{j+1}. If deeper revelation can be purchased sequentially, then*

$$V_J^{\text{dyn}} = C_{h_J}, \tag{5.13}$$

$$V_j^{\text{dyn}} = \max\left\{C_{h_j},\ \mathbb{E}[V_{j+1}^{\text{dyn}} \mid \mathscr{F}_{h_j}] - c_j\right\}. \tag{5.14}$$

Thus static information refinement becomes an optimal-stopping control problem over the revelation filtration. In a general fork–probe–promote system the state must also retain promoted-state information needed to make the transition Markov.

5.1 Margin and revealability: value variation versus decision nonredundancy

For binary actions, let

$$D = Q_{\text{Xi}} - Q_{\text{Keep}}, \qquad m_h = \mathbb{E}[D \mid \mathscr{F}_h]. \tag{5.15}$$

Suppose the refined Bayes margin admits the conditional location–scale representation

$$m_{h'} = m_h + \sigma_h Z, \tag{5.16}$$

where Z is independent of $\mathscr{F}_h$, symmetric about zero, integrable, and nondegenerate, while $\sigma_h \geq 0$ is $\mathscr{F}_h$-measurable. Define

$$\psi(t) = \mathbb{E}[(Z - t)_+], \qquad t \geq 0, \tag{5.17}$$

and define the conditional Bayes value of deeper revelation by

$$\mathcal{R}_{h\to h'} := \mathbb{E}[(m_{h'})_+ \mid \mathscr{F}_h] - (m_h)_+. \tag{5.18}$$

Theorem 5.5 (Two-coordinate revelation geometry)**.** *Under Equation* (5.16)*, the conditional Bayes value of deeper revelation is*

$$\boxed{\mathcal{R}_{h\to h'} = \sigma_h \psi\left(\frac{|m_h|}{\sigma_h}\right),} \tag{5.19}$$

with the convention $\mathcal{R}_{h\to h'} = 0$ *when* $\sigma_h = 0$*. Wherever differentiable,*

$$\frac{\partial \mathcal{R}}{\partial |m_h|} = -\Pr\left(Z > \frac{|m_h|}{\sigma_h}\right) \leq 0, \tag{5.20}$$

and

$$\frac{\partial \mathcal{R}}{\partial \sigma_h} = \psi(t) + t \Pr(Z > t) \geq 0, \qquad t = \frac{|m_h|}{\sigma_h}. \tag{5.21}$$

Hence, in the nonredundant regime, the cost-optimal continuation boundary is a curve in

$$\boxed{(|m_h|, \sigma_h),} \tag{5.22}$$

rather than a universal threshold on current decision margin. The derivative in σ_h *is strict whenever the refinement retains positive probability mass beyond the current normalized margin. The theorem itself does not require* σ_h *to remain empirically nonredundant after conditioning on* $|m_h|$*.*

Corollary 5.6 (Conditional value nonredundancy and scalar reduction)**.** *Let* $M_h = |m_h|$ *and write*

$$r(m, s) = s\,\psi(m/s), \tag{5.23}$$

with the same zero-scale convention as in Theorem 5.5.

(i) Scalar-redundant regime. *If* $\sigma_h = g(M_h)$ *almost surely on a decision-relevant event* E *for some measurable* g*, then*

$$\mathcal{R}_{h\to h'} = \widetilde{r}(M_h) \qquad \text{on } E, \tag{5.24}$$

for the scalar function $\widetilde{r}(m) = r(m, g(m))$*. Hence current margin is sufficient for the one-step oracle continuation value on* E*. A single monotone margin threshold requires the additional condition that* $\widetilde{r}$ *be monotone relative to the fixed compute price.*

(ii) Nonredundant regime. *If the conditional law of* σ_h *given* M_h *is nondegenerate on a set of positive probability and, on the corresponding conditional support,* $s \mapsto r(M_h, s)$ *is strictly increasing, then the conditional law of* $\mathcal{R}_{h\to h'}$ *given* M_h *is also nondegenerate on a set of positive probability. Consequently no measurable function of current margin alone can reproduce the oracle continuation value there.*

Corollary 5.7 (Cost-adjusted revealability crossing criterion)**.** *Assume the setting of Theorem 5.5, fix a continuation price* $c > 0$*, and suppose that for almost every* m *in the decision-relevant region the map* $s \mapsto r(m, s)$ *is continuous and strictly increasing on the conditional support of* $\sigma_h \mid M_h = m$*. Define the generalized critical revealability*

$$\sigma_c(m) = \inf\{s \geq 0 : r(m, s) > c\}, \tag{5.25}$$

with $\inf \varnothing = +\infty$*.*

(i) The scalar-margin decision is sufficient if and only if, for almost every M_h, either

$$\Pr(r(M_h, \sigma_h) \le c \mid M_h) = 1 \quad \textit{or} \quad \Pr(r(M_h, \sigma_h) \ge c \mid M_h) = 1. \tag{5.26}$$

Equivalently, conditional on almost every margin value, the priced continuation gain has no mass on both strict sides of zero. Under the stated continuity and monotonicity conditions, a transparent sufficient support condition is that $\sigma_h \mid M_h = m$ is contained in $[0, \sigma_c(m)]$ or in $[\sigma_c(m), \infty)$; zero-gain ties at the boundary are value-neutral and may be assigned to either optimal meta-action.

(ii) If there is a set of margin values with positive probability on which

$$\Pr(r(M_h, \sigma_h) < c \mid M_h) > 0 \quad \textit{and} \quad \Pr(r(M_h, \sigma_h) > c \mid M_h) > 0, \tag{5.27}$$

then the scalar-margin controller is strictly suboptimal:

$$V_{\text{meta}}(\mathscr{F}_h) > V_{\text{meta}}(M_h). \tag{5.28}$$

Under the same monotonicity conditions, Equation (5.27) *is implied by positive conditional mass strictly below and strictly above $\sigma_c(M_h)$. Thus conditional spread in revealability is* decision-nonredundant *exactly when a positive-mass scalar fiber contains both the strict* Stop *side and the strict* Continue *side of the priced continuation decision; zero-gain ties at the boundary do not affect value.*

Proposition 5.8 (Location–scale identification of revealability scale)**.** *Under Equation* (5.16)*,*

$$\mathbb{E}[|m_{h'} - m_h| \mid \mathscr{F}_h] = \sigma_h \, \mathbb{E}|Z|. \tag{5.29}$$

If additionally $\mathbb{E}[Z^2] = 1$, then

$$\mathbb{E}\big[(m_{h'} - m_h)^2 \mid \mathscr{F}_h\big] = \sigma_h^2. \tag{5.30}$$

At the population level, these identities identify revealability scale up to a fixed normalization from the conditional spread of the true deeper-margin increment. In applications the Bayes margins themselves may be latent or estimated; the identities then motivate development-only proxy construction rather than claiming direct identification from noisy fitted scores.

Definition 5.9 (Legal model-specific revealability proxy)**.** For a model family $\mathcal{M}$, let $X_h^{(\mathcal{M})}$ denote a declared shallow information map measurable with respect to $\mathscr{F}_h$. A statistic

$$R_h^{(\mathcal{M})} = r_{\mathcal{M}}\Big(X_h^{(\mathcal{M})}\Big) \tag{5.31}$$

is a *legal model-specific revealability proxy* when its functional form or fitted parameters are determined without using the target outcomes and are fixed before target evaluation. The proxy may estimate σ_h directly, rank the conditional spread in Proposition 5.8, or enter a predeclared estimator of $G_{h\to h'}$. No equality $R_h^{(\mathcal{M})} = \sigma_h$ is assumed, and different model families need not share the same proxy formula, scale, coefficient, or stopping threshold.

Protocol 5.10 (Model-specific Revelation-Control instantiation)**.** For a new learning-system family $\mathcal{M}$, an empirical instantiation should proceed as follows.

P1. Declare the decision problem. Fix the action menu, terminal utility, legal shallow filtration, admissible revelation depths, productive-promotion technology, and resource-price ledger before target outcomes are inspected.

P2. Separate development from target evaluation. Collect a development panel with the shallow and deeper potential outcomes needed to estimate continuation value. Keep the target panel disjoint at the exact-anchor level.

P3. Instantiate the observable continuation state. Fit either $G_{h\to h'}$ directly from legal shallow information or a legal proxy $R_h^{(\mathcal{M})}$ motivated by Proposition 5.8. Always retain a declared scalar comparator.

P4. Control representation selection. Choose a minimal learnable shallow summary using development data only. If a finite set of architectures remains eligible, predeclare that family and use simultaneous or familywise-valid inference on the target panel rather than selecting an unadjusted winner after target evaluation.

P5. Fix the controller before evaluation. Fix the fitted estimator, stop/continue threshold, tie and fallback rules,

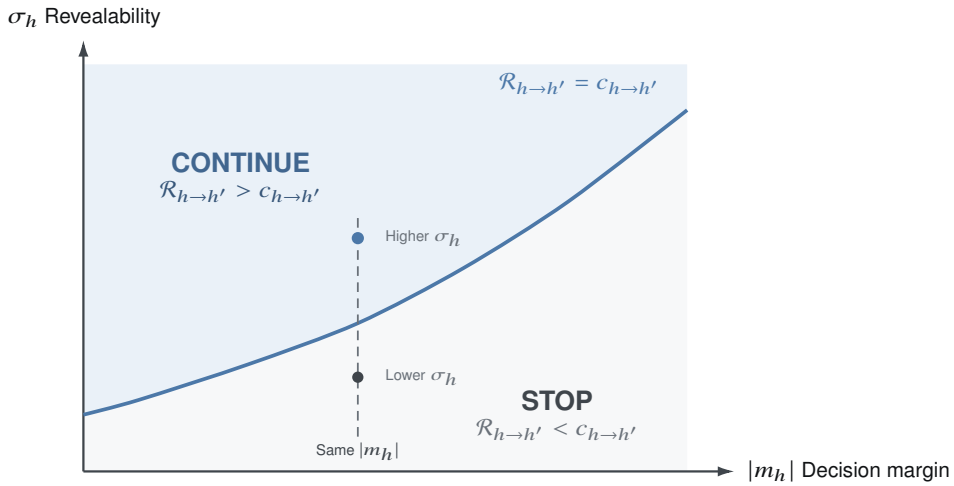

Figure 3. Schematic margin–revealability geometry implied by Theorem 5.5. At fixed compute price, larger current decision margins require greater state-dependent revealability to justify deeper future learning. The paired points illustrate the nonredundant case in which two states with the same current margin can fall on opposite sides of the stopping boundary because their revealability differs. The curve $\sigma_h = g(|m_h|)$ illustrates the stronger value-redundant special case in which continuation value itself reduces to a scalar function. More generally, scalar *decision* sufficiency requires only that the accessible states at a fixed margin remain on one side of the priced boundary; residual revealability variation is allowed. The boundary is theoretical and schematic, not an empirical fit.

productive compute accounting, risk target, resampling unit, and inferential criteria before evaluating target outcomes.

P6. Evaluate once and separate claims. Evaluate the fixed family on the disjoint target panel. Report decision value, compute-accounted utility, and decision-instability risk as distinct objects; a bounded-risk claim becomes an expected-utility certificate only under an explicit severity or tail bridge.

The protocol transports the theoretical structure, not the numerical proxy, coefficient, or threshold of an earlier model family.

Cross-model transport is structural, not parametric. The theorem-level transport target is the Revelation-Control relation between legal shallow information, continuation value, productive cost, and the stopping decision. A second model family need not reproduce the same empirical proxy used in the first. To claim a *decision-nonredundant revealability regime*, a legal proxy specified independently of target outcomes must add held-out control value beyond a scalar-margin comparator, or otherwise establish cost-adjusted boundary crossing as in Corollary 5.7. If the model fit on development data and fixed before target evaluation instead admits a successful scalar continuation rule and no legal second coordinate shows incremental control value, that outcome is consistent with scalar decision sufficiency under Theorem 5.3, without requiring the stronger assertion that revealability is functionally determined by margin. In either regime, target evaluation must use the same declared terminal utility and compute ledger, and the proxy must not be selected from target outcomes.

The theorem explains why current confidence and value of further revelation need not induce the same ordering in the nonredundant regime. A state can look locally decisive yet remain highly revealable under future training; conversely, a state near the current decision boundary can have little continuation value if its future response is stable. Corollary 5.6 characterizes value-level redundancy, while Theorem 5.3 and Corollary 5.7 give the sharper decision-level statement: scalar utility-aware control can remain optimal even with residual revealability variation, provided no positive-mass scalar fiber crosses the priced STOP/CONTINUE boundary.

6 From oracle revelation to learnable and certifiable control

A refinement can improve oracle Bayes value and still be a poor learned representation at finite sample size. To make this distinction explicit, fix a candidate acquisition scheme s with refined information $\mathcal{G}_s$ and promotion technology P_s. Let

$$\Delta_s^{\mathrm{rev}} = V_B(\mathcal{G}_s) - V_B(\mathcal{H}), \tag{6.1}$$

$$\Delta_s^{\mathrm{prom}} = V_{P_s}^{\star} - V_B(\mathcal{G}_s), \tag{6.2}$$

where $V_{P_s}^{\star}$ is the best value available under the refined information and the declared promotion technology. When restart remains an admissible fallback under that technology, $\Delta_s^{\mathrm{prom}} \geq 0$; otherwise the ledger remains algebraically valid with a signed technology term. For a learner class Π_s, let $R_s^{\mathrm{app}} \geq 0$ be the gap between $V_{P_s}^{\star}$ and the best policy in Π_s, let $R_{s,n}^{\mathrm{est}} \geq 0$ be the finite-sample gap between that class optimum and the learned policy, and let $C_s \geq 0$ be acquisition cost in the chosen utility units.

