

Gaussian Process Bandits for AI System Optimization with Selectively Verified Feedback

Shuang Wu
UCLA
shuangwu222@ucla.edu

Abstract

Gaussian process bandits exploit similarities among configurations to optimize prompts, retrieval systems, and agent workflows. In these applications, however, inexpensive model-based evaluations can disagree with trusted task outcomes. We study selective verification when an action is chosen before a diagnostic is observed and trusted feedback is informative on an unknown set of diagnostic states. A finite reliability class and a finite RKHS signal dictionary describe these two sources of uncertainty. We construct a fixed audit design that combines reliability coverage with reward separation and establish a finite-time regret bound that accounts for a hard audit budget and the rounds needed to obtain informative observations. A safe adaptive extension mixes this robust design with predictable Chernoff allocations while preserving a uniform identification floor and exact expected acquisition-cost accounting. For signed-coordinate kernels, structured and diffuse reliability classes require different audit scales despite sharing the same reward geometry. Synthetic experiments distinguish these certified designs from an adaptive posterior heuristic and GP controls. Safe adaptive allocation reduces aggregate regret relative to the fixed design under the exact dictionary, but its finite-budget error certificate is numerically loose. Conditional-feedback errors can remove the heuristic’s advantage and target perturbations can reverse it. Separate AI-task replays assess feedback correction and fitted-dictionary policies without establishing end-to-end validation of a learned selective-verification method.

1 Introduction

Language-model judges make repeated evaluation inexpensive, but they can reverse the ranking of prompts, retrieval pipelines, or agent configurations. A Gaussian process (GP) fitted to such scores can estimate the wrong objective precisely. Trusted evaluation—human adjudication, execution in a protected environment, or a high-quality test suite—repairs the target but can dominate total cost. This motivates a bandit that sees cheap post-action information and selectively purchases the trusted outcome.

Existing methods cover selective audits, prediction-powered inference, auxiliary-feedback kernel bandits, and post-surrogate acquisition [23, 29, 30, 42]. We ask how many trusted outcomes are required when the learner must discover *where* feedback is informative. For unrestricted outcome laws, target-kernel rank and a fixed finite collection of conditional statistics do not determine that information.

We introduce a finite conditional-reliability class. A known dictionary $\{C_h\}$ contains possible informative diagnostic sets, but h is unknown. Its fractional hitting number τ^* measures the audit mass needed to touch every candidate reliable set. We couple this uncertainty to a finite RKHS signal dictionary through a max–min experimental design. The resulting rate depends on reliability coverage and on the action energy needed to distinguish reward models with different optima.

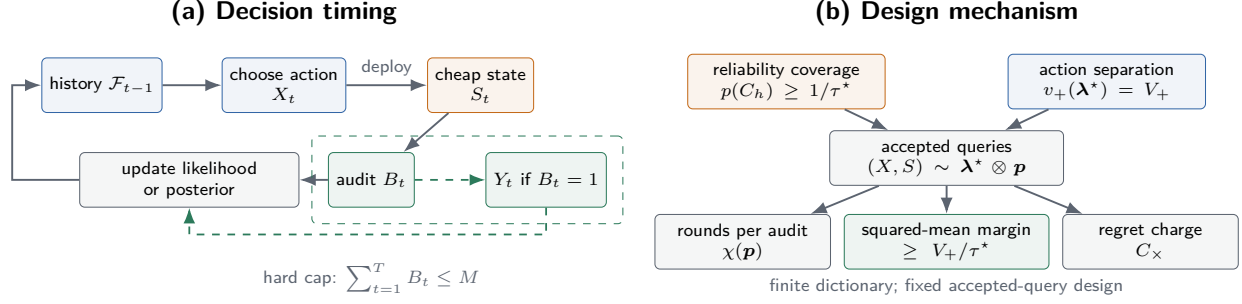


Figure 1: Selective-verification protocol. (a) Given the observed history $\mathcal{F}_{t-1}$, actions precede diagnostics, which inform the audit decision. Dashed arrows show when a purchased outcome enters the update. (b) Reliability coverage and action separation bound the squared-mean margin per audit. The factors $\chi(\mathbf{p})$ and $C_\times$ charge acquisition time and regret. Proposition 4 adds predictable adaptive allocation.

For example, RAG configurations differ in their prompts and retrieval pipelines. A cheap diagnostic can reveal request or output categories where trusted correctness distinguishes configurations. A selective policy must discover these categories and acquire informative observations while paying for every intervening deployment. This separates reward geometry, conditional reliability, and acquisition time.

Our contributions are:

- We prove a finite-time bound with audit scale $\tau^* \mathfrak{D}_+$ and characterize the exact expected acquisition cost of a joint action–state design. A safe adaptive mixture preserves a uniform Hellinger-information floor under a pathwise audit cap. A prior-mixture converse permits fully adaptive auditing and infinite signal classes.
- For signed rank- d signals, we establish matching polynomial audit scales of $d\tau^*/\Delta^2$ for structured reliability and $d(\tau^*)^2/\Delta^2$ for a diffuse reliability class. The deterministic structured model admits an exact asymptotic regret coefficient expressed by a coverage problem.
- We distinguish feedback-law error from target approximation. Two rank-one dictionaries match any prescribed finite collection of bounded feedback summaries and have identical targets, yet their regret differs by a factor growing as $K/\log K$ at a common $\log_2 K$ audit budget. Sequential audit-only controls provide reliability-free comparisons.
- We evaluate the proved designs, an adaptive-action posterior heuristic, and GP controls across kernel geometries, overlapping reliability sets, rare states, and model perturbations. AI-task replays examine calibration and acquisition costs without establishing end-to-end validation of a learned reliability model.

2 Related work

Selective verification and surrogate feedback. Prediction-powered best-arm identification and adaptive inference combine inexpensive predictions with selectively acquired trusted labels [29, 42]. Limited-verification bandits and surrogate-reward bandits study trusted feedback under attacks or machine-learning-generated rewards [19, 37]. ASB permits acquisition after observing a surrogate and accounts for learned refinements and hard budgets [30]. Confidence–budget matching

and costly-observation bandits provide reward-only controls [15, 44]. Our question concerns the additional cost of identifying an unknown reliable-state set while optimizing cumulative true reward.

Kernel bandits and AI evaluation. Auxiliary-feedback IGP-UCB, prediction-augmented GP-UCB, and robust multi-fidelity optimization exploit additional feedback under structural assumptions [22, 23, 39]. Kernel experimental design and misspecified GP bandits characterize reward geometry and target approximation [13, 14]. Related exploration methods include residual-bootstrap LinReBoot, graph-neural GNN-TS, and Laplacian kernelized bandits [17, 27, 36]. ARES and HbBoPs connect trusted evaluation and kernel optimization to RAG and prompt selection [25, 28]. We study how conditional reliability interacts with kernel decision geometry, rather than a new kernel posterior or a generic correction estimator.

Controlled sensing and decision-focused learning. Controlled sensing, decision-region active learning, and transductive design allocate measurements to distinguish consequential alternatives [1, 2, 4, 7, 18]. Best-arm identification with post-action context provides context-tracking and Track-and-Stop rules [41]. These ideas motivate our likelihood decoder and adaptive allocation. Our analysis separates reliability coverage, action separation, and the deployment cost imposed by choosing an action before observing the diagnostic.

Partial monitoring and identifiability. Kernelized partial monitoring, conditional mean embeddings, and cost-aware streaming labels provide broader sequential abstractions [11, 12, 20, 21]. Exact rejection sampling implements the desired accepted-query law [26]. Classical best-arm and mixture testing explain the polynomial rates in our structured-diffuse comparison [3, 8, 32], while contaminated and misspecified identification establish related identifiability failures [6, 16]. Our results specialize these tools to a finite reliability-signal dictionary, with an adaptive converse and matched-summary counterexamples. Appendix D compares the assumptions and guarantees in detail.

3 Problem formulation

3.1 AI evaluation protocol and objective

Each action is a fixed prompt, retrieval pipeline, or agent configuration used on a fresh request. A cheap diagnostic is observed after deployment, and a trusted assessment is available only if purchased. We optimize expected trusted quality under a stationary action set and kernel. Request-dependent routing and within-agent-step optimization are outside this model.

Let $\mathcal{A}$ and $\mathcal{S}$ be finite, with integer horizon $T \geq 1$ and audit cap $M \geq 1$. At round t , the learner chooses $X_t \in \mathcal{A}$, observes $S_t \in \mathcal{S}$, and then chooses $B_t \in \{0, 1\}$. If $B_t = 1$, it spends one audit and observes $Y_t \in \{0, 1\}$. The policy satisfies $\sum_{t=1}^T B_t \leq M$ almost surely. The audit decision may use S_t , so the lower bounds permit post-state acquisition.

3.2 Conditional reliability and kernel signals

Write $\Delta(\mathcal{X})$ for the probability simplex on a finite set $\mathcal{X}$ and $[x]_+ = \max\{x, 0\}$. States are i.i.d. from a known action-independent law μ , with $\mu_{\min} = \min_s \mu(s) > 0$. A known finite family $\{C_h : h \in \mathcal{H}\}$ has $\mu(C_h) = \rho > 0$. Its fractional hitting number is

$$\tau^* = \min_{x \geq 0} \left\{ \sum_{s \in \mathcal{S}} x_s : \sum_{s \in C_h} x_s \geq 1 \ \forall h \right\}. \quad (1)$$

For an optimizer, $p_s = x_s/\tau^*$ obeys $p(C_h) \geq 1/\tau^*$.

Let k be a known positive-semidefinite kernel on $\mathcal{A}$, with RKHS $\mathcal{H}_k$. A finite signal dictionary $\mathcal{U} = \{u_z : z \in \mathcal{Z}\} \subset \mathcal{H}_k$ satisfies

$$\|u_z\|_{\mathcal{H}_k} \leq B, \quad \|u_z\|_{\infty} \leq U \leq \frac{1}{4}, \quad a_z^* \in \arg \max_{a \in \mathcal{A}} u_z(a).$$

Under environment (h, z) , audited outcomes are conditionally independent with

$$Y_t \mid (S_t = s, X_t = a) \sim \text{Bern}\left(\frac{1}{2} + \mathbf{1}\{s \in C_h\}u_z(a)\right). \quad (2)$$

The target is $f_z(a) = \frac{1}{2} + \rho u_z(a)$ and

$$R_T = \sum_{t=1}^T [f_z(a_z^*) - f_z(X_t)], \quad R_T^* = \inf_{\text{causal policies}} \sup_{h,z} \mathbb{E}_{h,z}[R_T].$$

The known centering constant is $1/2$, and the kernel models the action-dependent target component. Trusted outcomes are informative about the best action within C_h and uninformative outside it. The diagnostic S_t need not be a reward prediction, and the model does not assume an arbitrary judge is unbiased.