Proposition 6.1 (Finite-sample revelation ledger)**.** *With the definitions above, the learned net value satisfies the exact identity*

$$\boxed{W_{s,n}^{\mathrm{net}} - V_B(\mathcal{H}) = \Delta_s^{\mathrm{rev}} + \Delta_s^{\mathrm{prom}} - R_s^{\mathrm{app}} - R_{s,n}^{\mathrm{est}} - C_s.} \tag{6.3}$$

The ledger separates an oracle question from an extraction question. More telemetry can increase Δ_s^{rev} while increasing effective estimation complexity enough to worsen $R_{s,n}^{\mathrm{est}}$. Development analyses with substantially broader telemetry exhibited this estimation obstruction, motivating a compact refinement rather than maximizing raw feature count.

6.1 Minimal learnable refinement

Define the expected finite-sample value

$$\Psi_n(s) = \Delta_s^{\mathrm{rev}} + \Delta_s^{\mathrm{prom}} - R_s^{\mathrm{app}} - \mathbb{E}[R_{s,n}^{\mathrm{est}}] - C_s. \tag{6.4}$$

Let $\kappa_n(s)$ be a learner-relative complexity measure and, for tolerance $\tau \geq 0$, let

$$\mathcal{S}_{n,\tau} = \left\{ s : \Psi_n(s) \geq \sup_{s'} \Psi_n(s') - \tau \right\}. \tag{6.5}$$

When $\mathcal{S}_{n,\tau}$ is nonempty and the minimum of κ_n is attained on it, a *minimal learnable refinement* is any

$$s_{n,\tau}^{\star} \in \arg\min_{s \in \mathcal{S}_{n,\tau}} \kappa_n(s). \tag{6.6}$$

Minimality is therefore relative to the learner, sample size, acquisition technology, and tolerance. It need not mean the fewest raw features or the mathematically coarsest sufficient sigma-field.

6.2 Learned frontier-victory ledger

Let a fixed frontier learner $\hat{\pi}_F$ use information $\mathcal{H}$ under restart evaluation, and let a fixed active learner $\hat{\pi}_A$ use $\mathcal{G} \supseteq \mathcal{H}$. Define

$$R_F = V_B(\mathcal{H}) - V_R(\hat{\pi}_F), \qquad R_A = V_B(\mathcal{G}) - V_R(\hat{\pi}_A), \tag{6.7}$$

where V_R is common restart-evaluated value, and let $P_A = V_P(\hat{\pi}_A) - V_R(\hat{\pi}_A)$ be the productive promotion value of the active policy.

Corollary 6.2 (Learned frontier victory)**.** *Writing* $\Delta_{\mathrm{rev}} = V_B(\mathcal{G}) - V_B(\mathcal{H})$*,*

$$\boxed{V_P(\hat{\pi}_A) - V_R(\hat{\pi}_F) = \Delta_{\mathrm{rev}} + P_A + R_F - R_A.} \tag{6.8}$$

If resources K_A, K_F are priced by λ, then net value satisfies

$$N_A - N_F = \Delta_{\mathrm{rev}} + P_A + R_F - R_A + \lambda^{\top}(K_F - K_A). \tag{6.9}$$

At equal resource cost, a sufficient condition for active victory is $\Delta_{\mathrm{rev}} + P_A > R_A - R_F$.

Equation (6.8) separates the two empirical questions studied later: whether deeper future learning reveals decision-relevant structure, and whether productive reuse converts that structure into superior end-to-end utility.

6.3 Certification boundary for adaptive depth

Finite-sample risk calibration for early stopping is established prior art: selective prediction, Learn-then-Test, Conformal Risk Control, early-time classification, and risk-controlled early-exit networks provide general mechanisms for calibrating bounded stopping risks [1, 2, 12, 19, 35, 41]. Productive-prefix reuse exposes a particularly natural structural loss here. If

$$F_{h\to h'} = \mathbf{1}\{\pi_h \neq \pi_{h'}\}, \tag{6.10}$$

then, under the declared same-path promotion technology,

$$F_{h\to h'} = 0 \quad \Longrightarrow \quad Q_{\pi_h} = Q_{\pi_{h'}}. \tag{6.11}$$

Thus shallow/deep decision instability can be calibrated with a bounded Bernoulli loss even when terminal utility gaps are heavy-tailed. Turning that probability statement into a strict expected-net-utility guarantee additionally requires control of the stopped-harm tail. Appendix B makes this boundary sharp: for any nonzero stop–flip risk, unrestricted severity makes the worst-case expected net utility unbounded below even if stop and flip probabilities are known exactly. Bounded-severity, moment, or integrable-tail assumptions provide sufficient bridges from bounded risk to expected utility.

7 Related work and novelty boundary

Bayesian decision theory and information value. The monotonic value of information under refinement is classical in statistical decision theory and comparison of experiments [4, 5, 15]. Expected information from experiments and Bayesian experimental design provide a complementary tradition in which observations are deliberately acquired under a utility criterion [6, 26]. Our Bayes-value notation is a specialization of these traditions, not a new notion of information value or experimental design. The distinctive question is which *learning-state* distinctions are worth purchasing through a future training intervention once finite-sample extraction and execution technology are included.

Terminology: revelation in decision analysis and mechanism design. The phrase *value of revelation* has prior use in influence-diagram decision analysis: Ezawa [10] defines it from values of evidence and relates it to value of control. We therefore do not claim that phrase as new; our conditional quantity in Equation (5.2) is a continuation value for an intervention-generated refinement that can also change the controlled system and create reusable progress. The word "revelation" also has a different established meaning in mechanism design, where the revelation principle concerns truthful direct mechanisms under private information [31]. Revelation Control does not assume strategic reporting or incentive compatibility. Its object is controlled acquisition of decision-relevant state through system response.

Partial observability, active sensing, and dual control. POMDPs represent decisions under partial observability through belief states [21]; dual control emphasizes that an action can simultaneously control a system and reveal uncertainty [11]; and active-sensing work studies task-directed information acquisition [39]. Revelation Control belongs to this broad family but does not replace it. Our narrower object is an internal learning execution state, the intervention is future training, and the formal target is decision factorization through a declared legal information sigma-field. The promotion term further distinguishes information gained by a trial from useful state change left by that trial.

Predictive state and decision-focused representations. Predictive State Representations encode state through action-conditional predictions of future observations [27]. The companion work is related in spirit because future responses reveal hidden learning-state distinctions, but the present paper does not propose a general predictive state representation. It asks only for the quotient needed to choose among a fixed action menu. This has conceptual kinship with state abstraction, where representations are judged by whether they retain distinctions needed for planning or learning [24]. Recent work on decision-relevant concept selection makes this connection explicit by requiring states that share a selected concept representation to preserve the optimal decision structure [34]. Decision-focused learning and predict-then-optimize methods likewise emphasize that predictive quality should be judged by downstream decision loss [8, 9, 29]; our finite-sample ledger adds an intervention/acquisition dimension in which the representation itself must be actively produced.

Function-equivalent states and self-intervention. Recent work on path-conditioned training makes especially clear that ReLU parameterizations implementing the same function can nevertheless induce different subsequent training dynamics [23]. This is closely aligned with the motivating non-sufficiency phenomenon, but it studies rescaling and conditioning rather than decision-relative future-learning experiments. Self-Interventional Learning perturbs a network's own functional organization, learns a predictive self-model from intervention consequences, and uses that model for later structural action [38]. Revelation Control instead treats the learner as the controlled object of an external decision problem, asks which future-training responses refine a declared action quotient, and accounts separately for information and reusable tested computation.

Partial training, early stopping, and resource allocation. Freeze–Thaw Bayesian optimization uses partial learning curves to pause and resume candidate models [37]; learning-curve extrapolation uses early trajectory segments to predict eventual performance and stop unpromising runs [7]; and non-stochastic best-arm allocation formalizes the use of intermediate learning performance to concentrate resources on promising configurations [18]. Hyperband extends this resource-allocation perspective at scale [25], while Population Based Training jointly adapts model populations and hyperparameters while reusing live trajectories [17]. These works establish strong precedents for partial training, continuation, early termination, promotion, and resource allocation. Our novelty therefore does not rest on any of those operations individually.

DynaMiCS and local prospective probes. Within the declared fine-tuning-probe problem class, DynaMiCS is the closest direct comparator because it explicitly performs short domain-specific fine-tuning probes to estimate local cross-domain slopes before optimizing a constrained data mixture [14]. We therefore concede short prospective training probes and local finite-difference summaries as prior art. Our theory asks a different question: can a local-probe information state fail to factorize the terminal decision, and can productive reuse make a deeper decision-relevant observation available under the same update budget? The prespecified H4 comparator is a strengthened task adaptation of that acquisition geometry, not a claim that the original DynaMiCS algorithm is globally dominated.

Metareasoning and value of computation. Rational metareasoning asks whether an additional computation is worth its cost because of the external decision it may change [36]; knowledge-gradient policies similarly choose measurements by expected increment in terminal value [13]. Recent adaptive-reasoning work operationalizes marginal-gain-versus-cost allocation through difficulty signals [43], while consequence-aware allocation shows that difficulty and error severity need not coincide when assigning test-time reasoning budgets [42]. Proposition 5.2 is a Revelation-Control specialization of value-of-computation reasoning, not a new generic stopping rule. The structural point specific to this paper is that Equation (5.3) makes the original hidden-learning-state refinement value the state-dependent reward of that control problem. The distinctive object is hidden learner execution state: the computation is controlled future training, and the selected experimental path can remain useful training.

Risk-controlled stopping. Reject-option and selective-classification methods trade prediction coverage against conditional risk [12]. Learn-then-Test and Conformal Risk Control provide finite-sample calibration of bounded risks for families of predictive rules [1, 2]. These tools have also been applied directly to sequential stopping: early-time classification can be calibrated for accuracy-gap control [35], risk-controlled neural early exits can use supervised or consistency losses [19], and recent work controls reasoning risk under a compute budget [41]. We therefore do not claim to introduce risk-controlled early stopping or early-versus-deep consistency. Our narrower contribution is to connect these ideas to productive future-training probes of hidden learning state, where deeper computation changes the learner and reveals a decision quotient.

Value of computation and finite-sample extraction. Our acquisition-cost and learner-regret terms also interact with finite-sample extraction: richer future telemetry can have higher oracle information value and lower learned value because estimation error grows. The minimal-learnable-refinement principle therefore treats representation complexity as part of the decision problem rather than assuming that more legal telemetry is automatically better.

Table 1. Novelty boundary relative to representative neighboring methods. Entries describe the central mechanism emphasized by each work, not every implementation variant.

Method / theory	Partial training	Prospective training signal	Path reuse	Hidden learning-state decision aliasing	Explicit finite-sample value ledger
Freeze–Thaw [37]	yes	learning curves	pause/resume	no	no
Hyperband [25]	yes	intermediate performance	continuation of survivors	no	no
PBT [17]	yes	population performance	yes	no	no
DynaMiCS [14]	yes	local cross-domain slopes	no	no	no
Risk-controlled early exit [19, 35]	no	shallow/deep prediction signal	no	no	risk control, not utility ledger
Revelation Control (this work)	yes	future-learning response	promotion separated from information value	central target	yes

Novelty claim. The paper therefore does *not* claim to invent future probing, promotion, Bayes value, value of computation, selective early exit, finite-sample risk calibration, or dynamic control. The strongest claimed contribution is the coupling

$$\boxed{\begin{array}{c}\text{hidden learning-state aliasing + productive future-training experiment}\\ \text{+ budget-indexed revelation depth + finite-sample extraction + comparative utility.}\end{array}} \tag{7.1}$$

The adaptive-control layer completes this chain by identifying current decision margin and learning-state revealability as joint control coordinates for purchasing deeper revelation, while also characterizing when revealability is conditionally redundant and scalar control suffices; the generic metareasoning and risk-calibration tools used around that result are explicitly treated as prior art.

8 Transformer instantiation

We evaluate Revelation Control in two independently instantiated decoder-only 7B Transformer families: Qwen2.5-7B [32, 33] and Mistral-7B-v0.3 [20, 30]. Both use LoRA adaptation [16] with AdamW optimizer state [22, 28], while the Mistral instantiation uses independently prepared task streams, training histories, and development data. The cross-model target is therefore structural rather than numerical.

In both families, exact anchors are generated from controlled training histories and the binary action menu is

$$\mathcal{A} = \{\textsc{Keep}, \textsc{Xi}\}, \tag{8.1}$$

where Xɪ denotes the prespecified matched optimizer-state intervention. The intervention matches the immediate adaptive field while altering hidden optimizer moments, allowing later common training to reveal a difference that is absent from the immediate update. Terminal utility is evaluated at H12. The active policy probes both candidate actions to H8, selects an action using a decision rule fit on development data and fixed before target evaluation, and promotes the selected tested path to H12. The strengthened short-probe frontier probes to H4, restores the anchor, selects an action from legal H4 information, and restarts that action to H12.

8.1 Qwen2.5-7B prespecified active policy

For Qwen2.5-7B, the final fixed-depth active policy is

$$\boxed{\begin{array}{c}\text{24-dimensional raw H8 response + Ridge}(\alpha = 10)\\ \text{+zero threshold + same-path promotion.}\end{array}} \tag{8.2}$$

Ties and nonfinite scores fall back to KEEP. The representation and decision rule were selected on development data and fixed before the two independent finite-library panels and the 432-anchor comparative panel.

Selection status. The method is prespecified, not a proof of global algorithmic optimality. Development-only comparisons evaluated H2, H4, H6, and H8; H8 was selected before target evaluation, while substantially broader telemetry increased effective estimation burden without a reliable compensating gain.

8.2 Independent Mistral-7B-v0.3 instantiation

For Mistral-7B-v0.3, the same H4/H8/H12 decision geometry, action menu, terminal utility, productive-promotion technology, and 20-update accounting are retained, while numerical heads and adaptive summaries are instantiated from Mistral-specific development data. The theory does not require Qwen's empirical revealability proxy, coefficients, or stopping threshold to transfer numerically. A separate 336-anchor development panel supplies the Mistral parameters, and an independent 480-anchor two-bank panel supplies target evaluation. The adaptive analysis is restricted to the scalar and two-coordinate continuation architectures already defined in the Qwen analysis, with Mistral parameters fit only on the development panel. The transported object is the Revelation-Control structure, not one fitted coordinate system. Statistical units, lower bounds, multiplicity control, and bank aggregation are specified in Appendix E.

9 Fixed-depth revelation beyond current-information comparators

Before turning to the independent model-family replication, we first establish that the selected H8 future-response representation carries decision value beyond a broad declared current-information library in Qwen2.5-7B. The same prespecified 24-dimensional H8 raw-response Ridge policy ($\alpha = 10$) is evaluated on two independent panels, each containing 336 exact anchors with exact position balance and an analysis fixed before target evaluation. The comparator library contains 22 legal current-information components: the two fixed actions plus five prespecified decision heads within each of four current-information feature families (position, order, anchor, and H0), as displayed in Figure 4. The two panels are analyzed independently rather than pooled.

Table 2. Independent Qwen2.5-7B validation against the 22-component current-information library. Each entry is the worst component across the declared comparator set; positive lower bounds are required simultaneously.

Quantity	Independent panel A (336)	Independent panel B (336)
Minimum point margin	1.2136812×10^{-4}	1.4878650×10^{-4}
Minimum Student-t LCB	6.9363323×10^{-5}	9.4261533×10^{-5}
Minimum bootstrap LCB	7.0252125×10^{-5}	9.5137645×10^{-5}
Minimum max-t LCB	4.4947532×10^{-5}	6.9209070×10^{-5}
Comparators satisfying criterion	22/22	22/22

Every componentwise comparison remains positive in both panels for both point margins and simultaneous max-t lower bounds; Figure 4 shows all 22 comparisons on a common scale.

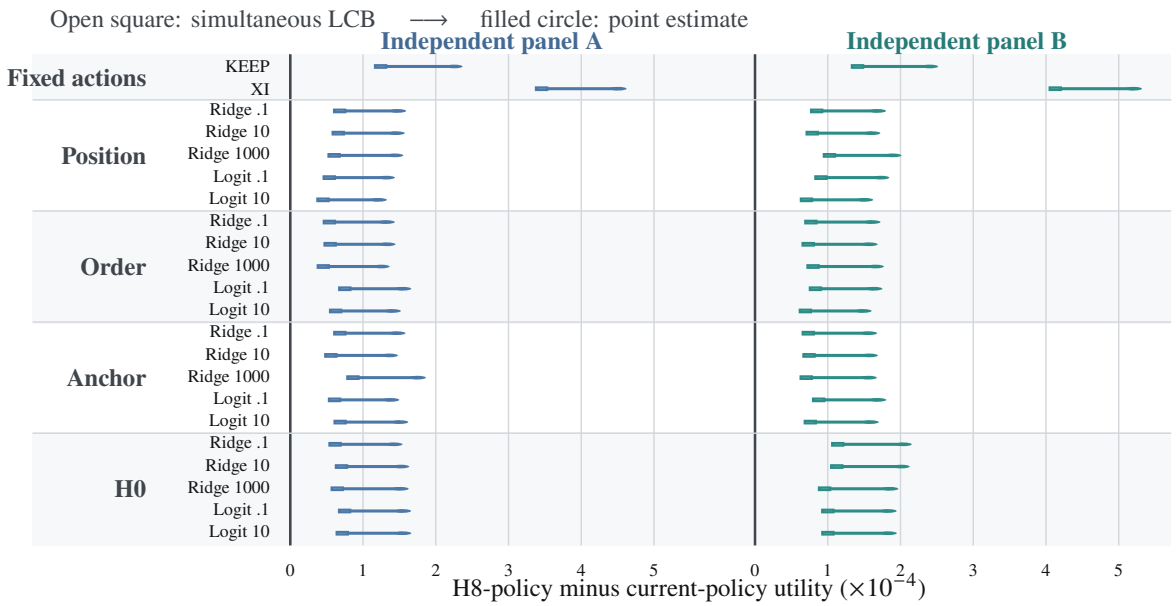

Figure 4. Two independent Qwen2.5-7B panels across all 22 current-information comparators. Open squares mark simultaneous max-t 95% lower confidence bounds and filled circles mark realized point margins. Every lower-bound endpoint remains strictly positive; the two panels are not pooled.

Productive-path value is also independently positive. Panel A gives

$$\widehat{\Delta}^{(A)}_{\text{path}} = 8.99131091346 \times 10^{-5}, \tag{9.1}$$

with one-sided Student-t and bootstrap lower bounds $4.22948353683 \times 10^{-5}$ and $4.41065688619 \times 10^{-5}$. Panel B independently gives

$$\widehat{\Delta}^{(B)}_{\text{path}} = 1.47763681856726 \times 10^{-4}, \tag{9.2}$$

with corresponding lower bounds $1.01800375097654 \times 10^{-4}$ and $1.03042544792948 \times 10^{-4}$. These path values are later combined one panel at a time with the separate 432-anchor common-restart comparison through Equation (4.6). The empirical claim in this section remains limited to the declared comparator library and the stated same-path execution technology.

10 Frontier comparison: DynaMiCS-style probing

DynaMiCS performs short domain-specific fine-tuning probes to estimate a slope matrix of local cross-domain effects and then optimizes mixture weights under performance constraints [14]. That makes it direct contemporary prior art for using future training itself as a prospective diagnostic. Our comparator is therefore not presented as the original DynaMiCS algorithm verbatim; it is a strengthened, task-adapted *DynaMiCS-style* frontier that preserves the relevant short-probe/local-slope/restart geometry while using a calibrated binary decision head tailored to the present terminal action problem.

Why strengthen the comparator? The source algorithm solves a constrained mixture-selection problem, whereas our terminal decision is binary. A literal transplant therefore requires an additional mapping from local slope information to the present action gap. If that mapping were deliberately weak, a positive frontier result could reflect head misspecification rather than a limitation of short-probe information. We instead retain the source method's short-probe acquisition structure—same anchor, temporary short probes, local finite-difference summaries, restore/restart—but give the frontier a decision head fit only on development data for the declared H12 action gap. This strengthens the competing decision rule without granting deeper or otherwise illegal information.

For Qwen2.5-7B, the legal H4 feature vector is

$$z_F = (L^0_A, L^0_B, L^0_C, S^4_{\text{KEEP},A}, S^4_{\text{KEEP},B}, S^4_{\text{KEEP},C}, S^4_{\text{XI},A}, S^4_{\text{XI},B}, S^4_{\text{XI},C}), \tag{10.1}$$

where

$$S^4_{a,d} = \frac{L_d(\theta^{H4}_a) - L_d(\theta_0)}{4}. \tag{10.2}$$

The Qwen decision head is fixed as `StandardScaler + Ridge(`$\alpha = 0.1$`)` with zero threshold and no test refit. Mistral-7B-v0.3 uses the same legal short-probe information class and restart technology with a model-specific numerical head fit on development data and fixed before target evaluation.

10.1 Equalized compute contract

Both methods receive exactly 20 policy-visible training updates:

Resource	Productive revelation	DynaMiCS-style restart
Probe updates	16	8
Fresh/promoted continuation	4	12
Total policy-visible updates	20	20
Discarded exploratory updates	8	8
Reused selected-probe updates	8	0
Terminal horizon	H12	H12
Terminal utility	identical	identical

Restore/save overhead is not charged against the comparator in the primary analysis, making the update accounting conservative for the productive-revelation method.

Theory-level comparison. Corollary 3.5 gives a conditional separation from an H4 short-probe information class: if a decision-changing direction is invisible through H4 yet revealed by H8, then the deeper refinement can have strictly larger Bayes value at the same update budget because productive work is reusable. The empirical sections below test the corresponding structural refinement and end-to-end utility consequences in both model families.

11 Structural decision refinement in the H4 frontier-ambiguity regime

The 432-anchor comparative panel is disjoint from development and from the two finite-library replication panels. Before terminal outcomes, the H4 frontier score defined the fixed ambiguity event

$$A = \{|f_F| \le \tau_F\}, \qquad \tau_F = 0.00052099142651661841. \tag{11.1}$$

The event contains 226/432 anchors, or 52.315% of the panel. Let

$$D_r = U_{T_r}(\text{XI}) - U_{T_r}(\text{KEEP}), \qquad r \in \{1, 2\}, \tag{11.2}$$

and let I_{AR} be the fixed H8 active decision. The two terminal banks are independent conditional replicates and expose both potential actions for every anchor.

11.1 Independent-bank refinement estimator

To test whether the H8 observation refines the coarse ambiguity cell in a decision-relevant way without reusing terminal outcomes for both choice and evaluation, bank T_1 chooses the best coarse constant action inside A,

$$\widehat{c}_1 = \mathbf{1}\{\overline{D}_{1,A} > 0\}, \tag{11.3}$$

and the independent bank T_2 evaluates the fixed H8 decisions relative to that coarse action:

$$\widehat{\Delta}_{1\to 2}(A) = \frac{1}{|A|}\sum_{i\in A} D_{2i}(I_{\mathrm{AR},i} - \widehat{c}_1). \tag{11.4}$$

The reverse direction $\widehat{\Delta}_{2\to 1}(A)$ is defined symmetrically, and the primary symmetric estimator is their average. In each exact-anchor bootstrap draw, the coarse action is reselected inside the resample before evaluation.

The two directions agree closely:

$$\widehat{\Delta}_{1\to 2}(A) = 1.92863119513 \times 10^{-4}, \qquad \mathrm{LCB}_t = 5.01319646865 \times 10^{-5}, \tag{11.5}$$

$$\widehat{\Delta}_{2\to 1}(A) = 1.93079378050 \times 10^{-4}, \qquad \mathrm{LCB}_t = 1.03136801205 \times 10^{-4}. \tag{11.6}$$

Their symmetric average is

$$\boxed{\widehat{\Delta}_{\mathrm{refine}}(A) = 1.92971248782 \times 10^{-4},} \tag{11.7}$$

with one-sided 95% Student-t lower bound $1.09966903938 \times 10^{-4}$ and 50,000-draw exact-anchor bootstrap lower bound $6.49311403490 \times 10^{-5}$. Weighting by the fixed ambiguity mass gives a full-panel contribution $1.00952551446 \times 10^{-4}$, with t lower bound $5.69892792476 \times 10^{-5}$ and bootstrap lower bound $3.39715294723 \times 10^{-5}$.

11.2 Opposite-action sign replication

The mechanism is visible independently in both terminal banks. Table 3 reports the terminal action gap conditional on the fixed H8 action within A.

Table 3. Independent-bank structural sign replication inside the fixed H4 frontier-ambiguity event. Bounds are one-sided 95% Student-t bounds.

Bank	Xɪ-selected mean	LCB	Keep-selected mean	UCB
T_1	2.9246×10^{-4}	6.5097×10^{-5}	-4.1166×10^{-4}	-2.2471×10^{-4}
T_2	3.6323×10^{-4}	9.5547×10^{-5}	-2.8389×10^{-4}	-8.2269×10^{-5}

Thus the H8 observation does not merely predict a continuous outcome more accurately: within a materially populated H4 frontier-ambiguity regime it reproducibly separates groups of states whose mean independent terminal outcomes favor opposite actions. Treating the ambiguity event as the coarse cell and the fixed H8 partition as the refinement gives a sample Bayes gap of $1.63115121441 \times 10^{-4}$ using the mean of T_1, T_2, with bootstrap lower bound $6.04584733171 \times 10^{-5}$. The same cell-refinement Bayes gap has a positive bootstrap lower bound in each terminal bank separately.

Secondary H4-versus-H8 diagnostic. A cross-bank diagnostic that preserves outcome separation compares the fixed nine-feature H4 Ridge representation with the fixed 24-feature H8 representation for terminal-gap prediction. Inside A, the H8 representation improves cross-bank mean squared error by 1.24677×10^{-7}, with positive t and bootstrap lower bounds. The corresponding learned-policy value difference has a confidence interval crossing zero, reflecting that the strengthened H4 frontier action is already a strong selector. We therefore interpret the main result as *decision-relevant H8 refinement inside the declared H4 frontier-ambiguity regime*, not as unrestricted dominance over every measurable function of complete H4 information.

11.3 Independent Mistral aliasing and deeper-revelation value

The independent Mistral target panel contains 480 exact anchors with two terminal banks. Its H4 ambiguity threshold, $\tau_F^{\mathrm{M}} = 0.00236920914414$, was determined from the separate development panel. Using Bank A's H4 score to define the ambiguity cell yields 216/480 anchors; among them, 88 favor Xɪ in both terminal banks and 58 favor Keep in both. Defining the cell with Bank B gives 210/480 anchors, with 83 stable-Xɪ and 58 stable-Keep anchors. Thus the same coarse shallow-information regime contains materially populated hidden-state subsets requiring opposite terminal actions.