Fix one maximizing action a_z^* for each signal, including a deterministic tie rule. The learner knows the dictionaries but neither h nor z . Equal reliable-set mass and action-independent μ make the target depend only on z , while h controls information acquisition. This factorization need not hold for action-dependent diagnostics. Appendix B.3 extends only the acquisition rule to a known action-dependent law. Neither result learns or validates the outcome dictionary.

3.3 Decision geometry and scope

For $\lambda \in \Delta(\mathcal{A})$, define the action-separation margin

$$v_+(\lambda) = \min \left\{ \min_z \mathbb{E}_{a \sim \lambda}[u_z(a)^2], \min_{a_z^* \neq a_{z'}^*} \mathbb{E}_{a \sim \lambda}[(u_z(a) - u_{z'}(a))^2] \right\}, \quad V_+ = \max_{\lambda} v_+(\lambda), \quad \mathfrak{D}_+ = V_+^{-1}. \quad (3)$$

Writing $u_z(a) = \langle \phi(a), \vartheta_z \rangle$, every constraint is linear in λ through $V(\lambda) = \sum_a \lambda_a \phi(a) \phi(a)^\top$. Hence V_+ is a pre-outcome linear program, not an oracle posterior quantity. We assume $V_+ > 0$ and at least two different designated optima. The features $\phi(a)$ and coefficients ϑ_z are vectors in a feature representation of k .

4 Selective Verification with Kernel Decision Geometry

4.1 Reliability coverage and kernel action selection

Two pre-outcome linear programs determine the policy. State allocation covers every candidate reliable set. Action allocation separates weak individual signals and reward pairs with different optima. Both use the known dictionaries, without knowledge of the true environment.

Let λ^* optimize (3). For an optimizer $\mathbf{p}$ of (1), define its action-before-state realization factor

$$\chi(\mathbf{p}) = \max_{s \in \mathcal{S}} \frac{p_s}{\mu(s)}.$$

Let $m_{h,z}(a, s) = \frac{1}{2} + \mathbf{1}\{s \in C_h\}u_z(a)$ and write $p_{h,z}(y \mid a, s) = m_{h,z}(a, s)^y [1 - m_{h,z}(a, s)]^{1-y}$. Algorithm 1 compiles the product law $\lambda^* \otimes \mathbf{p}$ by rejection sampling after the state arrives. The audit coin is drawn only after observing the state, independently of the unobserved outcome conditional

Algorithm 1 Reliability-Kernel-Design

1. **Input:** $T, M, \mu, \{C_h\}, \{u_z\}$. Solve (1) and (3) for $\mathbf{p}, \boldsymbol{\lambda}^*$.
 2. **Initialize:** $n \leftarrow 0, \ell_{h,z} \leftarrow 0$ for every (h, z) .
 3. **For** $t = 1, \dots, T$ **do**
 4. **If** $n < M$ **then** draw $X_t \sim \boldsymbol{\lambda}^*$ and observe $S_t = s$.
 5. Draw $B_t \sim \text{Bern}(p_s / [\chi(\mathbf{p})\mu(s)])$.
 6. **If** $B_t = 1$ **then** purchase Y_t and set

$$n \leftarrow n + 1, \quad \ell_{h,z} \leftarrow \ell_{h,z} + \log p_{h,z}(Y_t \mid X_t, s) \quad \text{for all } (h, z).$$
 7. **If** $n = M$ **then** set $(\hat{h}, \hat{z}) \in \arg \max_{h,z} \ell_{h,z}$.
 8. **Else** play $X_t = a_{\hat{z}}^*$, observe S_t , and set $B_t = 0$.
 9. **End if.**
 10. **End for.**
 11. **Return:** $a_{\hat{z}}^*$ if $n = M$ and “insufficient audits” otherwise.
-

on the action, state, and past. For $\Lambda \geq 1$, define the exact expected number of design rounds before M accepts or horizon end,

$$N_{T,M}(\Lambda) = \sum_{t=0}^{T-1} \Pr\{\text{Bin}(t, \Lambda^{-1}) < M\}.$$

Proposition 1 (Product separation). *For any (h, z) and (h', z') with $a_z^* \neq a_{z'}^*$,*

$$\mathbb{E}_{s \sim \mathbf{p}, a \sim \boldsymbol{\lambda}^*} [(m_{h,z}(a, s) - m_{h',z'}(a, s))^2] \geq \frac{V_+}{\tau^*},$$

where $m_{h,z}$ is the Bernoulli mean in (2).

The proof partitions the two reliable sets into their overlap and two differences. See Appendix A.1.

The product value is also controlled from above for every fixed accepted-query design. Define

$$V_\Delta = \max_{\lambda \in \Delta(\mathcal{A})} \min_{a_z^* \neq a_{z'}^*} \mathbb{E}_{a \sim \lambda} [(u_z(a) - u_{z'}(a))^2],$$

$$J^* = \max_{\nu \in \Delta(\mathcal{A} \times \mathcal{S})} \min_{a_z^* \neq a_{z'}^*} \mathbb{E}_{(a,s) \sim \nu} [(m_{h,z}(a, s) - m_{h',z'}(a, s))^2],$$

where the second minimum ranges over all h, h', z, z' with different designated optima.

Proposition 2 (Fixed-design sandwich). *The exact squared-mean fixed-design value obeys*

$$\frac{V_+}{\tau^*} \leq J^* \leq \frac{V_\Delta}{\tau^*}.$$

Consequently the product design is exact whenever its individual-energy constraints are nonbinding, $V_+ = V_\Delta$.

A product dual certificate gives the upper bound (Appendix A.2).

The individual energy constraints in (3) are essential. Pairwise action separation alone can be hidden outside the overlap of two reliability sets. The method remains finite-dictionary and exact-kernel:

$\mathcal{H}$, $\mathcal{U}$, and both designs must be frozen before outcomes are opened. Maximum-likelihood decoding uses the margin V_+/τ^* within controlled testing [1, 7, 18]. Adaptive sensing can outperform fixed designs in general multihypothesis problems [1], so this comparison does not establish adaptive optimality.

4.2 Finite-time regret with a hard audit budget

The pathwise audit cap and expected-regret bound account for all deployed actions. Independence of accepted samples and stationary interarrival counts gives exact finite-horizon accounting. Full proofs appear in Appendix A.

Theorem 1 (Finite-dictionary RKHS achievable rate). *Under (2) with $V_+ > 0$, Algorithm 1 uses at most M audits and satisfies*

$$\mathbb{E}[R_T] \leq \frac{C_\times}{\chi(\mathbf{p})} \mathbf{N}_{T,M}(\chi(\mathbf{p})) + 2\rho U [T - \mathbf{N}_{T,M}(\chi(\mathbf{p}))] |\mathcal{H}| |\mathcal{Z}| \exp\left(-\frac{M}{2\tau^* \mathfrak{D}_+}\right).$$

Here

$$C_\times = \rho \chi(\mathbf{p}) \max_z \mathbb{E}_{a \sim \lambda^*} [u_z(a_z^*) - u_z(a)] \leq 2\rho U \chi(\mathbf{p}).$$

Since $\mathbf{N}_{T,M}(\Lambda) \leq \min\{T, M\Lambda\}$, this refines the simpler untruncated charge whenever the horizon binds. Thus $M \geq 2\tau^* \mathfrak{D}_+ \log(2|\mathcal{H}| |\mathcal{Z}| T)$ makes the exploitation-error term at most ρU .

Appendix A.3 proves the result by factoring accepted triples from their geometric waiting times and applying a Hellinger likelihood bound.

4.3 Joint action–verification policies and acquisition costs

The audit scale $\tau^* \mathfrak{D}_+$ measures information per accepted query. In contrast, $\chi(\mathbf{p})$ controls how quickly those queries arrive, and $C_\times/\chi(\mathbf{p})$ is the worst expected regret per design round. Rare informative states can therefore make an audit-efficient design costly in calendar time. If $\mathbf{p} = \mu$, every round is accepted and $\chi(\mathbf{p}) = 1$.

The stopped bound charges exploitation only after all M observations arrive. For fixed M and a long horizon, the design charge approaches $MC_\times$.

The product construction is interpretable but can be conservative. To state the exact fixed-design cost, let $\mathcal{E}$ be a finite environment dictionary with at least two designated optima. Under $e \in \mathcal{E}$, fresh outcomes at (a, s) have law $P_e^{a,s}$, independently across rounds conditional on actions and states. Define $f_e(a) = \mathbb{E}_{S \sim \mu} [\mathbb{E}_e[Y \mid a, S]]$, choose $a_e^* \in \arg \max_a f_e(a)$, and put $\Delta_e(a) = f_e(a_e^*) - f_e(a)$. Put $G = \max_{e,a} \Delta_e(a)$. For $\nu \in \Delta(\mathcal{A} \times \mathcal{S})$ define

$$\begin{aligned} \Lambda(\nu) &= \sum_{a \in \mathcal{A}} \max_{s \in \mathcal{S}} \frac{\nu(a, s)}{\mu(s)}, & C(\nu) &= \max_e \sum_{a \in \mathcal{A}} \Delta_e(a) \max_s \frac{\nu(a, s)}{\mu(s)}, \\ I(\nu) &= \min_{e, e': a_e^* \neq a_{e'}^*} \sum_{a, s} \nu(a, s) H^2(P_e^{a,s}, P_{e'}^{a,s}), \end{aligned}$$

where $H^2(P, Q) = 1 - \int \sqrt{dP, dQ}$.

Proposition 3 (Action-before-state rejection-sampling identity). *There is a stationary action and post-state audit rule whose accepted pairs are i.i.d. ν and whose acceptance probability per round is $1/\Lambda(\nu)$. Every causal rule producing one accepted pair with law ν has expected waiting time at least $\Lambda(\nu)$ and sampling regret at least $C_e(\nu) := \sum_a \Delta_e(a) \max_s \nu(a, s)/\mu(s)$ under e ; the stationary rule attains both. After the M th audit, choose a maximum-likelihood environment and play its designated optimum without further audits. If the horizon arrives first, remain in the design phase. Then*

$$\sup_{e \in \mathcal{E}} \mathbb{E}_e[R_T] \leq \frac{C(\nu)}{\Lambda(\nu)} \mathbf{N}_{T,M}(\Lambda(\nu)) + G[T - \mathbf{N}_{T,M}(\Lambda(\nu))](|\mathcal{E}| - 1)e^{-MI(\nu)}.$$

The bound therefore holds after infimizing its right-hand side over ν .