More directly, the fixed H8 action has strictly positive terminal decision value relative to the fixed H4 action on the complete two-bank target panel. The exact-anchor mean is

$$\boxed{\widehat{\Delta}^{\mathrm{M}}_{H8-H4} = 2.25089539402 \times 10^{-4},} \tag{11.8}$$

with one-sided 95% Student-t lower bound $1.56391515827 \times 10^{-4}$ and 50,000-draw bootstrap lower bound $1.58226436880 \times 10^{-4}$. H4 and H8 actions differ on 100/480 exact anchors in at least one bank. Mistral therefore independently reproduces the core implication of decision-revelation depth: legal deeper future-learning information changes action choice on nontrivial mass and has strictly positive terminal decision value.

A stronger Qwen diagnostic in which the H8-selected partition itself must produce opposite-sign subgroup bounds in both directions is not promoted as a cross-model invariant. The theorem requires decision-relevant refinement, not numerical identity of every diagnostic partition.

12 Productive revelation and equal-compute frontier performance

We next ask whether additional revelation can be converted into higher end-to-end utility than the strengthened DynaMiCS-style H4 restart frontier under the common 20-update contract. Equation (4.6) gives the exact decomposition

$$\Delta_{\mathrm{end}} = \Delta^{\mathrm{AR-Frontier}}_{\mathrm{restart}} + \Delta^{\mathrm{AR}}_{\mathrm{path}}. \tag{12.1}$$

This separates what is gained by choosing better from what is gained because the selected deeper probe remains useful computation.

12.1 Qwen equal-compute closure

The 432-anchor Qwen panel estimates the common-restart selector component using the mean of two independent terminal banks:

$$\widehat{\Delta}^{\mathrm{AR-Frontier}}_{\mathrm{restart}} = -3.98358746688 \times 10^{-7}, \tag{12.2}$$

with standard deviation 4.67246×10^{-4}. The point estimate is therefore near zero; this is descriptive rather than an equivalence claim. Strict end-to-end advantage is supplied by productive reuse of the tested H8 path.

Combining the independent common-restart estimate separately with the two productive-path panels gives

$$\widehat{\Delta}^{(1)}_{\mathrm{end}} = 8.95147503879 \times 10^{-5}, \qquad \mathrm{LCB}_{0.95} = 2.92463884692 \times 10^{-5} > 0, \tag{12.3}$$

$$\widehat{\Delta}^{(2)}_{\mathrm{end}} = 1.47365323110 \times 10^{-4}, \qquad \mathrm{LCB}_{0.95} = 8.83939050151 \times 10^{-5} > 0. \tag{12.4}$$

The productive-path values themselves are 8.9913×10^{-5} and 1.4776×10^{-4}, respectively, with positive one-sided t and bootstrap lower bounds in both panels.

12.2 Independent Mistral equal-compute replication

The same decomposition is evaluated on the independent Mistral target panel under the identical 20-update visible-compute contract. The common-restart selector component is

$$\widehat{\Delta}^{\mathrm{M}}_{\mathrm{restart}} = -1.191350097 \times 10^{-5}, \tag{12.5}$$

whereas same-path reuse contributes

$$\boxed{\widehat{\Delta}^{\mathrm{M}}_{\mathrm{path}} = 3.09392744795 \times 10^{-4},} \tag{12.6}$$

with one-sided Student-t and bootstrap lower bounds $1.96017743591 \times 10^{-4}$ and $2.01665987607 \times 10^{-4}$. The resulting end-to-end equal-compute advantage is

$$\boxed{\widehat{\Delta}^{\mathrm{M}}_{\mathrm{end}} = 2.97479243825 \times 10^{-4},} \tag{12.7}$$

with one-sided Student-t lower bound $2.00162462592 \times 10^{-4}$ and bootstrap lower bound $2.04870462705 \times 10^{-4}$.

Table 4. Equal-compute productive-revelation evidence across the two Transformer families. The two Qwen rows use independent productive-path panels and are not pooled.

Model / evidence block	Productive path	End-to-end point	One-sided 95% LCB
Qwen, independent panel 1	8.9913×10^{-5}	8.9515×10^{-5}	2.9246×10^{-5}
Qwen, independent panel 2	1.4776×10^{-4}	1.4737×10^{-4}	8.8394×10^{-5}
Mistral, independent family	3.0939×10^{-4}	2.9748×10^{-4}	2.0016×10^{-4}

The same decomposition is supported in both families: the restart-only selector difference is not the source of the main gain, while productive reuse of deeper tested computation creates a strict equal-budget end-to-end advantage. This cross-family pattern is consistent with the productive-revelation mechanism formalized in Equation (4.6).

13 Adaptive Revelation Control across model families

The dynamic theory predicts a state-dependent stopping decision based on the conditional utility of deeper revelation, not a universal population ambiguity threshold. The empirical question is therefore whether legal shallow information can support compute-saving continuation decisions without changing the terminal utility definition or productive compute ledger. The theory also permits different model families to occupy different conditional revealability regimes.

13.1 Qwen2.5-7B: decision-nonredundant revealability and held-out safe compute

In Qwen2.5-7B, the observable H4 revealability proxy is

$$R_4 = \left\| \left(S^4_{\mathrm{XI},A} - S^4_{\mathrm{KEEP},A}, S^4_{\mathrm{XI},B} - S^4_{\mathrm{KEEP},B}, S^4_{\mathrm{XI},C} - S^4_{\mathrm{KEEP},C} \right) \right\|_2 . \tag{13.1}$$

We treat R_4 only as a legal model-specific proxy; no equality $R_4 = \sigma_4$ is assumed. In a supporting analysis with fitting and evaluation performed on separate collection blocks under the same safety protocol, the two-coordinate continuation rule improves paired cost-aware value over the scalar $|f_{H4}|$ rule by 3.2514×10^{-7}, with one-sided Student-t and bootstrap lower bounds 4.0357×10^{-8} and 3.6127×10^{-8}, and no harmful stops in either held-out direction. This is evidence for a decision-nonredundant regime at the declared compute price, not a prospective claim that this particular R_4 is unique or universally necessary.

A separate 480-history full-grid study on new exact learning-state anchors prospectively fixed the utility-aware continuation architecture, H4/H8 policies, update price, and 16/20-update compute ledger. With an independent second future bank on the same anchors, the H8-over-H4 terminal decision value is

$$\widehat{\mathbb{E}}[Q_{\pi_8} - Q_{\pi_4}]_{AB} = 1.229 \times 10^{-4}, \qquad \mathrm{LCB}_{t,0.95} = 8.687 \times 10^{-5} > 0, \qquad \mathrm{LCB}_{\mathrm{boot},0.95} = 8.769 \times 10^{-5} > 0. \tag{13.2}$$

The fixed two-coordinate point controller has positive two-bank net point value but its strict one-sided net-value lower bounds cross zero, illustrating the finite-sample tail sensitivity emphasized by the theory.

The bounded-risk certification layer provides a complementary held-out safe-compute result. In a theory-constrained calibration/held-out evaluation, one bank is used for calibration and the other for evaluation. The resulting calibrated rule stops on 4/480 histories, produces no H4/H8 action changes and no harmful stops, has a 95% Bernoulli–KL upper joint-risk bound of $0.6222\% < 2\%$, and saves 0.1667% of visible updates. Its realized panel-level update-priced net value is

$$\widehat{\Delta V}_{\mathrm{cert}} = 2.890 \times 10^{-7}, \qquad \mathrm{LCB}_{t,0.95} = 5.161 \times 10^{-8} > 0, \qquad \mathrm{LCB}_{\mathrm{boot},0.95} = 7.225 \times 10^{-8} > 0. \tag{13.3}$$

13.2 Mistral-7B-v0.3: scalar continuation control at the tested resolution

Mistral-7B-v0.3 does not require numerical transfer of Qwen's R_4. A Mistral-specific continuation head is instantiated using only the separate 336-anchor development panel, while the independent 480-anchor two-bank panel is used for target evaluation without target-panel refitting. Among the two pre-existing shallow continuation architectures evaluated as a finite family, the scalar margin-based estimator retains positive adaptive value at the observed resolution.

Relative to always continuing to H8, the scalar utility-aware controller achieves exact-anchor update-priced net value

$$\boxed{\widehat{\Delta V}^{\mathrm{M}}_{\mathrm{scalar}} = 1.13877630257 \times 10^{-5}.} \tag{13.4}$$

Its ordinary one-sided 95% Student-t and 50,000-draw bootstrap lower bounds are $2.37141363069 \times 10^{-6}$ and $5.41900958243 \times 10^{-6}$. Because both scalar and two-coordinate continuation architectures existed before the Mistral target evaluation, we treat them as a two-element finite architecture family. Bonferroni-adjusted familywise 95% lower bounds for the scalar rule remain positive: $6.37735915928 \times 10^{-7}$ for Student-t inference and $5.23837592968 \times 10^{-6}$ for the exact-anchor bootstrap. Across 960 bank-level realizations it stops 166 times, changes the H8 action once, and has zero harmful stops under the declared update price. The raw Qwen-style R_4 coordinate does not improve Mistral adaptive value over the scalar controller; exploratory development-only proxy diagnostics are excluded from the target-panel claim. We therefore interpret Mistral as consistent with scalar decision sufficiency under Theorem 5.3 and Corollary 5.7, rather than as a failure of the two-coordinate theorem. This empirical conclusion does not assert that σ_4 is literally a function of margin; it says only that the finite pre-existing architecture family did not require a second coordinate to improve the priced target-panel stopping decision at the tested resolution.

13.3 Risk–severity separation in Mistral

A separate risk-calibrated Mistral shallow-stop rule provides an empirical illustration of Theorem B.3. It makes 351/960 early stops while the 95% Bernoulli–KL joint stop/action-change upper bound is $1.5924\% < 2\%$ in each bank. Yet the exact-anchor update-priced net point value is slightly negative,

$$\widehat{\Delta V}^{\mathrm{M}}_{\mathrm{risk}} = -3.3428 \times 10^{-7}, \tag{13.5}$$

with negative one-sided lower bounds. Only four stopped action changes occur, but the largest absolute continuation value is 9.811×10^{-3}. Thus low action-instability probability and positive utility are empirically distinct objects: rare high-severity misses can dominate many cheap correct stops. This is precisely the risk–severity separation formalized in Appendix B.

Taken together, the two model families support the same adaptive-control principle while occupying different shallow-state regimes. Qwen exhibits evidence for a decision-nonredundant revealability coordinate; Mistral admits a successful scalar continuation rule. What transports is the conditional revelation-value decision, not a universal empirical coordinate system.

14 Cross-model structural validation

The two model families are not expected to reproduce identical coefficients, thresholds, or empirical revealability coordinates. The relevant replication object is the sequence of theory-level implications: shallow information can alias decision-relevant state; deeper revelation can have positive decision value; productive reuse can convert that information into equal-compute utility; revelation depth can be controlled state by state; and risk control is not equivalent to utility control when severity is unrestricted.

Table 5. Theory-level evidence across the two Transformer families. The fixed-depth and equal-compute Mistral results use an independent target panel; the adaptive row evaluates a pre-existing two-architecture family with familywise-adjusted inference.

Theoretical implication	Qwen2.5-7B evidence	Mistral-7B-v0.3 evidence
Decision-relevant shallow aliasing	Fixed H4 ambiguity cell contains H8-refined groups with opposite terminal action gaps in two independent banks.	Using the H4 threshold fixed from development data, the bank-A-defined ambiguity cell contains 216/480 anchors, including 88 stable-XI and 58 stable-KEEP anchors whose terminal-gap signs agree across both banks (descriptive structural check).
Positive value of deeper revelation	H8 beats the declared current-information library on two independent 336-anchor panels; the separate 480-history H8-over-H4 value is strictly positive.	Exact-anchor H8-over-H4 value 2.2509×10^{-4} with positive t and bootstrap lower bounds.
Productive information-to-utility closure	Productive path value is positive in two independent panels and yields strict 20-update end-to-end advantage over the strengthened restart frontier.	Same-path reuse 3.0939×10^{-4} and equal-compute end-to-end advantage 2.9748×10^{-4}, both with positive lower bounds.
Adaptive Revelation Control	Supporting evidence for incremental decision value from a nonredundant revealability coordinate; a held-out bounded-risk safe-compute evaluation has positive panel net-value lower bounds.	Consistent with scalar decision sufficiency at the tested resolution; a familywise structural validation of the scalar continuation architecture, fit only on the separate development panel, retains positive Bonferroni-adjusted exact-anchor t and bootstrap lower bounds within the two-element pre-existing architecture family.
Risk–severity boundary	Positive held-out safe-compute behavior is explicitly limited to panel-level evidence without an unrestricted tail guarantee.	Sub-2% stop–flip risk coexists with nonpositive risk-controller net value because a few flips have high severity.

The common pattern is therefore stronger than numerical parameter reuse. In both families, future learning supplies decision-relevant information, deeper tested paths carry productive value, and the resulting information can improve compute-accounted utility. The model-specific difference occurs one level lower: Qwen provides evidence of incremental decision value from an additional shallow revealability coordinate, whereas Mistral is consistent with scalar decision sufficiency at the observed resolution. Theorem 5.3 shows why these outcomes belong to the same theory: an extra coordinate is required for control only when states sharing the scalar summary cross the priced STOP/CONTINUE boundary. Corollary 5.7 gives the corresponding location–scale criterion, while Definition 5.9 allows model-specific empirical realizations.

Accordingly, the cross-model conclusion is not that one fitted controller or proxy transfers unchanged. It is that the *Revelation-Control relation* between legal shallow information, continuation value, productive cost, and terminal decision utility survives an independent model-family change. This is the level at which the empirical evidence supports the general theory's central testable implications.

15 Potential application domains

The empirical results in this paper concern learning systems, but the decision-theoretic structure is more general. A Revelation-Control problem can be formulated when five ingredients can be defined: a consequential terminal action; hidden state not fully resolved by current legal information; a priced intervention that can be executed before commitment; an observable response that can refine the relevant decision quotient; and a cost ledger that distinguishes information acquisition from any useful state change left by the intervention. The strongest form of the framework arises when the experiment is *productive*: if the tested path is selected, some of the work performed to reveal state is itself part of the useful continuation. These conditions are substantially narrower than "any sequential decision problem,"

and the examples below are prospective application domains rather than deployment claims established by the present experiments.

15.1 Learning and computational systems

Foundation-model adaptation and data-domain allocation. Fine-tuning a large model often requires choosing among domains, task mixtures, adapters, or continuation depths under a limited compute budget. Two checkpoints can look similar under current loss or held-out behavior yet respond differently to the same candidate domain update. A short domain-specific training intervention can therefore serve as a future-learning experiment: its response can refine the decision quotient over which domain or adapter should be promoted, while the selected probe trajectory can remain useful training. This differs from using a learning curve only to forecast final loss; the target is whether newly revealed learning state changes a downstream adaptation decision. The DynaMiCS-style comparison studied here is one concrete instance of this broader allocation problem [14].

Continual learning, curriculum control, rehearsal allocation, and AutoML. In continual or curriculum learning, present task performance need not reveal how much plasticity, interference, or recoverability remains for the next phase. A bounded amount of rehearsal or target-task training can be treated as an experiment on that hidden state before the curriculum decision is fixed. The same principle applies to hyperparameter and configuration search: partial-training methods already pause, terminate, or promote candidate runs from intermediate performance [7, 17, 25, 37], but Revelation Control asks whether additional computation should be purchased because its *response* is expected to change the decision. A run with a mediocre current metric can justify continuation when deeper revelation has positive decision value, while a promising-looking run can be stopped when additional information is unlikely to alter the selected action. For long-horizon or lifelong learning, an important extension is to identify a minimal revelation state that remains dynamically sufficient across successive stages; if such a state satisfies multi-stage Markov closure, Revelation Control could be composed recursively rather than applied only as a one-step continuation rule. Productive promotion further distinguishes useful exploratory work from discarded evaluation cost.

Intervention-aware data acquisition. The intervention can also be a training batch, synthetic-data block, augmentation regime, rehearsal set, or other legal update whose response is observable before a larger commitment is made. Classical Bayesian experimental design and active sensing choose observations by expected decision value [6, 26, 39]. In a learning system, the experiment can instead be an update to the learner itself. Development-only future responses can be used to estimate a model-specific revealability coordinate or continuation state and to determine whether current information is already decision-sufficient. This suggests a route to choosing not merely which data are informative, but which data reveal something about the learner's *future response* that changes the action menu.

Adaptive computation beyond training. The depth variable need not literally count gradient updates. In an iterative solver, multi-fidelity simulation, branch-and-bound routine, Monte Carlo computation, or other staged numerical procedure, a coarse computation may be insufficient to choose a downstream action while additional computation both reveals information and accumulates reusable work. If the current approximation, the refinement response, and the terminal decision can be placed in a common utility ledger, revelation depth becomes computation depth. The same continuation principle then asks whether another refinement level is worth purchasing because of the decision it may change, connecting Revelation Control to value-of-computation ideas [13, 36] while retaining the productive-reuse term.

15.2 Beyond machine learning: productive experiments in sequential decisions

Control, system identification, and robotics. A physical system can occupy internal states with nearly identical current outputs but different future responses and therefore different optimal controls. Examples include uncertain friction, payload, contact mode, actuator health, or local dynamics. Dual control already formalizes the fact that an action can simultaneously control a system and reveal uncertainty [11], while POMDP and active-sensing formulations treat task-directed information acquisition under partial observability [21, 39]. Revelation Control suggests a decision-relative refinement of this idea when a short controlled motion or excitation can be priced, its response can reveal which control action is appropriate, and the selected probing trajectory can be retained rather than discarded. The theory does not require complete system identification: only the hidden distinctions that change the terminal action need to be resolved.

Industrial process control, diagnostics, and maintenance. Machines or production processes can present similar current telemetry while differing in latent wear, loading, material state, or failure proximity. A short controlled operating segment, calibration cycle, or diagnostic excitation can reveal whether the correct action is to continue production, derate, switch operating mode, inspect, or maintain. Such settings are especially close to productive revelation when the diagnostic segment also performs useful work. The relevant accounting must, however, include downtime, safety margin, irreversible wear, and state-restoration cost; if probing itself damages the system or changes the target regime, the promotion geometry used in the present experiments no longer applies without modification.

Scientific experimentation and automated laboratories. In materials, chemistry, biology, and other experimental sciences, an early observation may leave several latent mechanisms compatible with the data even though those mechanisms imply different next experiments or operating decisions. Bayesian experimental design already provides a mature decision-theoretic language for choosing informative experiments [6, 26]. Revelation Control is potentially relevant when an experiment can be staged: an initial intervention is run only far enough to determine whether deeper measurement or continuation is decision-relevant, and the partial trajectory can be retained if that branch is selected. Automated laboratories, adaptive measurement pipelines, and multi-stage physical experiments are natural settings in which "how much experiment to purchase" may be as important as "which experiment to run." Establishing this connection rigorously would require domain-specific transition models, measurement error, and experimental-cost accounting beyond the present paper.

Operations, resource allocation, and pilot-to-scale decisions. Many operational decisions involve latent regimes that cannot be resolved from passive current data alone. A limited allocation, pilot deployment, test market, routing perturbation, or operating-policy trial can reveal response heterogeneity before a larger commitment is made. The Revelation-Control abstraction is appropriate when the pilot has a well-defined terminal decision, its response can be evaluated before scale-up, and at least part of the pilot's value or infrastructure is reusable if promoted. The central distinction is between *learning about the world* and *learning enough to choose the action*: a pilot need not identify the full latent regime if it already resolves the relevant decision quotient. Strategic behavior, interference, nonstationarity, and general-equilibrium effects can break the simple controlled-probe assumptions and would have to be modeled explicitly.

High-stakes human decision systems. The mathematical pattern also appears, at least abstractly, in adaptive treatment, public-policy pilots, and personalized education: two individuals or populations with similar current observables can respond differently to a small intervention, and the response can change the preferred next action. These domains are intentionally placed at the outer boundary of the present paper's application claims. A valid extension would require causal identification, ethical constraints, uncertainty about intervention harm, fairness considerations, and stronger safety guarantees than the bounded-risk analysis developed here. In medicine or policy especially, "productive probing" cannot be assumed merely because an intervention also has intended benefit; the value ledger must price adverse effects and irreversibility, and experimentation may be impermissible even when its statistical information value is high. The present work therefore supplies only a possible decision-theoretic template, not a deployment prescription.

15.3 A domain-level applicability test

The most useful transfer question is not whether a new field resembles model fine-tuning superficially, but whether its decision structure matches the theory. For a new system, one should first declare the legal current information, the terminal action menu, terminal utility, admissible probe interventions, and all costs or safety penalties. Development-only probe responses can then be used to estimate the conditional continuation value and to test whether the current scalar summary is already decision-sufficient. If scalar fibers cross the priced STOP/CONTINUE boundary, additional revealability information is decision-nonredundant and a richer controller is required; if they do not, a simpler controller can be sufficient even when hidden-state variation remains. The resulting architecture and cost ledger should be fixed before target evaluation. If the probe leaves reusable work, information value and productive reuse must be reported separately; if it does not, the problem reduces to discard-and-restart information acquisition. This procedure is the general method that transfers across domains: the decision factorization, continuation-value criterion, and accounting principles are invariant, whereas observable proxies, dynamics, coefficients, safety constraints, and prices are system-specific.

16 Limitations

Scope of the empirical systems. The empirical study now spans two independently instantiated 7B decoder-only Transformer families, but both use LoRA adaptation, AdamW-style optimizer state, the same binary intervention menu, and the same H4/H8/H12 control geometry. Cross-optimizer, substantially different-scale, non-Transformer, and materially different-horizon generality remain open.

Comparator scope. The finite-library result does not imply dominance over every measurable current-information policy. The frontier comparison is against one strengthened DynaMiCS-style short-probe/restart class under a common policy-visible update budget, not against every possible probing or metareasoning method.

Productive revelation assumptions. Promotion is useful only when tested work remains legally and scientifically reusable. If a probe corrupts the deployment path, creates irreversible risk, changes the target distribution, or incurs additional state-I/O or safety cost, restart may be the appropriate technology and the budget identity must be repriced.

Conditional rather than universal revealability dimension. The location–scale theorem is a structural statement about $(|m_h|, \sigma_h)$, not a claim that every model must exhibit two empirically nonredundant shallow coordinates. The exact scalar-control factorization theorem further shows that residual variation in σ_h need not be decision-relevant: a second coordinate is required only when a positive-mass scalar fiber crosses the priced STOP/CONTINUE boundary. Qwen provides evidence for a decision-nonredundant regime, while Mistral is consistent with scalar decision sufficiency at the observed resolution. The empirical R_4 used in Qwen is one legal model-specific proxy, not a universal statistic.

Tail-robust utility certification. Controlling the probability that an early stop changes the deeper action does not by itself certify positive expected utility when stopped-flip severity is unrestricted. Theorem B.3 shows that this is an identification boundary rather than merely a power limitation: a population net-utility certificate additionally requires a predeclared severity, moment, or integrable-tail class. Positive held-out panel-level utility is therefore not a universal tail-robust guarantee.

Applications beyond learning systems. The mathematical decision structure can be instantiated whenever a priced controlled intervention reveals decision-relevant hidden state before commitment, but the empirical evidence in this paper is confined to learning systems. Physical, scientific, operational, medical, policy, and educational settings introduce domain-specific causality, irreversibility, strategic response, measurement error, safety, fairness, and ethical constraints that are not resolved here. The broader domains in Section 15 are therefore prospective uses of the framework, not validated deployments or claims of domain-universal optimality.

Method optimality. The evaluated policies are selected using development data and fixed before target evaluation. No result establishes global optimality over all representations, learners, stopping rules, probe schedules, intervention menus, or acquisition policies.

Compute and deployment value. The main normalization counts policy-visible training updates under a common model and batch construction. Restore/save overhead is not charged against the frontier in the main comparison. Device-seconds, wall-clock, memory, state-I/O, monetary value, safety value, and live-deployment utility are different resources and are not collapsed into the primary claim.

Dynamic-state compression. The location–scale result is an exact one-step statement. One-step low-dimensional decision sufficiency does not imply multi-stage Markov closure. A multi-depth controller can therefore use (m_h, σ_h) as a complete low-dimensional state only under an additional Markov-closure condition for future margin–revealability dynamics. Characterizing minimal dynamically sufficient revelation states, and the conditions under which they induce Markov closure across revelation depths, is a natural next theoretical problem.

17 Conclusion

The companion work established that present behavior can hide distinctions in future learning. The present paper closes the decision-theoretic step: such distinctions can have action value, can be actively revealed by controlled future learning, and can become a state-dependent control variable. We call this problem Revelation Control.

The theory identifies the relevant object as a decision quotient rather than the full hidden state. A coarse observation is sufficient only when the refined optimal action factorizes through it. The fork–probe–promote realization studied here uses future learning as an experiment on hidden learner state, while productive revelation keeps the selected tested path as useful computation. The resulting value ledger separates pure information gain from the value of already executed work. Static Bayes refinement value then closes exactly into a conditional revelation value, yielding an optimal-stopping rule over revelation depth.

The empirical evidence supports this structure in two independently instantiated Transformer families. In Qwen2.5-7B, a fixed H8 policy strictly beats a declared 22-component current-information library on two independent panels, refines a fixed H4 ambiguity regime into groups with opposite terminal action value, and obtains strict equal-compute advantage over a strengthened DynaMiCS-style H4 restart frontier through productive reuse. In Mistral-7B-v0.3, an independent two-bank panel reproduces positive H8-over-H4 decision value and the same productive equal-compute mechanism with positive one-sided lower bounds. The cross-model result is therefore not numerical reuse of one fitted policy: it is replication of the information-to-utility structure predicted by the theory.

Adaptive control exhibits the regime dependence predicted by the theory. Qwen provides evidence that a second shallow revealability coordinate can matter beyond current margin; a familywise-adjusted Mistral structural validation, restricted to the finite pre-existing architecture family and fit only on independent development data, supports a successful scalar continuation rule at the tested resolution. The exact scalar-control factorization result makes the common structure precise: an additional coordinate is necessary for the stopping decision only when states sharing the scalar summary fall on both sides of the cost-adjusted continuation boundary. Thus predictive or value variation can remain present without being decision-nonredundant. A separate Mistral risk-calibrated controller further illustrates the impossibility boundary: sub-2% decision-instability risk can coexist with nonpositive net utility when a few stopped flips have large severity.