Appendix A.4 proves both the causal cost identity and the finite-horizon decoding bound.

4.4 Adaptive verification with an information guarantee

The fixed design can be adapted without losing worst-case identifiability. Given a fixed ν_0 with $I_0 := I(\nu_0) > 0$, fix $0 < \varepsilon \leq 1$ and mix any predictable design as $\nu_m = \varepsilon \nu_0 + (1 - \varepsilon) \tilde{\nu}_m$. Our implementation uses the Chernoff max-min law against different-optimum alternatives to the current maximum-likelihood environment.

Proposition 4 (Safe adaptive joint design). *Assume an exact finite observation dictionary, a known action-independent state law, and the same predictable rule under every environment. At stage m , hold ν_m fixed until the next accepted audit and apply the row-max compiler of Proposition 3. After M audits, exploit the designated optimum of a maximum-likelihood environment. Define the stage sequence by continuing the same rule beyond T if necessary. Then at most M audits are used pathwise and, for every $e \in \mathcal{E}$,*

$$\mathbb{E}_e[R_T] \leq \mathbb{E}_e \left[\sum_{m=1}^M C_e(\nu_m) \right] + GT \min \left\{ 1, (|\mathcal{E}| - 1)e^{-\varepsilon MI_0} \right\}.$$

The conditional Hellinger information is at least εI_0 for every different-optimum pair, and the process extended until M audits arrive has $\mathbb{E}_e[\tau_M] = \mathbb{E}_e[\sum_{m=1}^M \Lambda(\nu_m)]$.

Appendix A.5 combines accepted-history contraction with the horizon coupling. The result composes controlled sensing [1, 18] and density-ratio rejection sampling [26] for this protocol. It requires an exact dictionary and differs from ASB’s inverse-propensity variance objective [30].

Corollary 1 (Independent calibration transfer). *Let independent calibration data $\mathcal{C}$ return a finite dictionary $\hat{\mathcal{E}}(\mathcal{C})$ containing the exact true environment with probability at least $1 - \delta$. Let G_{all} bound every true one-round gap, including when calibration fails ($G_{\text{all}} = 1$ is valid for rewards in $[0, 1]$). If the instantiated policy has a nonnegative conditional bound $B_T(\mathcal{C})$ on the coverage event, then $\mathbb{E}[R_T] \leq \mathbb{E}[B_T(\mathcal{C})] + \delta G_{\text{all}} T$. Calibration with n_{cal} trusted labels gives total trusted-label use at most $n_{\text{cal}} + M$.*

Appendix A.6 proves this elementary conditioning transfer. Exact coverage is essential; continuously refit signals, misspecification, and population shift remain outside its scope.

5 Adaptive limits and reliability sensitivity

5.1 An adaptive converse and a matched coordinate family

The converse permits actions based on the entire history and audits chosen after the current state. It therefore covers adaptive policies, rather than only the fixed designs analyzed above.

For a converse beyond finite signal dictionaries, let Π_0 contain every probability law π on a measurable collection $\mathcal{U}_0 \subset \mathcal{H}_k$ with $\|u\|_{\mathcal{H}_k} \leq B$ and $\|u\|_\infty \leq U \leq 1/4$ and

$$\mathbb{E}_{u \sim \pi}[u(a)] = 0 \quad \forall a \in \mathcal{A}. \quad (4)$$

For $\pi \in \Pi_0$, put

$$G_\pi = \mathbb{E}_{u \sim \pi}[\max_a u(a)], \quad W_\pi = \max_a \mathbb{E}_{u \sim \pi}[u(a)^2].$$

Theorem 2 (Prior-mixture adaptive lower bound). *For every $\pi \in \Pi_0$, the minimax regret over $\mathcal{H} \times \mathcal{U}_0$ satisfies*

$$R_T^* \geq \rho T \left[G_\pi - 4U \sqrt{\frac{MW_\pi}{3\tau^*}} \right]_+.$$

When $W_\pi > 0$, $M \leq 3\tau^* G_\pi^2 / (64U^2 W_\pi)$ implies $R_T^* \geq \rho G_\pi T / 2$.

The proof compares every environment with a fair-outcome null law and averages the resulting KL divergence under a zero-barycenter prior. See Appendix A.7.

The prior may have infinite support in the RKHS. The bound respects the hard audit cap and counts every action used while waiting for an informative state. Its applicability depends on the existence of a zero-barycenter prior.

For the finite balanced signed specialization, let $\{g_j : j \in \mathcal{J}\} \subset \mathcal{H}_k$ satisfy $\|g_j\|_{\mathcal{H}_k} \leq B$ and include both signs $u_{j,\theta} = \theta g_j$. Assume $\max_a \theta g_j(a) \geq g_{\min} > 0$ for every (j, θ) and $\max_j \|g_j\|_\infty = U \leq 1/4$. Define

$$V_- = \max_{\lambda \in \Delta(\mathcal{A})} \min_j \mathbb{E}_{a \sim \lambda}[g_j(a)^2]. \quad (5)$$

LP duality supplies a law on j with $\max_a \mathbb{E}[g_j(a)^2] = V_-$. Combining it with an independent uniform sign gives a member of Π_0 with $G_\pi \geq g_{\min}$ and $W_\pi = V_-$. Therefore Theorem 2 yields

$$R_T^* \geq \rho T \left[g_{\min} - 4U \sqrt{\frac{MV_-}{3\tau^*}} \right]_+.$$

In particular, $M \leq 3\tau^* g_{\min}^2 / (64U^2 V_-)$ implies $R_T^* \geq \rho g_{\min} T / 2$.

The coordinate family makes the two design quantities coincide. For $d \geq 1$ and $0 < \Delta \leq \min\{B, 1/4\}$, let $\mathcal{A} = [d] \times \{\pm 1\}$,

$$k((i, a), (j, b)) = \mathbf{1}\{i = j\}ab, \quad u_{J,\theta}(i, a) = \Delta \theta a \mathbf{1}\{i = J\}. \quad (6)$$

Then $V_+ = V_- = \Delta^2/d$, $\mathfrak{D}_+ = d/\Delta^2$, and $g_{\min} = U = \Delta$. For $d \geq 2$, $V_\Delta = 2\Delta^2/d$, so Proposition 2 also makes the product design two-approximate for the exact squared-mean fixed-design game (the rank-one ratio is four). Appendix A.8 derives these identities. Theorems 1–2 match at the polynomial audit scale $d\tau^*/\Delta^2$, up to logarithmic factors.

5.2 Structured and diffuse reliability

The coordinate example separates reward-direction uncertainty from reliability uncertainty. Knowing the active coordinate reduces $\mathfrak{D}_+$ from d/Δ^2 to $1/\Delta^2$ without changing the kernel. Kernel rank alone therefore does not specify the decision complexity.

The finite-dictionary assumption is consequential. Let $\mathcal{S} = [L]$ be uniform, let $1 \leq k \leq L/2$ be an integer and $\rho = k/L$, and take as reliability class every k -subset $C \subseteq [L]$. This class has $\tau^* = L/k = 1/\rho$.

The structured and diffuse families share the target kernel, rank, signal family, and hitting number, but their polynomial audit scales differ by τ^* . Both satisfy (2) exactly. In the diffuse family, state labels rarely repeat when $L \geq 8T^2$, so their identities provide little reusable information about reliability. This regime is essential to the lower bound. The next theorem shows that selective audits cannot improve its polynomial rate.

Theorem 3 (Diffuse-reliability frontier). *Use the signed coordinate kernel and signals in (6). If $M \leq T$, a balanced pooled policy satisfies*

$$\mathbb{E}[R_T] \leq 2\rho\Delta M + 4\rho\Delta Td \exp\left\{-\frac{\lfloor M/d \rfloor \rho^2 \Delta^2}{2}\right\}.$$

Conversely, if $d \geq 32$, $L \geq 8T^2$, and

$$M \leq \frac{3d}{4096\rho^2\Delta^2},$$

every adaptive action-before-state policy with post-state audits obeys $\sup_{C,J,\theta} \mathbb{E}[R_T] \geq \rho\Delta T/4$. Hence this family has audit scale $d(\tau^*)^2/\Delta^2$ up to constants and logarithms.

Appendix A.9 proves the upper bound by balanced audit-only sampling and the lower bound by an exchangeable unseen-state mixture.

5.3 Robustness and limits of summary statistics

Keep the independent Bernoulli protocol, with nominal and true means $m_e(a, s)$ and $\tilde{m}_e(a, s)$. The true target is $f_e(a) = \mathbb{E}_{S \sim \mu}[\tilde{m}_e(a, S)]$, and $\tilde{R}_T$ denotes regret against its maximizing action. Define

$$D(\boldsymbol{\nu}) = \min_{a_e^* \neq a_{e'}} \left\{ \sum_{a,s} \nu(a, s) [m_e(a, s) - m_{e'}(a, s)]^2 \right\}^{1/2}.$$

Proposition 5 (Feedback-law misspecification boundary). *Assume*

$$\sup_{e,a,s} |\tilde{m}_e(a, s) - m_e(a, s)| \leq \delta_{\text{fb}}, \quad \sup_{e,a} |\tilde{f}_e(a) - f_e(a)| \leq \varepsilon_{\text{tar}}.$$

Replace maximum likelihood in Proposition 3 by the nominal squared-loss decoder. Then

$$\sup_e \mathbb{E}_e[\tilde{R}_T] \leq \frac{C(\boldsymbol{\nu})}{\Lambda(\boldsymbol{\nu})} \mathbf{N}_{T,M}(\Lambda(\boldsymbol{\nu})) + G[T - \mathbf{N}_{T,M}(\Lambda(\boldsymbol{\nu}))](|\mathcal{E}| - 1) \exp\left\{-\frac{3M}{40}[D(\boldsymbol{\nu}) - 2\delta_{\text{fb}}]_+^2\right\} + 2\varepsilon_{\text{tar}}T.$$

Conversely, for every $0 < \rho \leq 1/2$ and $0 < \Delta \leq 1/4$, there is a two-action rank-one instance with $\varepsilon_{\text{tar}} = 0$ and $\delta_{\text{fb}} = (1 - \rho)\Delta$ such that every causal policy obeys

$$\sup_e \mathbb{E}_e[R_T] \geq \frac{\rho\Delta T}{2} e^{-M\kappa_\rho\Delta}, \quad \kappa_x = 2x \log \frac{1+2x}{1-2x} \leq 16x^2 \quad (0 \leq x \leq 1/4).$$

Appendix A.10 gives the squared-loss concentration argument and the target-preserving two-point construction.