The resulting conclusion is specific but broad in implication. Hidden learning state is not merely predictive structure. Controlled future learning can reveal distinctions that change decisions; productive reuse can convert revelation into higher utility at the same visible training-update budget; and the amount of revelation can itself be controlled according to state-dependent expected value. More generally, Section 15 identifies the same decision pattern in control and system identification, robotics and industrial diagnostics, scientific experimentation, adaptive computation, and pilot-to-scale operational decisions, with high-stakes human domains requiring additional causal, ethical, and safety theory. What must transport across systems is the decision structure and value accounting, not a universal proxy, coefficient, threshold, or even a fixed empirical state dimension. Revelation Control therefore provides a theory-first framework for deciding not only *what* action to take, but *how much controlled future interaction it is worth purchasing in order to know*.

A Conservative replicate-denoised Bayes-regret bridge

This appendix records an optional stronger certification device. It is not required for the main empirical conclusions in Sections 11–12. Independently generated future outcomes can be used to separate persistent conditional signal from one-bank realization noise and to upper-bound coarse-policy Bayes regret under the stated assumptions.

Consider binary actions with terminal gap

$$D = Q_{\mathrm{XI}} - Q_{\mathrm{KEEP}}. \tag{A.1}$$

Let future banks A and B share a pre-bank sigma-field $\mathscr{X}$ and satisfy

$$\mathbb{E}[D_A \mid \mathscr{X}] = \mathbb{E}[D_B \mid \mathscr{X}] =: \mu. \tag{A.2}$$

The legal coarse information obeys $\mathscr{H} \subseteq \mathscr{X}$, and

$$m_{\mathscr{H}} = \mathbb{E}[\mu \mid \mathscr{H}] \tag{A.3}$$

provides the coarse-information Bayes score. Let f be any fixed $\mathscr{H}$-measurable score in the same utility units as D, and let

$$\pi_f = \mathbf{1}\{f > 0\}. \tag{A.4}$$

Here and below, action indicator one denotes XI and zero denotes KEEP.

Define the cross-replicate residual moment

$$M_\times(f) = \mathbb{E}[(D_A - f)(D_B - f)]. \tag{A.5}$$

Let

$$c_{\mathscr{X}} = \mathrm{Cov}(D_A, D_B \mid \mathscr{X}). \tag{A.6}$$

Conditional independence gives $c_{\mathscr{X}} = 0$; allowing $c_{\mathscr{X}} \geq 0$ yields the same conservative direction.

Theorem A.1 (Replicate-denoised Bayes-regret bridge)**.** *Assume square integrability, a shared conditional mean across banks, and $c_{\mathscr{X}} \geq 0$ almost surely. Then*

$$M_\times(f) = \mathbb{E}[(\mu - f)^2] + \mathbb{E}[c_{\mathscr{X}}] \tag{A.7}$$

$$= \mathbb{E}[(m_{\mathscr{H}} - f)^2] + \mathbb{E}[\mathrm{Var}(\mu \mid \mathscr{H})] + \mathbb{E}[c_{\mathscr{X}}], \tag{A.8}$$

and therefore

$$\boxed{V_B(\mathscr{H}) - \mathcal{V}(\pi_f) \leq \sqrt{M_\times(f)}.} \tag{A.9}$$

Consequently, for any policy π_G measurable with respect to a refinement $\mathscr{G} \supseteq \mathscr{H}$,

$$\mathcal{V}(\pi_G) - V_B(\mathscr{H}) \geq \{\mathcal{V}(\pi_G) - \mathcal{V}(\pi_f)\} - \sqrt{M_\times(f)}. \tag{A.10}$$

The decomposition shows why the bridge is conservative. Even when $f = m_{\mathscr{H}}$, persistent hidden heterogeneity contributes the nonnegative floor

$$V_{\mathrm{hid}} := \mathbb{E}[\mathrm{Var}(\mu \mid \mathscr{H})]. \tag{A.11}$$

The bridge is therefore sufficient rather than necessary: not clearing it does not imply that the active value is below the coarse-information Bayes value.

Corollary A.2 (Conditional bridge)**.** *Let $A \in \mathscr{H}$ with $\mathbb{P}(A) > 0$. Replacing all expectations by expectations conditional on A gives*

$$V_B(\mathscr{H}; A) - \mathcal{V}(\pi_f; A) \leq \sqrt{M_\times(f; A)}, \tag{A.12}$$

where

$$M_\times(f; A) = \mathbb{E}[(D_A - f)(D_B - f) \mid A]. \tag{A.13}$$

B Adaptive-depth certification details

This appendix records the structural certification layer used to interpret adaptive revelation depth. The generic risk-calibration tools are established prior art; the purpose here is to state how they interact with productive-prefix reuse and terminal decision severity in the present execution technology.

Let $S \in \{0, 1\}$ denote an H4 early-stop decision, let $F = \mathbf{1}\{\pi_4 \neq \pi_8\}$, and define $Y = Q_{\pi_8} - Q_{\pi_4}$. Each H4 stop saves compute value $c = 4\lambda_0$ relative to always-H8.

Proposition B.1 (Productive-prefix decision invariance)**.** *Under the declared same-path promotion technology, if $F = 0$ then $Y = 0$. For the binary action menu,*

$$Y = D\left(\mathbf{1}\{\pi_8 = \mathrm{XI}\} - \mathbf{1}\{\pi_4 = \mathrm{XI}\}\right), \qquad |Y| = |D|F. \tag{B.1}$$

Proposition B.2 (Risk–severity factorization)**.** *Let $\rho = \Pr(S = 1)$ and $r = \Pr(S = 1, F = 1)$. When $r > 0$, define the signed stopped-flip severity $\eta_s = \mathbb{E}[Y \mid S = 1, F = 1]$ and the positive-harm severity $\eta_+ = \mathbb{E}[Y_+ \mid S = 1, F = 1]$, where $Y_+ = \max(Y, 0)$. When $r = 0$, set $\eta_s = \eta_+ = 0$ by convention. Then*

$$\boxed{\Delta V(S) = \mathbb{E}[S(c - Y)] = c\rho - r\eta_s.} \tag{B.2}$$

Moreover,

$$\Delta V(S) \geq c\rho - r\eta_+. \tag{B.3}$$

Theorem B.3 (Risk-only impossibility and the harm-tail boundary)**.** *Let $E = \{S = 1, F = 1\}$ and define the stopped positive-harm mass*

$$H_+ := \mathbb{E}[Y_+ \mathbf{1}_E] \in [0, \infty]. \tag{B.4}$$

Then

$$H_+ = \int_0^\infty \Pr(E,\, Y_+ > t)\, dt, \qquad \Delta V(S) \geq c\rho - H_+. \tag{B.5}$$

Moreover, fix any $0 < r \leq \rho \leq 1$. For every $M > 0$ there exists a joint law of (S, F, Y) satisfying productive-prefix decision invariance, with $\Pr(S = 1) = \rho$, $\Pr(E) = r$, and finite $\mathbb{E}[Y_+]$, such that

$$\Delta V(S) < -M. \tag{B.6}$$

Hence, whenever nonzero stop–flip risk is permitted, exact knowledge of (ρ, r) alone gives no finite distribution-robust lower bound on expected net utility over a class with unrestricted stopped-flip severity. Any positive expected-utility certificate must additionally control H_+, or a sufficient severity/tail surrogate for it.

Corollary B.4 (Minimal sufficient harm-tail bridges)**.** *The positive-harm component left uncontrolled by stop–flip risk is summarized by the integrated stopped-harm tail H_+ in Equation* (B.4)*. The following declared conditions are sufficient ways to control it for a lower utility certificate.*

If $Y_+ \leq B_+$ almost surely, then

$$\Delta V(S) \geq c\rho - B_+ r. \tag{B.7}$$

If instead $q > 1$ and $\mathbb{E}[|D|^q] \leq M_q$, then Hölder's inequality gives

$$\Delta V(S) \geq c\rho - M_q^{1/q} r^{1-1/q}. \tag{B.8}$$

More generally, if an integrable envelope $u : [0, \infty) \to [0, \infty)$ satisfies

$$\Pr(E,\, Y_+ > t) \leq u(t) \qquad \text{for all } t \geq 0, \tag{B.9}$$

then

$$\Delta V(S) \geq c\rho - \int_0^\infty u(t)\, dt. \tag{B.10}$$

Thus a finite-sample event $\rho \geq L_\rho$, $r \leq U_r$ yields certified lower bounds $cL_\rho - B_+ U_r$ or $cL_\rho - M_q^{1/q} U_r^{1-1/q}$ under the corresponding declared severity class; an independently valid upper bound U_H on H_+ yields the assumption-matched bound $cL_\rho - U_H$.

Theorem B.3 is a control-specific boundary, not a claim that bounded-risk calibration is impossible. The Bernoulli event E can be calibrated without utility-tail assumptions; what cannot be obtained from that event probability alone is a nontrivial lower bound on an unbounded mean severity. At the statistical level, this distinction is aligned with classical nonparametric impossibility results for unrestricted mean inference [3].

Bounded decision-instability calibration. The loss $L = \mathbf{1}\{S = 1, F = 1\}$ is Bernoulli. Hence fixed-sequence calibration or other established bounded-risk procedures such as Learn-then-Test and Conformal Risk Control can be used prospectively once the risk model, threshold family, calibration data, and risk target are fixed [1, 2]. This controls how often early stopping can change the deeper action, not how severe those changed decisions are.

Fixed-data risk–severity diagnostic. A fixed-data nested evaluation illustrates the distinction. A two-coordinate H4 flip-risk score is trained on one collection block, calibrated on one future bank of the other block to target 2% joint stop-and-flip risk, and checked on the remaining future bank. Table 6 reports the held-out outcomes. The first direction has positive realized net value; the second has a lower flip rate but negative realized net value because its single harmful flip has substantially larger severity. This is a diagnostic validation of Equation (B.2), not a prospective controller claim.

Table 6. Held-out risk–severity diagnostic for a derived risk-calibrated stopping rule. Net value is relative to always-H8.

Direction	Stop rate	Stop–flip rate	Harm events	Max harm severity	Net value
A → B	26.67%	0.833%	1	2.083×10^{-4}	$+1.186 \times 10^{-5}$
B → A	11.67%	0.417%	1	2.010×10^{-3}	-4.331×10^{-6}

The diagnostic shows that a probability budget is not a utility budget: decision-instability risk controls how often early stopping can differ from H8, while the integrated harm tail controls how costly those differences are. Theorem B.3 makes this separation sharp. A fully prospective expected-utility certificate therefore needs both a bounded-risk calibration and a predeclared severity, moment, or integrable tail class.

C Proofs

C.1 Proof of Proposition 2.2

Let $M_a = \mathbb{E}[Q_a \mid \mathcal{G}]$. Because $\mathcal{A}$ is finite, a $\mathcal{G}$-Bayes action exists after fixing a measurable tie rule. If a $\mathcal{G}$-Bayes action π_H is $\mathcal{H}$-measurable, then

$$V_B(\mathcal{G}) = \mathbb{E}[M_{\pi_H}] = \mathbb{E}[Q_{\pi_H}] \le V_B(\mathcal{H}) \le V_B(\mathcal{G}), \tag{C.1}$$

so equality holds. Conversely, let $\pi_H^\star$ be an $\mathcal{H}$-Bayes policy. If $V_B(\mathcal{H}) = V_B(\mathcal{G})$, then

$$0 = V_B(\mathcal{G}) - \mathcal{V}(\pi_H^\star) = \mathbb{E}\Big[\max_a M_a - M_{\pi_H^\star}\Big]. \tag{C.2}$$

The integrand is nonnegative, hence it vanishes almost surely and $\pi_H^\star$ is also $\mathcal{G}$-Bayes optimal.

C.2 Proof of Proposition 2.3

By definition,

$$O_t^{\mathrm{DF}} = L_t + \max_a\{\nu_t(a) + \gamma\omega_t(a)\} - \nu_t^\star \tag{C.3}$$

$$= L_t + \max_a\{\gamma\omega_t(a) - (\nu_t^\star - \nu_t(a))\} \tag{C.4}$$

$$= L_t + \max_a\{\gamma\omega_t(a) - d_t(a)\}. \tag{C.5}$$

The maximum term is nonnegative because any coarse-optimal action has $d_t(a) = 0$ and $\omega_t(a) \ge 0$. Since $L_t \ge 0$, the sum is zero if and only if both $L_t = 0$ and every term inside the maximum is nonpositive.

C.3 Proof of Proposition 4.1

Temporary probing spends mc updates across m candidates and then restarts the selected action for the full horizon H, so $K_F(c) = mc + H$. Probe-and-promote revelation spends mh updates across candidates and then only $H - h$ further updates on the selected path, so $K_{\mathrm{AR}}(h) = H + (m-1)h$. Equating the budgets gives $(m-1)h = mc$.

C.4 Proof of Proposition 6.1

Let $V_{\Pi_s}^\star = V_{P_s}^\star - R_s^{\mathrm{app}}$ be the best value in the learner class and let $W_{s,n} = V_{\Pi_s}^\star - R_{s,n}^{\mathrm{est}}$ be the learned gross value. Since $V_{P_s}^\star = V_B(\mathcal{H}) + \Delta_s^{\mathrm{rev}} + \Delta_s^{\mathrm{prom}}$ and $W_{s,n}^{\mathrm{net}} = W_{s,n} - C_s$, substitution yields Equation (6.3).