Thus target error contributes an additive comparison term, while feedback-law error erodes the testing margin and can destroy identifiability. Related identifiability failures arise in contaminated best-arm identification and misspecified linear identification [6, 16]. For the product design, Proposition 1 gives $D(\boldsymbol{\lambda}^* \otimes \mathbf{p}) \geq \sqrt{V_+/\tau^*}$.

Beyond Bernoulli outcomes, identical targets and any prescribed finite feedback summaries can coexist with a regret ratio at least $K/(4\log_2 K)$ under equal $\log_2 K$ audit budgets (Proposition 6, Appendix B.2). Appendix B.1 gives an exact deterministic coverage coefficient.

6 Experiments

6.1 Evaluation protocol and comparison methods

We evaluate whether conditional reliability improves decisions under a fixed trusted-label budget, whether adaptive verification improves the proved fixed design, and how these benefits depend on model specification. Cumulative target regret R_T charges every deployed action, including rounds without an audit. Terminal simple regret is the gap of the recommended action. Lower values are better. We separately report trusted-label counts and the round of the final audit, since equal label budgets need not imply equal acquisition time.

Synthetic tasks. The primary study uses signed-coordinate kernels of ranks two and eight and a periodic RBF kernel on 16 actions. It crosses one or four reliability groups, candidate-state probabilities 1 and .25, and audit caps 128, 512, and 2,048, with horizon 8,192 and 256 paired replications per setting. The independent GP comparison uses 24 settings, $T = 4,096$, $M \in \{64, 256, 1,024\}$, and 128 paired replications. A separate adaptive-design study uses eight actions, four squared-exponential kernel signals, three reliability families, $(T, M) \in \{(512, 32), (2048, 64)\}$, and 512 paired replications. Each primary evaluation uses prespecified dictionaries, policies, and seeds.

Policies and controls. Reliability-Kernel-Design is Algorithm 1. The safe adaptive policy in Proposition 4 mixes a fixed robust joint law with Chernoff allocation using $\varepsilon = .25$. Conditional TS instead samples a reliability-signal pair from its posterior to select actions and audits all candidate-state outcomes until the cap. Its gate is fixed, so it is an adaptive-action heuristic, not the proved adaptive-verification policy. Controls include marginal-dictionary pooled TS, direct-arm UCB, likelihood decoders, GP policies, and a known-reliability oracle. Actions precede states, and only purchased outcomes update the models. GP controls use a Gaussian working likelihood with variance 1/4. Appendix C gives the grids, default parameters, and separately development-selected parameters.

Uncertainty and aggregation. Policies share environment and noise streams. Per-setting paired 95% intervals are descriptive and unadjusted. Aggregate ratios are geometric means of per-setting mean-regret ratios. Standard errors and bootstrap intervals condition on the fixed task grid, not a population of new tasks.

6.2 Regret reduction and component comparisons

Conditional feedback and adaptive action selection. In the 36-setting primary study, conditional TS has the lowest mean cumulative regret in 27 settings. Its paired comparison with pooled TS favors the conditional policy in 22 settings, is unresolved in 14, and never favors the pooled policy. The geometric regret ratios are .556, .339, and .369 as the audit cap increases (Figure 2). Direct-arm UCB has higher regret in 34 settings and no resolved improvement. The overlapping-reliability study gives a ratio of .569, supporting the mechanism beyond disjoint reliable states.

The fixed design improves terminal identification: mean simple regret falls from .0243 to .00053 across budgets. Its mean cumulative regret is nevertheless 196.1, versus 84.0 for conditional TS (at cap 2,048: 261.5 versus 24.7). The information guarantee therefore does not imply adaptive-regret optimality. With 75% uninformative states, only 51.3% of either policy’s trajectories reach that cap by the horizon.

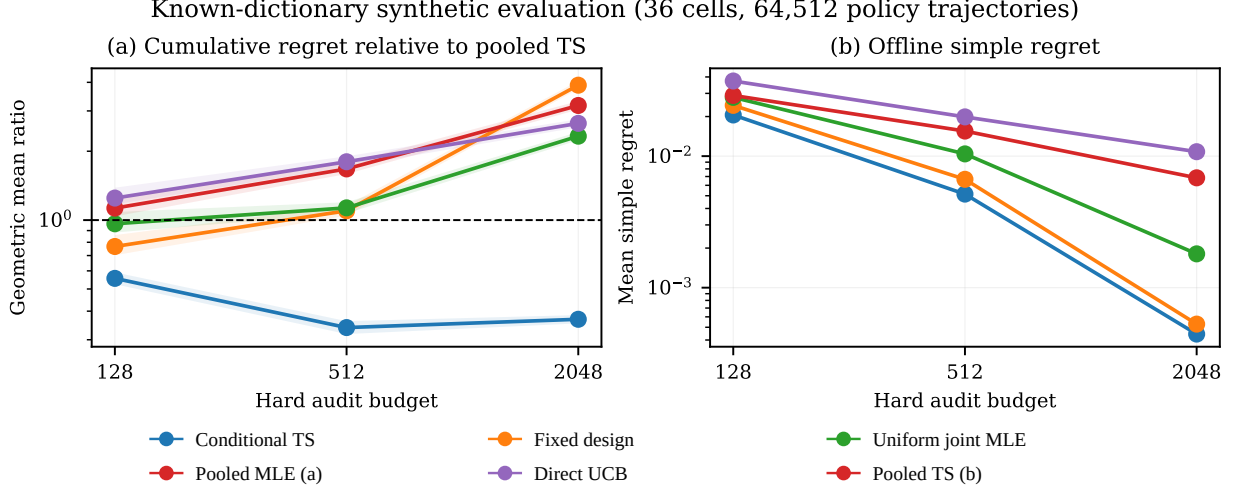


Figure 2: Conditional feedback under a hard audit cap, with 256 paired replications in each of 36 settings. Left: geometric mean cumulative-regret ratios relative to **pooled TS**. Right: mean terminal simple regret. Lower values are better. Shading in (a) is a paired-bootstrap 95% interval conditional on the fixed settings. The full setting-level comparison and audit-completion panel appear in Figure 3.

Table 1: Independent GP comparison on 24 settings and 128 paired replications per setting. Entries average settings equally. Cumulative regret is $\text{mean} \pm \text{Monte Carlo standard error}$, computed as $[\sum_{c=1}^{24} s_c^2/128]^{1/2}/24$, where s_c^2 is the within-setting sample variance. Bold marks the lowest mean in each regret column, not statistical significance. GP and identity rows use default parameters. Development-selected variants and audit counts appear in Table 4.

Policy	Cumulative regret ↓	Simple regret ↓
Conditional TS	92.24 ± 2.58	.0198
Candidate-gated pooled TS	99.92 ± 2.58	.0219
All-round joint TS	106.09 ± 2.62	.0241
All-round pooled TS	124.30 ± 2.66	.0287
GP-UCB	119.27 ± 2.64	.0275
GP-TS	130.40 ± 2.65	.0292
Identity-UCB	124.83 ± 2.78	.0286
Identity-TS	144.58 ± 2.76	.0325
Reliability-Kernel-Design	138.46 ± 2.74	.0226

GP controls and component effects. Table 1 reports absolute regret and uncertainty for the independent comparison. **Conditional TS** has lower mean cumulative and simple regret than all listed controls. Its geometric regret ratios are .663 against default **GP-UCB** and .612 against default **GP-TS**, although UCB wins two settings. Development-selected GP parameters do not improve the aggregate result. Matched kernel-versus-identity ratios are .925 for UCB and .886 for TS, with most paired comparisons unresolved. Joint versus pooled likelihood under the same candidate-state gate has ratio .901. The full conditional policy therefore benefits from the combination of its model and acquisition rule, while the isolated kernel effect is smaller.

Adaptive verification and acquisition time. On the separately seeded exact-dictionary study, safe adaptive allocation has regret ratio .933 relative to the fixed robust joint design. Nine settings

favor safe adaptive, 15 are unresolved, and none favor the fixed design. A subsequent comparison on new paired streams evaluates stronger controls (Table 2). Most comparisons with audit-only TS and posterior greedy are unresolved, and pure adaptive allocation is comparable without the robust floor. The supplementary study is post-measurement evidence, not independent confirmation. Equal T deployments and M labels conceal longer acquisition times under selective verification. The exact-model information floor holds, but its error bound clips to one in all 24 settings.

Table 2: Supplementary safe-adaptive comparisons on 24 settings with 512 paired replications. Regret and time are geometric ratios to the comparator. Time is the round of the M th audit. Ratios below one favor safe adaptive. Every policy uses T deployments and M labels. Table 3 gives per-setting uncertainty classifications and decision errors.

Comparator	Regret ratio ↓	Acquisition-time ratio ↓
Fixed robust joint design	.916	2.554
Pure adaptive design	1.006	.827
Product design	.916	4.764
Uniform action and verification	.907	4.764
Optimized action, uniform verification	.921	4.764
Audit-only dictionary TS	.943	4.764
Audit-only posterior greedy	.966	4.764

6.3 Model sensitivity and AI-task evaluation

Reliability errors and target perturbations. The robustness study separates changes to the conditional feedback law from changes to the reward target. Target-preserving feedback error gradually erodes the conditional policy’s advantage. The conditional-to-pooled regret ratio is 1.027 at feedback error .20 and 1.068 after exact feedback flattening. Target perturbations of .05 and .10 reverse the comparison, with ratios 1.398 and 1.799. Correct target geometry alone therefore does not ensure informative conditional feedback, as Proposition 5 predicts. Appendix C.4 gives the laws and uncertainty.

AI evaluation with fitted dictionaries. JudgeBench and the 32-judge ChaosNLI study test whether calibration can supply the model required by the online policy. The former uses 144 calibration labels and 1,584 archived judge outputs, while the latter uses 591 labels and 18,912 outputs. Online reveal budgets are additional to these calibration costs. On JudgeBench, fitted dictionaries predict worse than a pooled model and give regret ratios 11.63 and 16.46 against uniform verified TS. In the 32-judge study, clipping nearly every fitted signal entry makes the action design collapse to a tie, and selective verification changes acquisition time without improving regret over its matched uniform-state control.

A postdiagnostic 32-judge study separates reliability-set coverage from action-model expansion. Retaining two fitted signals while completing the reliability class improves all four settings relative to expanding both classes. The two-state reliability model has ratio .581 against uniform verified TS, whereas the one-state model has ratio 1.143. This contrast identifies action-model constraints as a concrete source of the earlier loss, but it is fitted-table development evidence and does not validate a learned selective gate. Appendix C gives the complete comparisons, offline replays, calibration uncertainty, and data-access provenance. The evidence supports the exact-model mechanism, while practical gains remain sensitive to model estimation and acquisition time.

References

- [1] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013.
- [2] S. Javdani, Y. Chen, A. Karbasi, A. Krause, J. A. Bagnell, and S. Srinivasa. Near optimal Bayesian active learning for decision making. In *AISTATS*, 2014.
- [3] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *COLT*, 2016.
- [4] Y. Chen, S. H. Hassani, and A. Krause. Near-optimal Bayesian active learning with correlated and noisy tests. In *AISTATS*, 2017.
- [5] I. Bárány and P. Soberón. Tverberg’s theorem is 50 years old: a survey. *Bulletin of the American Mathematical Society*, 55(4):459–492, 2018.
- [6] J. Altschuler, V.-E. Brunel, and A. Malek. Best-arm identification for contaminated bandits. *JMLR*, 20(91):1–39, 2019.
- [7] T. Fiez, L. Jain, K. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. In *NeurIPS*, 2019.
- [8] J. C. H. Lee and P. Valiant. Uncertainty about uncertainty: optimal adaptive algorithms for estimating mixtures of unknown coins. arXiv:1904.09228, 2019.
- [9] S. Raghavendran and V. Narayanan. The Prouhet Tarry Escott problem: a review. *Mathematics*, 7(3):227, 2019.
- [10] Y. Nie, X. Zhou, and M. Bansal. What can we learn from collective human opinions on natural language inference data? In *EMNLP*, 2020.
- [11] J. Kirschner, T. Lattimore, and A. Krause. Information directed sampling for linear partial monitoring. In *COLT*, 2020.
- [12] S. Ray Chowdhury, R. Oliveira, and F. Ramos. Active learning of conditional mean embeddings via Bayesian optimisation. In *UAI*, 2020.
- [13] I. Bogunovic and A. Krause. Misspecified Gaussian process bandit optimization. In *NeurIPS*, 2021.
- [14] R. Camilleri, K. Jamieson, and J. Katz-Samuels. High-dimensional experimental design and kernel bandits. In *ICML*, 2021.
- [15] Y. Efroni, N. Merlis, A. Saha, and S. Mannor. Confidence–budget matching for sequential budgeted learning. In *ICML*, 2021.
- [16] C. Réda, A. Tirinzoni, and R. Degenne. Dealing with misspecification in fixed-confidence linear Top- m identification. In *NeurIPS*, 2021.
- [17] S. Wu, C.-H. Wang, Y. Li, and G. Cheng. Residual bootstrap exploration for stochastic linear bandit. In *UAI*, volume 180 of *Proceedings of Machine Learning Research*, pages 2117–2127, 2022.
- [18] S. Mukherjee, A. S. Tripathy, and R. Nowak. Chernoff sampling for active testing and extension to active regression. In *AISTATS*, 2022.
- [19] A. Rangi et al. Stochastic bandits with limited verification. In *AAAI*, 2022.
- [20] S. Takemori. Distributionally-aware kernelized bandit problems for risk aversion. In *ICML*, 2022.
- [21] T. Cai and K. Kandasamy. Active cost-aware labeling of streaming data. In *AISTATS*, 2023.
- [22] P. Mikkola et al. Robust multi-fidelity Bayesian optimization. In *AISTATS*, 2023.
- [23] A. Verma et al. Exploiting correlated auxiliary feedback in parameterized bandits. In *NeurIPS*, 2023.
- [24] C. Wang, Z. Ye, Z. Feng, A. Badanidiyuru, and H. Xu. Follow-ups also matter: improving contextual bandits via post-serving contexts. In *NeurIPS*, 2023.
- [25] S. Saad-Falcon et al. ARES: an automated evaluation framework for retrieval-augmented generation systems. In *NAACL*, 2024.
- [26] Y. Chandak, S. Shankar, V. Syrgkanis, and E. Brunskill. Adaptive instrument design for indirect experiments. In *ICLR*, 2024.

- [27] S. Wu and A. A. Amini. Graph neural Thompson sampling. *arXiv:2406.10686*, 2024.
- [28] R. Liu et al. HbBoPs: prompt selection with Bayesian optimization and Hyperband. In *ICML*, 2025.
- [29] R. Ao, H. Chen, S. Gao, H. Li, and D. Simchi-Levi. Best arm identification with LLM judges and limited human. *arXiv:2601.21471*, 2026.
- [30] K. Lee and M. R. Kosorok. The surrogate is not the reward: post-surrogate primary-outcome acquisition in contextual bandits. *arXiv:2610.06610v1*, 2026. October 5.
- [31] A. N. Sridhar, R. Singh, M. Mahdavi, and V. Narayanan. Correlation-aware contextual bandits with surrogate rewards for LLM routing. *arXiv:2607.09015*, 2026.
- [32] K. Balagopalan, Y. Li, Y. Zhao, T. N. Nguyen, A. Daitche, H. Nassif, and K.-S. Jun. Fixed budget is no harder than fixed confidence in best-arm identification up to logarithmic factors. In *ICML*, 2026.
- [33] X. Gong and J. Zhang. The cost of information: phase transitions in contextual bandits with paid observations. In *ICML*, 2026.
- [34] M. O. Harding, V. Singh, and K. Kandasamy. Learning from biased and costly data sources: minimax-optimal data collection under a budget. In *COLT*, 2026.
- [35] K. W. Zhang, T. Baldwin-McDonald, K. Ciosek, L. Maystre, and D. Russo. Impatient bandits: optimizing for the long-term without delay. *JMLR*, 27:1–58, 2026.
- [36] S. Wu and A. A. Amini. Laplacian kernelized bandit. In *ICLR*, 2026.
- [37] W. Ji, Y. Pan, R. Zhu, and L. Lei. Multi-armed bandits with machine learning-generated surrogate rewards. *arXiv:2506.16658*, 2026.
- [38] Z. Du, H. Zhang, H. Zhu, and B. Zhang. Optimal design for active preference learning with biased LLM judges. *arXiv:2609.38860*, 2026.
- [39] X. J. Chen and Y. Tong. Gaussian process bandit optimization with machine learning predictions and application to hypothesis generation. *arXiv:2601.22315*, 2026.
- [40] C. Li, Y. Yu, and Y. Li. How many humans are 32 LLM judges worth? *arXiv:2609.21277*, 2026.
- [41] M. Shahverdikondori, A. M. Abouei, A. Rezaeimoghadam, and N. Kiyavash. Pure exploration beyond reward feedback: the role of post-action context. *arXiv:2502.03061v2*, 2026. May 12.
- [42] M. Sargent et al. Prediction-powered adaptive inference. In *ICML*, 2026.
- [43] L. Reyzin, A. Saha, and S. Wu. Expert advice with costly observations. In *UAI*, 2026.
- [44] F. Schur, J. Lago, and T. Fiez. The price of knowledge: optimal algorithms for costly bandits. In *UAI*, 2026.
- [45] W. Wei, S. Pal, K. Mukherjee, F. Deroncourt, R. A. Rossi, B. Kveton, and H. Eldardiry. Online learning with LLM experts from limited feedback. *arXiv:2609.05820*, 2026.
- [46] Q. Zeng, X. Wang, J. Shen, X. Liu, F. Kong, and J. Zuo. Best arm identification in generalized linear bandits via hybrid feedback. *arXiv:2605.05745*, 2026.

A Proofs

The proofs follow the order of the main statements. The final three proofs establish the supplementary results in Appendix B.

A.1 Proof of Proposition 1

Fix $e = (h, z)$ and $e' = (h', z')$ with $a_z^* \neq a_{z'}^*$. Let

$$H = C_h \cap C_{h'}, \quad A = C_h \setminus C_{h'}, \quad B = C_{h'} \setminus C_h.$$

The constant $1/2$ in the conditional means cancels. Independence of the action and state under the product design therefore gives

$$\begin{aligned} & \mathbb{E}_{s \sim \mathbf{p}, a \sim \lambda^*} [(m_e(a, s) - m_{e'}(a, s))^2] \\ &= p(H) \mathbb{E}_{a \sim \lambda^*} [(u_z(a) - u_{z'}(a))^2] + p(A) \mathbb{E}_{a \sim \lambda^*} [u_z(a)^2] + p(B) \mathbb{E}_{a \sim \lambda^*} [u_{z'}(a)^2] \\ &\geq V_+ \{p(H) + p(A) + p(B)\} = V_+ p(C_h \cup C_{h'}) \geq V_+ p(C_h) \geq \frac{V_+}{\tau^*}. \end{aligned}$$

The first inequality uses all three classes of constraints in (3). The last inequality follows from the feasibility of $\mathbf{p}$ in (1). This proves the proposition. $\square$

A.2 Proof of Proposition 2

Taking $\boldsymbol{\nu} = \lambda^* \otimes \mathbf{p}$ in the definition of J^* and applying Proposition 1 gives $J^* \geq V_+/\tau^*$.