C.5 Proof of Corollary 6.2

Using $V_P(\hat{\pi}_A) = V_B(\mathcal{G}) - R_A + P_A$ and $V_R(\hat{\pi}_F) = V_B(\mathcal{H}) - R_F$,

$$V_P(\hat{\pi}_A) - V_R(\hat{\pi}_F) = [V_B(\mathcal{G}) - V_B(\mathcal{H})] + P_A + R_F - R_A. \tag{C.6}$$

Subtracting priced resource costs gives the net-value version.

C.6 Proof of Theorem 3.2

Relative to KEEP, the refined Bayes increment is $\mathbb{E}[(m_G)_+]$ and the coarse Bayes increment is $\mathbb{E}[(m_H)_+]$. Conditioning the refined increment on $\mathscr{H}$ gives a_H. Since

$$m_H = \mathbb{E}[m_G \mid \mathscr{H}] = a_H - b_H, \tag{C.7}$$

we have pointwise

$$a_H - (m_H)_+ = a_H - (a_H - b_H)_+ = \min\{a_H, b_H\}. \tag{C.8}$$

Taking expectations proves Equation (3.8). On the event A in Equation (3.9),

$$a_H \geq \delta\alpha, \qquad b_H \geq \delta\beta, \tag{C.9}$$

so $\min\{a_H, b_H\} \geq \delta \min(\alpha, \beta)$ there. Integrating over A gives Equation (3.10).

C.7 Proof of Proposition 3.3

By the constant-rank theorem, the local level set $\Phi_c^{-1}(\Phi_c(x_0))$ is a submanifold with tangent space $\ker D\Phi_c(x_0)$. Hence there is a C^1 curve $\gamma(s)$ in that level set with $\gamma(0) = x_0$ and $\gamma'(0) = v$. Taylor expansion gives

$$g(\gamma(s)) = s\, Dg(x_0)v + o(s), \tag{C.10}$$

so for sufficiently small positive and negative s the terminal gap has opposite signs. Likewise

$$Y_h(\gamma(s)) = Y_h(x_0) + s\, DY_h(x_0)v + o(s), \tag{C.11}$$

which changes to first order because $DY_h(x_0)v \neq 0$.

C.8 Proof of the replicate-denoised bridge

By iterated expectation,

$$M_\times(f) = \mathbb{E}[\mathbb{E}[(D_A - f)(D_B - f) \mid \mathscr{X}]] \tag{C.12}$$

$$= \mathbb{E}[(\mu - f)^2] + \mathbb{E}[c_{\mathscr{X}}]. \tag{C.13}$$

Since f is $\mathscr{H}$-measurable and $m_{\mathscr{H}} = \mathbb{E}[\mu \mid \mathscr{H}]$, the conditional Pythagorean identity gives

$$\mathbb{E}[(\mu - f)^2] = \mathbb{E}[(m_{\mathscr{H}} - f)^2] + \mathbb{E}[\mathrm{Var}(\mu \mid \mathscr{H})]. \tag{C.14}$$

For binary actions with the declared tie rule $\pi_f = \mathbf{1}\{f > 0\}$, the regret relative to the $\mathscr{H}$-Bayes rule $\pi_m = \mathbf{1}\{m_{\mathscr{H}} > 0\}$ is

$$V_B(\mathscr{H}) - \mathcal{V}(\pi_f) = \mathbb{E}\big[|m_{\mathscr{H}}|\, \mathbf{1}\{\pi_f \neq \pi_m\}\big]. \tag{C.15}$$

On action disagreement, $|m_{\mathscr{H}}| \leq |m_{\mathscr{H}} - f|$; ties at $m_{\mathscr{H}} = 0$ contribute zero regret. Cauchy–Schwarz therefore yields

$$V_B(\mathscr{H}) - \mathcal{V}(\pi_f) \leq \mathbb{E}|m_{\mathscr{H}} - f| \leq \sqrt{\mathbb{E}[(m_{\mathscr{H}} - f)^2]} \leq \sqrt{M_\times(f)}. \tag{C.16}$$

Adding and subtracting $\mathcal{V}(\pi_f)$ proves the policy-improvement inequality. Conditioning throughout on $A \in \mathscr{H}$ proves the conditional version.

C.9 Proof of Proposition 5.1

By the tower property,

$$\mathbb{E}[\mathcal{G}_{h\to h'}] = \mathbb{E}[\mathbb{E}[C_{h'} \mid \mathscr{F}_h] - C_h] \tag{C.17}$$

$$= \mathbb{E}[C_{h'}] - \mathbb{E}[C_h] \tag{C.18}$$

$$= V_B(\mathscr{F}_{h'}) - V_B(\mathscr{F}_h). \tag{C.19}$$

The final equality is the definition of Bayes value under each information state.

C.10 Proof of Proposition 5.4

At the final admissible depth h_J, no deeper revelation is available, so $V_J^{\mathrm{dyn}} = C_{h_J}$. At an earlier depth, the controller has exactly two admissible choices in the stated finite-depth problem: commit immediately for value C_{h_j}, or buy the next refinement, pay c_j, and receive conditional expected value $\mathbb{E}[V_{j+1}^{\mathrm{dyn}} \mid \mathscr{F}_{h_j}]$. Taking the larger value gives Equation (5.14). Backward induction completes the recursion.

C.11 Proof of Proposition 5.2

Given $\mathscr{F}_h$, the conditional net increment from continuing is $G_{h\to h'} - c_{h\to h'}$; hence continuation is optimal exactly when this quantity is positive.

C.12 Proof of Theorem 5.3

Write $\Gamma = \Gamma_{h\to h'}$ and condition on the scalar summary S_h. Since

$$\Gamma = \Gamma_+ - (-\Gamma)_+, \tag{C.20}$$

we have

$$\mathbb{E}[\Gamma \mid S_h] = a(S_h) - b(S_h). \tag{C.21}$$

Therefore

$$\begin{aligned} V_{\mathrm{meta}}(\mathscr{F}_h) - V_{\mathrm{meta}}(S_h) &= \mathbb{E}[a(S_h)] - \mathbb{E}[(a(S_h) - b(S_h))_+] && \text{(C.22)}\\ &= \mathbb{E}[a(S_h) - (a(S_h) - b(S_h))_+] && \text{(C.23)}\\ &= \mathbb{E}[\min\{a(S_h), b(S_h)\}], && \text{(C.24)} \end{aligned}$$

which proves Equation (5.11). Because $a, b \geq 0$, the gap is zero if and only if $\min\{a, b\} = 0$ almost surely. For an integrable random variable, $a(S_h) = 0$ exactly when $\Pr(\Gamma > 0 \mid S_h) = 0$ almost surely on that scalar fiber, and similarly $b(S_h) = 0$ exactly when $\Pr(\Gamma < 0 \mid S_h) = 0$. Thus equality holds if and only if no positive-mass scalar fiber contains both strictly positive and strictly negative net continuation gains. Equivalently, there exists a full-information optimal meta-action whose tie assignment at $\Gamma = 0$ is measurable with respect to S_h; that optimal meta-action then factorizes through S_h. This is precisely value-preserving decision factorization for the two-action meta-menu $\{\mathrm{STOP}, \mathrm{CONTINUE}\}$.

C.13 Proof of Theorem 5.5

By symmetry and integrability of Z, $\mathbb{E}[Z] = 0$. If $m_h \geq 0$, then

$$\begin{aligned} \mathbb{E}[(m_h + \sigma_h Z)_+ \mid \mathscr{F}_h] - m_h &= \mathbb{E}[(-m_h - \sigma_h Z)_+ \mid \mathscr{F}_h] && \text{(C.25)}\\ &= \sigma_h \mathbb{E}[(Z - m_h/\sigma_h)_+], && \text{(C.26)} \end{aligned}$$

where the last equality uses symmetry. If $m_h < 0$, the coarse positive-part value is zero and

$$\mathbb{E}[(m_h + \sigma_h Z)_+ \mid \mathscr{F}_h] = \sigma_h \mathbb{E}[(Z - |m_h|/\sigma_h)_+]. \tag{C.27}$$

This proves Equation (5.19). Since $\psi'(t) = -\Pr(Z > t)$ at continuity points,

$$\frac{\partial \mathcal{R}}{\partial |m_h|} = \psi'(t) = -\Pr(Z > t). \tag{C.28}$$

For $t = |m_h|/\sigma_h$,

$$\frac{\partial \mathcal{R}}{\partial \sigma_h} = \psi(t) - t\psi'(t) = \psi(t) + t\Pr(Z > t) \geq 0 \tag{C.29}$$

with strict inequality whenever the refinement retains positive tail mass beyond t.

C.14 Proof of Corollary 5.6

If $\sigma_h = g(M_h)$ on E, substitution into Equation (5.19) gives

$$\mathcal{R}_{h\to h'} = g(M_h)\psi(M_h/g(M_h)) = \widetilde{r}(M_h) \tag{C.30}$$

on E, proving scalar sufficiency for the one-step continuation value. A threshold representation is a stronger statement and requires the resulting scalar value function to be monotone relative to cost.

For the second part, condition on $M_h = m$. By assumption, the conditional distribution of σ_h is nondegenerate on a set of m values with positive probability, and $s \mapsto r(m, s)$ is strictly increasing over that conditional support. A strictly monotone transform of a nondegenerate random variable is nondegenerate. Equation (5.19) identifies this transform with $\mathcal{R}_{h\to h'}$, so the conditional law of $\mathcal{R}_{h\to h'}$ given M_h is nondegenerate on that set and no M_h-measurable scalar function can equal the continuation value almost surely there.

C.15 Proof of Corollary 5.7

For fixed $M_h = m$, let

$$\Gamma = \mathcal{R}_{h\to h'} - c = r(m, \sigma_h) - c. \tag{C.31}$$

Theorem 5.3 gives scalar sufficiency exactly when, conditional on almost every m, Γ does not have both strictly positive and strictly negative mass. This is equivalent to Equation (5.26): either there is no strictly positive mass, or there is no strictly negative mass. Zero-gain ties may be assigned to either optimal side without changing value. Under continuity and strict monotonicity, $r(m, s) \le c$ for $s \le \sigma_c(m)$ and $r(m, s) \ge c$ for $s \ge \sigma_c(m)$ on the relevant support (with equality possible only at the crossing), which gives the stated support-side sufficient condition.

For the second statement, Equation (5.27) implies that on a positive-probability set of margin fibers there is positive conditional probability of both $\Gamma < 0$ and $\Gamma > 0$. Hence both conditional positive-part expectations $a(M_h)$ and $b(M_h)$ from Theorem 5.3 are strictly positive on that set. Their minimum therefore has strictly positive expectation, so Equation (5.11) yields

$$V_{\text{meta}}(\mathscr{F}_h) - V_{\text{meta}}(M_h) > 0. \tag{C.32}$$

C.16 Proof of Proposition 5.8

Equation (5.16) gives $m_{h'} - m_h = \sigma_h Z$. Because σ_h is $\mathscr{F}_h$-measurable and Z is independent of $\mathscr{F}_h$,

$$\mathbb{E}[|m_{h'} - m_h| \mid \mathscr{F}_h] = \sigma_h \mathbb{E}|Z|. \tag{C.33}$$

If $\mathbb{E}[Z^2] = 1$, the same argument yields

$$\mathbb{E}[(m_{h'} - m_h)^2 \mid \mathscr{F}_h] = \sigma_h^2 \mathbb{E}[Z^2] = \sigma_h^2. \tag{C.34}$$

C.17 Proof of Proposition B.1

If $F = 0$, then $\pi_4 = \pi_8$ and both policies promote the same already-tested path under the declared same-path technology. Their H12 terminal outcome is therefore identical, so $Y = Q_{\pi_8} - Q_{\pi_4} = 0$. For the binary action menu,

$$Q_{\pi_j} = Q_{\text{KEEP}} + D\mathbf{1}\{\pi_j = \text{X1}\}, \tag{C.35}$$

for $j \in \{4, 8\}$. Subtracting the two identities gives Equation (B.1); because the two action indicators differ exactly when $F = 1$, $|Y| = |D|F$.

C.18 Proof of Proposition B.2

By Proposition B.1, $SY = 0$ outside $E = \{S = 1, F = 1\}$. Hence

$$\Delta V(S) = \mathbb{E}[S(c - Y)] \tag{C.36}$$

$$= c\Pr(S = 1) - \mathbb{E}[Y\mathbf{1}_E] \tag{C.37}$$

$$= c\rho - r\mathbb{E}[Y \mid E] = c\rho - r\eta_s, \tag{C.38}$$

when $r > 0$. Since $Y \le Y_+$ pointwise,

$$\mathbb{E}[Y\mathbf{1}_E] \le \mathbb{E}[Y_+\mathbf{1}_E] = r\eta_+, \tag{C.39}$$

which yields Equation (B.3). The case $r = 0$ reduces to $\Delta V(S) = c\rho$.

C.19 Proof of Theorem B.3

For the nonnegative random variable $Y_+\mathbf{1}_E$, the layer-cake identity gives

$$\mathbb{E}[Y_+\mathbf{1}_E] = \int_0^\infty \Pr(Y_+\mathbf{1}_E > t)\, dt = \int_0^\infty \Pr(E,\, Y_+ > t)\, dt. \tag{C.40}$$

Combining this identity with Equation (B.3) proves Equation (B.5).