For the reverse inequality, let $\mathcal{P} = \{(z, z') : a_z^* \neq a_{z'}^*\}$. Finite-game duality yields distributions $\mathbf{q}$ on $\mathcal{H}$ and $\boldsymbol{\xi}$ on $\mathcal{P}$ such that

$$\sum_h q_h \mathbf{1}\{s \in C_h\} \leq \frac{1}{\tau^*} \quad \text{for every } s, \quad \sum_{(z, z') \in \mathcal{P}} \xi_{z, z'} [u_z(a) - u_{z'}(a)]^2 \leq V_\Delta \quad \text{for every } a.$$

Fix an arbitrary joint design $\boldsymbol{\nu}$. Restricting the alternatives to $h = h'$ and then averaging them with $\mathbf{q} \otimes \boldsymbol{\xi}$ gives

$$\begin{aligned} & \min_{h, h', (z, z') \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\nu}} [(m_{h, z}(X, S) - m_{h', z'}(X, S))^2] \\ &\leq \sum_{s, a} \nu(a, s) \left(\sum_h q_h \mathbf{1}\{s \in C_h\} \right) \left(\sum_{(z, z') \in \mathcal{P}} \xi_{z, z'} [u_z(a) - u_{z'}(a)]^2 \right) \\ &\leq \frac{V_\Delta}{\tau^*} \sum_{s, a} \nu(a, s) = \frac{V_\Delta}{\tau^*}. \end{aligned}$$

Maximizing over $\boldsymbol{\nu}$ proves the upper bound. When $V_+ = V_\Delta$, the lower and upper bounds agree. $\square$

A.3 Proof of Theorem 1

Accepted observations and waiting times. Fix the true environment $e = (h, z)$ and write $\chi = \chi(\mathbf{p})$. Extend the design phase to an infinite sequence for the purpose of defining its stopping time. In each round, let I_t denote the acceptance indicator generated by the rejection rule. The actual policy uses this sequence only until M acceptances or the horizon. In one design round,

$$\Pr_e\{X_t = a, S_t = s, I_t = 1\} = \lambda_a^* \mu(s) \frac{p_s}{\chi \mu(s)} = \frac{\lambda_a^* p_s}{\chi}.$$

Thus I_t are independent Bernoulli($1/\chi$) variables and

$$\Pr_e\{X_t = a, S_t = s \mid I_t = 1\} = \lambda_a^* p_s.$$

Write $\tau_M = \inf\{t : \sum_{j=1}^t I_j = M\}$. The actual number of audits is at most M on every sample path. Moreover,

$$\begin{aligned} \mathbb{E}_e[\min\{T, \tau_M\}] &= \sum_{t=1}^T \Pr_e\{\tau_M \geq t\} \\ &= \sum_{t=0}^{T-1} \Pr\{\text{Bin}(t, 1/\chi) < M\} = \mathbf{N}_{T,M}(\chi). \end{aligned}$$

Regret during acquisition. Let $\mathcal{F}_{t-1}$ contain all variables generated before design round t , and define the target gap $\Delta_z(a) = \rho[u_z(a_z^*) - u_z(a)]$. The event $\{\tau_M \geq t\}$ is $\mathcal{F}_{t-1}$ -measurable, while X_t has law λ^* independently of that history. Hence the regret incurred before completion satisfies

$$\begin{aligned} \mathbb{E}_e \left[\sum_{t=1}^T \mathbf{1}\{\tau_M \geq t\} \Delta_z(X_t) \right] &= \sum_{t=1}^T \Pr_e\{\tau_M \geq t\} \mathbb{E}_{a \sim \lambda^*}[\Delta_z(a)] \\ &\leq \frac{C_\times}{\chi} \mathbf{N}_{T,M}(\chi). \end{aligned}$$

Likelihood comparison. It remains to control the recommendation error. Let Q_e be the law of one accepted triple (X, S, Y) , so that

$$Q_e(a, s, y) = p_s \lambda_a^* m_e(a, s)^y [1 - m_e(a, s)]^{1-y}, \quad y \in \{0, 1\}.$$

The accepted triples are independent with law Q_e . For any $e' = (h', z')$ with a different designated optimum, let $L_M(e')/L_M(e)$ be the likelihood ratio of their first M accepted triples. Markov's inequality and independence imply

$$\begin{aligned} \Pr_e\{L_M(e') \geq L_M(e)\} &\leq \mathbb{E}_e \left[\sqrt{L_M(e')/L_M(e)} \right] \\ &= \left(\sum_{s,a,y} \sqrt{Q_e(a, s, y) Q_{e'}(a, s, y)} \right)^M \\ &= \left(1 - \mathbb{E}_{s \sim p, a \sim \lambda^*} [H^2(P_e^{a,s}, P_{e'}^{a,s})] \right)^M \\ &\leq \exp \left\{ -\frac{M}{2} \mathbb{E}_{s \sim p, a \sim \lambda^*} [(m_e(a, s) - m_{e'}(a, s))^2] \right\} \\ &\leq \exp \left(-\frac{MV_+}{2\tau^*} \right). \end{aligned}$$

Here $H^2(P, Q) = 1 - \sum_y \sqrt{P(y)Q(y)}$. The penultimate line uses $1 - x \leq e^{-x}$ and the Bernoulli inequality $H^2(\text{Bern}(x), \text{Bern}(y)) \geq (x - y)^2/2$. The final line is Proposition 1.

Let E be the event that the maximum-likelihood recommendation has a different designated optimum. On E , some such e' has likelihood at least that of e , including ties. A union bound gives

$$\Pr_e(E) \leq |\mathcal{H}||\mathcal{Z}| \exp \left(-\frac{MV_+}{2\tau^*} \right).$$

Stopping and exploitation. To combine this error probability with the remaining horizon, let G_j be the waiting time for the j th accepted triple $(\bar{X}_j, \bar{S}_j, \bar{Y}_j)$. The repeated independent rejection

rule gives, for every integer $g \geq 1$ and measurable set D of triples,

$$\Pr_e\{G_j = g, (\bar{X}_j, \bar{S}_j, \bar{Y}_j) \in D\} = (1 - 1/\chi)^{g-1} \chi^{-1} Q_e(D).$$

The pairs indexed by j are independent. Thus $\tau_M = \sum_{j=1}^M G_j$ is independent of the accepted sample and of E . Since each target gap is at most $2\rho U$, exploitation regret is at most

$$\begin{aligned} 2\rho U \mathbb{E}_e[(T - \tau_M)_+ \mathbf{1}\{E\}] &= 2\rho U \{T - \mathbf{N}_{T,M}(\chi)\} \Pr_e(E) \\ &\leq 2\rho U \{T - \mathbf{N}_{T,M}(\chi)\} |\mathcal{H}| |\mathcal{Z}| e^{-M/(2\tau^* \mathfrak{D}_+)}. \end{aligned}$$

Adding the design and exploitation terms proves the regret bound. Finally, $\mathbb{E}_e[\tau_M] = M\chi$ gives $\mathbf{N}_{T,M}(\chi) \leq \min\{T, M\chi\}$. Substituting the stated audit threshold makes the last exponential at most $1/(2|\mathcal{H}||\mathcal{Z}|T)$, so the exploitation term is at most ρU . $\square$

A.4 Proof of Proposition 3

Write

$$m_a = \max_{s \in \mathcal{S}} \frac{\nu(a, s)}{\mu(s)}, \quad \Lambda = \sum_{a \in \mathcal{A}} m_a.$$

Actions with $m_a = 0$ are never selected. For $m_a > 0$, define

$$q_a = \frac{m_a}{\Lambda}, \quad \alpha_{a,s} = \frac{\nu(a, s)}{m_a \mu(s)}.$$

Because $m_a \geq \nu(a, s)/\mu(s)$, we have $0 \leq \alpha_{a,s} \leq 1$. A design round draws $X_t \sim \mathbf{q}$, observes S_t , and audits with probability α_{X_t, S_t} . Hence

$$\begin{aligned} \Pr\{X_t = a, S_t = s, B_t = 1\} &= q_a \mu(s) \alpha_{a,s} = \frac{\nu(a, s)}{\Lambda}, \\ \Pr\{B_t = 1\} &= \frac{1}{\Lambda}, \quad \Pr\{X_t = a, S_t = s \mid B_t = 1\} = \nu(a, s). \end{aligned}$$

Independent repetition therefore produces i.i.d. accepted pairs with law $\boldsymbol{\nu}$.

Causal lower bound. We next prove the causal lower bound. Consider any policy that produces one audited pair with law $\boldsymbol{\nu}$. Let $\mathcal{F}_{t-1}$ contain the history before action X_t is chosen, let τ be the first audited round, and define

$$N_a = \sum_{t \geq 1} \mathbf{1}\{\tau \geq t, X_t = a\}.$$

The event $\{\tau \geq t, X_t = a\}$ is measurable with respect to $\sigma(\mathcal{F}_{t-1}, X_t)$. Since $S_t \sim \mu$ independently of $\mathcal{F}_{t-1}$ and X_t , Tonelli's theorem gives, for every (a, s) ,

$$\begin{aligned} \mathbb{E} \left[\sum_{t \geq 1} \mathbf{1}\{\tau \geq t, X_t = a, S_t = s\} \right] &= \sum_{t \geq 1} \mathbb{E}[\mathbf{1}\{\tau \geq t, X_t = a\} \Pr\{S_t = s \mid \mathcal{F}_{t-1}, X_t\}] \\ &= \mu(s) \mathbb{E}[N_a]. \end{aligned}$$

The audited occurrence (X_τ, S_τ) is one term in this sum and has law $\boldsymbol{\nu}$. Therefore

$$\nu(a, s) \leq \mu(s) \mathbb{E}[N_a], \quad \mathbb{E}[N_a] \geq \max_s \frac{\nu(a, s)}{\mu(s)} = m_a.$$

Summing over a yields

$$\mathbb{E}[\tau] = \sum_a \mathbb{E}[N_a] \geq \sum_a m_a = \Lambda.$$

Under environment e , weighting the same inequality by $\Delta_e(a)$ yields the sampling-regret bound

$$\mathbb{E}_e \left[\sum_{t=1}^{\tau} \Delta_e(X_t) \right] = \sum_a \Delta_e(a) \mathbb{E}_e[N_a] \geq \sum_a \Delta_e(a) m_a = C_e(\boldsymbol{\nu}).$$

For the stationary construction, $\mathbb{E}[\tau] = \Lambda$ and $\mathbb{E}[N_a] = q_a \Lambda = m_a$, so both inequalities are equalities. The argument applies conditionally at each stage when the next design law is measurable with respect to the accepted history.

Finite-horizon regret. It remains to prove the finite-horizon regret bound. Before the M th audit, one design round has expected regret

$$\sum_a q_a \Delta_e(a) = \frac{C_e(\boldsymbol{\nu})}{\Lambda} \leq \frac{C(\boldsymbol{\nu})}{\Lambda}.$$

The accept indicators are i.i.d. Bernoulli($1/\Lambda$). Repeating the stopping-time calculation in Appendix A.3, the expected design length is $\mathbf{N}_{T,M}(\Lambda)$ and the expected exploitation length is $T - \mathbf{N}_{T,M}(\Lambda)$.