For the impossibility statement, fix $0 < r \le \rho \le 1$ and $M > 0$. Choose a joint law with

$$\Pr(S = 1, F = 1) = r, \qquad \Pr(S = 1, F = 0) = \rho - r, \tag{C.41}$$

and allocate the remaining probability to $S = 0, F = 0$. Set $Y = 0$ whenever $F = 0$ and set

$$Y = K := \frac{M + c\rho + 1}{r} \tag{C.42}$$

on $E = \{S = 1, F = 1\}$. This law satisfies productive-prefix decision invariance and has finite $\mathbb{E}[Y_+] = rK$. Its net value is

$$\Delta V(S) = c\rho - rK = -M - 1 < -M. \tag{C.43}$$

Because M is arbitrary while (ρ, r) are fixed, no finite lower bound depending only on (c, ρ, r) can hold uniformly over unrestricted stopped-flip severities.

C.20 Proof of Corollary B.4

If $Y_+ \le B_+$ almost surely, then

$$H_+ = \mathbb{E}[Y_+\mathbf{1}_E] \le B_+ \Pr(E) = B_+ r, \tag{C.44}$$

which gives Equation (B.7). If $\mathbb{E}[|D|^q] \le M_q$ with $q > 1$, Proposition B.1 and Hölder's inequality yield

$$\begin{aligned} H_+ &\le \mathbb{E}[|D|\mathbf{1}_E] && \text{(C.45)}\\ &\le \mathbb{E}[|D|^q]^{1/q} \Pr(E)^{1-1/q} && \text{(C.46)}\\ &\le M_q^{1/q} r^{1-1/q}, && \text{(C.47)} \end{aligned}$$

proving Equation (B.8). Finally, Equation (B.5) and the assumed tail envelope give

$$H_+ \le \int_0^\infty u(t)\, dt, \tag{C.48}$$

which proves Equation (B.10). Substituting simultaneous confidence bounds for ρ, r, or H_+ gives the stated finite-sample lower certificates.

D Notation and decision objects

Symbol	Meaning
$\mathscr{H}, \mathscr{G}$	Coarse legal information and a refinement
$V_B(\mathcal{I})$	Bayes value under information $\mathcal{I}$
$\mathscr{F}_h, C_h$	Revelation filtration and Bayes commit value at depth h
Q_a	Terminal utility for action a
D	Binary terminal gap $Q_{\text{XI}} - Q_{\text{KEEP}}$
Δ^{rev}	Pure information-refinement value
$\Delta^{\text{prom}}, P_A$	Productive path-reuse / promotion value
$R^{\text{app}}, R^{\text{est}}$	Approximation and finite-sample estimation regret
C_s	Acquisition cost
Φ_c	Complete legal short-probe frontier information map
Y_h	Deeper future-learning observation
$\tau^\star(v)$	First depth at which direction v is revealed
$M_\times(f)$	Cross-replicate residual moment used in the optional conservative bridge
A	Fixed H4 frontier-ambiguity event in the 432-anchor structural analysis
$\Delta_{\text{restart}}, \Delta_{\text{path}}$	Common-restart policy difference and productive path-reuse value
$\mathcal{G}_{h\to h'}$	Oracle conditional Bayes value of purchasing a deeper information refinement
$G_{h\to h'}$	Policy gain from deeper revelation, conditional on $\mathscr{F}_h$
$\Gamma_{h\to h'}$	Cost-adjusted conditional continuation gain $G_{h\to h'} - c_{h\to h'}$
σ_h	State-dependent revealability scale in the location–scale stopping model
$\sigma_c(m)$	Critical revealability at which the priced STOP/CONTINUE decision changes for margin m
$R_h^{(\mathcal{M})}$	Legal model-specific empirical proxy for revealability, fixed without target outcomes
S, F	Early-stop indicator and shallow/deep action-instability indicator
ρ, r	Stop probability and joint stop–flip probability in the risk–severity bridge

E Statistical estimands and simultaneous inference

This appendix collects the inferential rules used by the empirical claims. The unit of resampling and inference is the *exact anchor*, not a bank-level realization, readout, or policy action. When two conditionally independent future banks are available for one anchor, bank-specific contributions are averaged within the anchor before estimating a population mean.

One-sided mean lower bounds. For exact-anchor contributions $Z_1, \dots, Z_n$, let $\bar{Z}$ and s_Z denote the sample mean and standard deviation. The reported one-sided Student-t lower bound is

$$\text{LCB}_{t,1-\alpha} = \bar{Z} - t_{1-\alpha,n-1}\frac{s_Z}{\sqrt{n}}. \tag{E.1}$$

The percentile bootstrap resamples exact anchors with replacement. Unless otherwise stated, 50,000 draws are used and the lower bound is the empirical α quantile of the bootstrap means. These uncertainty summaries target the declared exact-anchor sampling design and panel-generating regime; they are not universal guarantees over arbitrary model families, task streams, or deployment distributions.

Finite-library multiplicity. The two 336-anchor Qwen finite-library panels each test the same 22 prespecified current-information comparator components. Every component must have a positive point margin, one-sided t lower bound, and exact-anchor bootstrap lower bound. In addition, a 200,000-draw max-t procedure supplies simultaneous

95% lower bounds across all 22 components. The two panels are independent replications and are never pooled for this conjunction.

Multiplicity scope. Multiplicity is controlled within each prespecified family to which a conjunction or model-selection claim is attached: the 22-component current-information library and, separately, the two-architecture Mistral adaptive family. The manuscript does not assert a single paper-wide familywise-error guarantee across heterogeneous theory-targeted estimands, which answer distinct scientific questions and are reported with their own declared inferential units and lower bounds.

Equal-compute decomposition. The 432-anchor common-restart comparison and each 336-anchor productive-path panel are independent. Equation (4.6) is therefore evaluated once with productive-path panel I and once with panel II. Standard errors use the independence-aware Welch–Satterthwaite construction; the two productive-path panels are not pooled.

Mistral two-bank evaluation. The fixed-depth and equal-compute Mistral estimands average banks A and B within each of 480 exact anchors before t or bootstrap inference. The adaptive target analysis likewise forms $Z_i^{AB} = (Z_{i,A} + Z_{i,B})/2$ before inference.

Finite architecture family for Mistral adaptive control. Scalar and two-coordinate continuation architectures were both specified in the Qwen adaptive-control analysis before the Mistral target analysis. Their Mistral parameters were fit using only the separate 336-anchor development panel. To protect the target-panel claim against selecting between these two architectures, they are treated as a two-element finite family. The scalar target-panel point estimate is

$$\widehat{\Delta V}^{\mathrm{M}}_{\mathrm{scalar}} = 1.13877630257 \times 10^{-5}. \tag{E.2}$$

Its ordinary one-sided 95% t and bootstrap lower bounds are $2.37141363069 \times 10^{-6}$ and $5.41900958243 \times 10^{-6}$. Bonferroni familywise 95% bounds use one-sided level $1 - 0.05/2 = 0.975$ for each architecture and remain positive for the scalar rule:

$$\mathrm{LCB}_t^{\mathrm{FWER}} = 6.37735915928 \times 10^{-7}, \qquad \mathrm{LCB}_{\mathrm{boot}}^{\mathrm{FWER}} = 5.23837592968 \times 10^{-6}. \tag{E.3}$$

Later exploratory revealability proxies are not used to support this target-panel result.

Bounded stop–flip risk. For a fixed stop rule, the joint event $\{S = 1, F = 1\}$ is Bernoulli. Reported risk certificates are one-sided Bernoulli–KL upper confidence bounds computed bankwise under the prespecified calibration role. These bounds concern the probability of changing the deeper action; Theorem B.3 explains why they do not by themselves bound the utility severity of those changes.

F Experimental protocol details

Qwen comparative panel. The Qwen comparative panel contains 432 exact anchors disjoint from development and from the two finite-library panels, with three prespecified task-stream strata of 144 anchors each and nine-position balance. The active policy, strengthened DynaMiCS-style frontier comparator, 20-update policy-visible compute contract, and ambiguity threshold

$$\tau_F = 0.00052099142651661841 \tag{F.1}$$

were fixed before terminal evaluation. Each anchor contains both Keep and Xi H12 outcomes in two conditionally independent terminal banks, T_1 and T_2.

Mistral development and target panels. Mistral-specific numerical heads are fit on a separate 336-anchor development panel. The independent 480-anchor target panel uses model revision `c03fc1dabc3d31b96271626f15a76a6779fb4037` of `mistralai/Mistral-7B-v0.3`, two independent future banks, the same Keep/Xi action menu, H4/H8/H12 geometry, and 20-update policy-visible compute contract. The H4 ambiguity threshold 0.00236920914414 is estimated from the development panel and fixed for target evaluation.

References

[1] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

[2] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi:10.1214/24-AOAS1998.

[3] R. R. Bahadur and Leonard J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956. doi:10.1214/aoms/1177728077.

[4] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 2 edition, 1985. doi:10.1007/978-1-4757-4286-2.

[5] David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi:10.1214/aoms/1177729032.

[6] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi:10.1214/ss/1177009939.

[7] Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 3460–3468, 2015.

[8] Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems 30*, 2017.

[9] Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022. doi:10.1287/mnsc.2020.3922.

[10] Kazuo J. Ezawa. Evidence propagation and value of evidence on influence diagrams. *Operations Research*, 46(1):73–83, 1998. doi:10.1287/opre.46.1.73.

[11] Alexander A. Feldbaum. Dual control theory. i. *Automation and Remote Control*, 21:874–880, 1960.

[12] Vojtech Franc and Daniel Prusa. On discriminative learning of prediction uncertainty. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1963–1971, 2019.

[13] Peter I. Frazier, Warren B. Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008. doi:10.1137/070693424.

[14] Eleonora Gualdoni, Sonia Laguna, Louis Bethune, João Monteiro, Pierre Ablin, and Marco Cuturi. DynaMiCS: Fine-tuning LLMs with performance constraints using dynamic mixtures. *arXiv preprint arXiv:2605.10770*, 2026. doi:10.48550/arXiv.2605.10770.

[15] Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi:10.1109/TSSC.1966.300074.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. doi:10.48550/arXiv.2106.09685.

[17] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017. doi:10.48550/arXiv.1711.09846.

[18] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 240–248. PMLR, 2016.

[19] Metod Jazbec, Alexander Timans, Tin Hadži Veljković, Kaspar Sakmann, Dan Zhang, Christian A. Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[20] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. doi:10.48550/arXiv.2310.06825.

[21] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi:10.1016/S0004-3702(98)00023-X.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. doi:10.48550/arXiv.1412.6980.

[23] Arthur Lebeurrier, Titouan Vayer, and Rémi Gribonval. Path-conditioned training: A principled way to rescale ReLU neural networks. *arXiv preprint arXiv:2602.19799*, 2026. doi:10.48550/arXiv.2602.19799.

[24] Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*, 2006.

[25] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

[26] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi:10.1214/aoms/1177728069.

[27] Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems 14*, pages 1555–1561, 2001.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. doi:10.48550/arXiv.1711.05101.

[29] Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024. doi:10.1613/JAIR.1.15320.

[30] Mistral AI. Mistral-7B-v0.3 model card. Hugging Face model repository, 2024. URL https://huggingface.co/mistralai/Mistral-7B-v0.3. Accessed 2026-08-24.
[31] Roger B. Myerson. Incentive compatibility and the bargaining problem. *Econometrica*, 47(1):61–73, 1979. doi:10.2307/1912346.
[32] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi:10.48550/arXiv.2412.15115.
[33] Qwen Team. Qwen2.5-7B model card. Hugging Face model repository, 2024. URL https://huggingface.co/Qwen/Qwen2.5-7B. Accessed 2026-08-24.
[34] Naveen Raman, Stephanie Milani, and Fei Fang. Selecting decision-relevant concepts in reinforcement learning. *arXiv preprint arXiv:2604.04808*, 2026. doi:10.48550/arXiv.2604.04808.
[35] Liran Ringel, Regev Cohen, Daniel Freedman, Michael Elad, and Yaniv Romano. Early time classification with accumulated accuracy gap control. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42584–42600, 2024.
[36] Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1–3):361–395, 1991. doi:10.1016/0004-3702(91)90015-C.
[37] Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze–thaw bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014. doi:10.48550/arXiv.1406.3896.
[38] Michał Tomaszewski. Can neural networks learn by experimenting on themselves? self-interventional learning from functional consequences to predictive self-knowledge. *arXiv preprint arXiv:2608.14894*, 2026. doi:10.48550/arXiv.2608.14894.
[39] Tiago Veiga and Jennifer Renoux. From reactive to active sensing: A survey on information gathering in decision-theoretic planning. *ACM Computing Surveys*, 55(13s):1–22, 2023. doi:10.1145/3583068.
[40] Qinyou Wang. Fiber fingerprints of hidden learning-state dynamics. *arXiv preprint arXiv:2608.15976*, 2026. doi:10.48550/arXiv.2608.15976.
[41] Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Benjamin Van Durme, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. Conformal thinking: Risk control for reasoning on a compute budget. *arXiv preprint arXiv:2602.03814*, 2026. doi:10.48550/arXiv.2602.03814.
[42] Jingbo Wen, Liang He, and Ziqi He. Not all errors are equal: Consequence-aware reasoning compute allocation. *arXiv preprint arXiv:2606.04402*, 2026. doi:10.48550/arXiv.2606.04402.
[43] Siye Wu, Jian Xie, Yikai Zhang, and Yanghua Xiao. CODA: Difficulty-aware compute allocation for adaptive reasoning. *arXiv preprint arXiv:2603.08659*, 2026. doi:10.48550/arXiv.2603.08659.