Let Q_e denote the law of one accepted triple. For a wrong-decision environment e' , its Bhattacharyya affinity with Q_e is

$$\begin{aligned} \sum_{a,s} \nu(a,s) \int \sqrt{dP_e^{a,s} dP_{e'}^{a,s}} &= 1 - \sum_{a,s} \nu(a,s) H^2(P_e^{a,s}, P_{e'}^{a,s}) \\ &\leq 1 - I(\boldsymbol{\nu}) \leq e^{-I(\boldsymbol{\nu})}. \end{aligned}$$

For the first M accepted triples, Markov's inequality applied to the square root of the likelihood ratio and independence give

$$\Pr_e \{L_M(e') \geq L_M(e)\} \leq e^{-MI(\boldsymbol{\nu})}.$$

A maximum-likelihood decision with the wrong optimum implies this event for at least one of the at most $|\mathcal{E}| - 1$ alternatives. A union bound therefore gives error probability at most $(|\mathcal{E}| - 1)e^{-MI(\boldsymbol{\nu})}$. The accepted triples are independent of their geometric waiting times, so this event is independent of the exploitation length. Multiplying its probability by the expected exploitation length and the largest one-round gap G , then adding the design regret, proves the displayed bound. If fewer than M audits arrive by time T , the policy remains in the design phase and has no exploitation segment. $\square$

A.5 Proof of Proposition 4

Fix the true environment e and a different-optimum alternative e' . Let $\mathcal{F}_{m-1}^{\text{aud}}$ be the observed history before stage m , including internal randomization, and let $L_m(e')/L_m(e)$ be the likelihood ratio after m accepted observations. The law $\boldsymbol{\nu}_m$ is measurable with respect to $\mathcal{F}_{m-1}^{\text{aud}}$ and is selected by the same rule under both environments. Therefore the design factors cancel from the likelihood

ratio. Conditional on $\mathcal{F}_{m-1}^{\text{aud}}$, we have

$$\begin{aligned} \mathbb{E}_e \left[\sqrt{\frac{L_m(e')}{L_m(e)}} \middle| \mathcal{F}_{m-1}^{\text{aud}} \right] \\ = \sqrt{\frac{L_{m-1}(e')}{L_{m-1}(e)}} \sum_{a,s} \nu_m(a,s) \int \sqrt{dP_e^{a,s} dP_{e'}^{a,s}} \\ = \sqrt{\frac{L_{m-1}(e')}{L_{m-1}(e)}} \left[1 - \sum_{a,s} \nu_m(a,s) H^2(P_e^{a,s}, P_{e'}^{a,s}) \right]. \end{aligned}$$

By nonnegativity and the mixture definition,

$$\sum_{a,s} \nu_m(a,s) H^2(P_e^{a,s}, P_{e'}^{a,s}) \geq \varepsilon \sum_{a,s} \nu_0(a,s) H^2(P_e^{a,s}, P_{e'}^{a,s}) \geq \varepsilon I_0.$$

Iterating conditional expectations, using $1 - x \leq e^{-x}$, and applying Markov's inequality gives

$$\Pr_e \{L_M(e') \geq L_M(e)\} \leq \mathbb{E}_e \left[\sqrt{\frac{L_M(e')}{L_M(e)}} \right] \leq (1 - \varepsilon I_0)^M \leq e^{-\varepsilon M I_0}.$$

A maximum-likelihood recommendation with the wrong optimum requires this event for at least one of at most $|\mathcal{E}| - 1$ alternatives. Hence its probability is at most

$$\min \left\{ 1, (|\mathcal{E}| - 1) e^{-\varepsilon M I_0} \right\}.$$

Acquisition cost and the finite horizon. Extend the design process beyond T until it collects all M audits. Conditional on $\mathcal{F}_{m-1}^{\text{aud}}$, Proposition 3 gives expected waiting time $\Lambda(\boldsymbol{\nu}_m)$ and expected design-stage regret $C_e(\boldsymbol{\nu}_m)$. The tower property and Tonelli's theorem give

$$\mathbb{E}_e[\tau_M] = \mathbb{E}_e \left[\sum_{m=1}^M \Lambda(\boldsymbol{\nu}_m) \right], \quad \mathbb{E}_e[R_{\text{design}}^{\text{ext}}] = \mathbb{E}_e \left[\sum_{m=1}^M C_e(\boldsymbol{\nu}_m) \right].$$

The horizon- T design regret is at most $R_{\text{design}}^{\text{ext}}$ pathwise. If all audits arrive by T , erroneous exploitation implies an error of the extended M -audit decoder and costs at most GT . If they do not arrive, there is no exploitation segment. Adding the two terms proves the regret bound. This coupling does not assume independence between completion time and decoding error. The algorithm checks the remaining budget before every purchase, so the cap holds pathwise. $\square$

A.6 Proof of Corollary 1

Let $E = \{e_\star \in \hat{\mathcal{E}}(\mathcal{C})\}$ be the exact-coverage event. Because E is measurable with respect to the independent calibration sample, the conditional online guarantee and the pathwise bound $0 \leq R_T \leq G_{\text{all}}T$ give

$$\begin{aligned} \mathbb{E}[R_T] &= \mathbb{E}[R_T \mathbf{1}\{E\}] + \mathbb{E}[R_T \mathbf{1}\{E^c\}] \\ &\leq \mathbb{E}[B_T(\mathcal{C}) \mathbf{1}\{E\}] + G_{\text{all}}T \Pr(E^c) \\ &\leq \mathbb{E}[B_T(\mathcal{C})] + \delta G_{\text{all}}T. \end{aligned}$$

The calibration phase consumes n_{cal} trusted labels before the online audit cap M , so the total trusted-label endowment is $n_{\text{cal}} + M$. $\square$

A.7 Proof of Theorem 2

Fix an arbitrary adaptive policy with at most M audits on every sample path. The dual of (1) supplies a distribution $\mathbf{q}$ over h such that

$$\Pr_{h \sim \mathbf{q}} \{s \in C_h\} \leq 1/\tau^\star \quad \forall s. \quad (7)$$

Fix $\pi \in \Pi_0$ and put the product prior $\mathbf{q} \otimes \pi$ on $E = (h, u)$. For each $e = (h, u)$, let P_e be the policy transcript law under the conditional mean in (2). Let P_0 be the transcript law obtained by replacing every audited outcome with an independent $\text{Bern}(1/2)$ variable while leaving the state process and the policy's internal randomization unchanged.

Let $\mathcal{F}_{t-1}$ contain the transcript before X_t is chosen, and let B_t be the audit indicator selected after observing S_t . Under P_0 , the latent variable E is independent of $\mathcal{F}_{t-1}$, X_t , S_t , and B_t . By (7), the zero-barycenter condition, and the definition of W_π , every audited query satisfies

$$\mathbb{E}_E \left[\mathbf{1}\{S_t \in C_h\} u(X_t)^2 \mid \mathcal{F}_{t-1}, X_t, S_t, B_t = 1 \right] \leq \frac{W_\pi}{\tau^\star}.$$

For $|x| \leq 1/4$,

$$\begin{aligned} \text{kl}(\text{Bern}(1/2), \text{Bern}(1/2 + x)) &= -\frac{1}{2} \log(1 - 4x^2) \\ &\leq \frac{8}{3} x^2, \end{aligned}$$

where the inequality holds because $4x^2 \leq 1/4$ and $-\log(1 - y) \leq 4y/3$ for $0 \leq y \leq 1/4$. The adaptive KL chain rule, followed by the preceding conditional bound and the pathwise constraint $\sum_{t=1}^T B_t \leq M$, yields

$$\begin{aligned} \mathbb{E}_E [\text{KL}(P_0 \| P_E)] &= \mathbb{E}_{0,E} \left[\sum_{t=1}^T B_t \text{kl}(\text{Bern}(1/2), \text{Bern}\{1/2 + \mathbf{1}(S_t \in C_h)u(X_t)\}) \right] \\ &\leq \frac{8}{3} \mathbb{E}_0 \left[\sum_{t=1}^T B_t \frac{W_\pi}{\tau^\star} \right] \leq \frac{8MW_\pi}{3\tau^\star}. \end{aligned}$$

For a fixed environment $e = (h, u)$ and transcript, define the target regret functional

$$L_e = \rho \sum_{t=1}^T \left\{ \max_a u(a) - u(X_t) \right\}.$$

It satisfies $0 \leq L_e \leq 2\rho UT$. Under P_0 , the policy actions are independent of u , so (4) gives

$$\begin{aligned} \mathbb{E}_E [\mathbb{E}_0[L_E]] &= \rho \sum_{t=1}^T \left\{ \mathbb{E}_{u \sim \pi} [\max_a u(a)] - \mathbb{E}_0 [\mathbb{E}_{u \sim \pi} [u(X_t)]] \right\} \\ &= \rho G_\pi T. \end{aligned}$$

For each e , the variational characterization of total variation gives

$$\mathbb{E}_e[L_e] \geq \mathbb{E}_0[L_e] - 2\rho UT \text{TV}(P_0, P_e).$$

Averaging over E , using Pinsker's inequality, Jensen's inequality, and the KL bound above, we obtain

$$\begin{aligned}\mathbb{E}_{e \sim \mathbf{q} \otimes \pi}[\mathbb{E}_{P_e}[L_e]] &\geq \rho G_\pi T - 2\rho UT \mathbb{E}_E[\text{TV}(P_0, P_E)] \\ &\geq \rho G_\pi T - 2\rho UT \sqrt{\frac{1}{2} \mathbb{E}_E[\text{KL}(P_0 \| P_E)]} \\ &\geq \rho T \left(G_\pi - 4U \sqrt{\frac{MW_\pi}{3\tau^*}} \right).\end{aligned}$$

The left-hand side is the Bayes regret of the fixed policy, and Bayes regret lower-bounds its worst-environment regret. Since regret is nonnegative, taking the positive part and then infimizing over policies proves the first claim. If $W_\pi = 0$, every action has $u(a) = 0$ almost surely under π , so $G_\pi = 0$ and the bound is trivial. Otherwise, if $M \leq 3\tau^* G_\pi^2 / (64U^2 W_\pi)$, the square-root term is at most $G_\pi/2$, which proves the second claim. $\square$

A.8 Coordinate specialization of the design bounds

For a design λ on $[d] \times \{\pm 1\}$, write

$$r_i = \lambda(i, +1) + \lambda(i, -1), \quad \sum_{i=1}^d r_i = 1.$$

The sign assigned within a coordinate does not affect any squared signal. For $u_{J,\theta}$,

$$\mathbb{E}_{(i,a) \sim \lambda}[u_{J,\theta}(i, a)^2] = \Delta^2 r_J.$$

Hence the energy part of $v_+(\lambda)$ is $\Delta^2 \min_J r_J \leq \Delta^2/d$. Under the uniform coordinate design $r_i = 1/d$. If two signals use distinct coordinates, their squared contrast is $\Delta^2(r_J + r_{J'}) = 2\Delta^2/d$. If they use the same coordinate with opposite signs, it is $4\Delta^2 r_J = 4\Delta^2/d$. All pairs with different designated optima fall into one of these cases. The energy constraint is therefore the active one and $V_+ = \Delta^2/d$.

The functions $g_J(i, a) = \Delta a \mathbf{1}\{i = J\}$ satisfy $\mathbb{E}_\lambda[g_J^2] = \Delta^2 r_J$, so the identical max-min calculation in (5) gives $V_- = \Delta^2/d$. Finally, if $d \geq 2$, the pure pairwise value V_Δ is at most $2\Delta^2/d$ because the two smallest coordinate masses sum to at most $2/d$. Uniform coordinate mass attains this bound, as the two contrast cases above show. Thus $V_\Delta = 2\Delta^2/d$. When $d = 1$, only the opposite-sign case remains and $V_\Delta = 4\Delta^2$. $\square$

A.9 Proof of Theorem 3

For the all- k -subset class, uniform weights $x_s = 1/k$ are feasible in (1). Conversely, averaging any feasible constraint over a uniformly random k -subset gives $(k/L) \sum_s x_s \geq 1$. Hence $\tau^* = L/k = 1/\rho$. For $M < d$, the stated upper bound exceeds the maximal regret $2\rho\Delta T$, so any recommendation suffices. For $M \geq d$, audit the first M rounds and assign at least $n = \lfloor M/d \rfloor$ sign-+1 queries to every coordinate. Marginalizing the i.i.d. state gives, independently across queries,

$$Y \mid I = i, A = +1 \sim \text{Bern}\left(\frac{1}{2} + \rho\theta\Delta \mathbf{1}\{i = J\}\right).$$

Choose the coordinate with largest absolute centered empirical mean and choose its sign from that mean. The active coordinate has centered expectation $\theta\rho\Delta$, while all other coordinates have expectation zero. If every empirical mean differs from its expectation by strictly less than $\rho\Delta/2$, the active coordinate and sign are recovered. For n Bernoulli samples, Hoeffding's inequality and a

union bound give

$$\Pr\{\hat{J} \neq J \text{ or } \hat{\theta} \neq \theta\} \leq 2d \exp\left(-\frac{n\rho^2\Delta^2}{2}\right).$$

Each exploration round and each round after an incorrect recommendation incurs at most $2\rho\Delta$. The first M rounds therefore contribute at most $2\rho\Delta M$, and the remaining rounds contribute at most the second term in the theorem.

Diffuse-reliability lower bound. For the converse, let P_0 be the transcript law with every audited outcome fair. For each coordinate j , let Q_j mix the sign-+1 environments uniformly over all k -subsets. Define $Q_j^\dagger$ by using the same state sequence and policy randomization as Q_j , but returning a fair outcome whenever a state label repeats. Before a repeat, after $t-1$ distinct labels have appeared, a newly observed label has posterior active probability at most

$$\frac{k}{L-t+1} \leq 2\rho.$$

This remains true after conditioning on adaptive actions and post-state audit decisions because the unobserved labels are exchangeable and at most k active labels remain. Conditional on a first-seen state, an audit of coordinate j under $Q_j^\dagger$ has mean $1/2 + \sigma p_t \Delta$, where $\sigma \in \{\pm 1\}$ is the played sign and $0 \leq p_t \leq 2\rho$. Thus its absolute displacement from $1/2$ is at most $2\rho\Delta \leq 1/4$. Audits of other coordinates are fair. For $|x| \leq 1/4$,

$$\text{kl}\{\text{Bern}(1/2), \text{Bern}(1/2 + x)\} \leq \frac{8}{3}x^2.$$

Let N_j be the number of audits on coordinate j . The adaptive KL chain rule, with all conditionals taken under P_0 , and the pathwise budget yield

$$\sum_{j=1}^d \text{KL}(P_0 \| Q_j^\dagger) \leq \frac{32}{3} \rho^2 \Delta^2 \mathbb{E}_0[\sum_j N_j] \leq \frac{32}{3} \rho^2 \Delta^2 M.$$

The coordinate play counts T_j satisfy $\sum_j T_j = T$. If j is uniform, Markov's inequality shows that each of the events $\{\mathbb{E}_0[T_j] > 4T/d\}$ and $\{\text{KL}(P_0 \| Q_j^\dagger) > 128\rho^2\Delta^2 M/(3d)\}$ contains fewer than one quarter of the coordinates. Hence at least one j satisfies both

$$\mathbb{E}_0[T_j] \leq \frac{4T}{d}, \quad \text{KL}(P_0 \| Q_j^\dagger) \leq \frac{128\rho^2\Delta^2 M}{3d} \leq \frac{1}{32}.$$

For $E_j = \{T_j \leq T/2\}$, Markov's inequality gives $P_0(E_j^c) \leq 8/d \leq 1/4$. The budget condition makes the displayed KL at most $1/32$, so Pinsker's inequality gives

$$Q_j^\dagger(E_j) \geq P_0(E_j) - \sqrt{\text{KL}(P_0 \| Q_j^\dagger)/2} \geq \frac{3}{4} - \frac{1}{8} = \frac{5}{8}.$$

The state sequence is common under Q_j and $Q_j^\dagger$. Their transcripts can differ only after a repeated state, whose probability is at most $T(T-1)/(2L) \leq 1/16$. Hence $Q_j(E_j) \geq 9/16 \geq 1/2$. On E_j , at least $T/2$ actions use another coordinate, and every such action incurs target regret $\rho\Delta$. The Bayes regret under Q_j is therefore at least $\rho\Delta(T/2)(1/2) = \rho\Delta T/4$. Since Q_j is a uniform mixture over fixed k -subsets, at least one fixed C attains this lower bound. The sign-+1 subfamily lies inside the stated signed class. $\square$

A.10 Proof of Proposition 5

Fix the true environment e and a wrong-decision e' . For one accepted query $x = (a, s) \sim \nu$, put $d(x) = m_e(x) - m_{e'}(x)$ and

$$W = \{Y - m_{e'}(x)\}^2 - \{Y - m_e(x)\}^2 = 2d(x) \left\{ Y - \frac{m_e(x) + m_{e'}(x)}{2} \right\}.$$

Writing $D_{e,e'}^2 = \mathbb{E}_\nu[d(x)^2]$ and $r_e(x) = \tilde{m}_e(x) - m_e(x)$, Cauchy–Schwarz gives

$$\overline{W} := \mathbb{E}_e[W] = D_{e,e'}^2 + 2\mathbb{E}_\nu[d(x)r_e(x)] \geq D_{e,e'}(D_{e,e'} - 2\delta_{\text{fb}}).$$

When $D_{e,e'} \leq 2\delta_{\text{fb}}$ the displayed upper bound is trivial. Otherwise, $|W| \leq 2|d(x)|$, $\mathbb{E}[W^2] \leq 4D_{e,e'}^2$, $\overline{W} \leq D_{e,e'}^2 + 2\delta_{\text{fb}}D_{e,e'} < 2D_{e,e'}^2$, and $|W - \overline{W}| \leq 4$. Bernstein’s inequality yields

$$\Pr_e \left\{ \sum_{m=1}^M W_m \leq 0 \right\} \leq \exp \left\{ -\frac{M\overline{W}^2}{2(4D_{e,e'}^2 + 4\overline{W}/3)} \right\} \leq \exp \left\{ -\frac{3M}{40}(D_{e,e'} - 2\delta_{\text{fb}})^2 \right\}.$$

A union bound over wrong-decision environments, followed by the same schedule calculation as in Proposition 3, proves the nominal-regret term. For every action, replacing f_e by $\tilde{f}_e$ changes its gap by at most $2\varepsilon_{\text{tar}}$, which adds at most $2\varepsilon_{\text{tar}}T$.

Target-preserving converse. For the converse, take actions $a \in \{-1, +1\}$, signs $\theta \in \{-1, +1\}$, and a known set C with $\mu(C) = \rho$. The nominal and true conditional means are

$$m_\theta(a, s) = \frac{1}{2} + \mathbf{1}\{s \in C\}\theta a\Delta, \quad \tilde{m}_\theta(a, s) = \frac{1}{2} + \rho\theta a\Delta.$$

Both have target $f_\theta(a) = \frac{1}{2} + \rho\theta a\Delta$, so $\varepsilon_{\text{tar}} = 0$, while their uniform conditional discrepancy is $(1 - \rho)\Delta$. The true feedback is state independent. Let P_+ and P_- be the complete transcript laws under the two signs. Conditional on any audited action, the outcome laws are $\text{Bern}(1/2 + \rho\Delta)$ and $\text{Bern}(1/2 - \rho\Delta)$, possibly in the opposite order. Unaudited rounds have identical conditional laws. The adaptive KL chain rule and the pathwise constraint therefore give

$$\text{KL}(P_+ \| P_-) \leq \mathbb{E}_+ \left[\sum_{t=1}^T B_t \right] \kappa_{\rho\Delta} \leq M\kappa_{\rho\Delta}.$$

Apply Bretagnolle–Huber to the event $\{X_t = +1\}$ using the transcript available before X_t is chosen. Data processing bounds its KL by the full-transcript KL, so

$$P_+\{X_t = -1\} + P_-\{X_t = +1\} \geq \frac{1}{2}e^{-M\kappa_{\rho\Delta}}.$$

The two one-round regret gaps are $2\rho\Delta$. Summing over t and averaging the two signs yields

$$\frac{\mathbb{E}_+[R_T] + \mathbb{E}_-[R_T]}{2} = \rho\Delta \sum_{t=1}^T [P_+\{X_t = -1\} + P_-\{X_t = +1\}] \geq \frac{\rho\Delta T}{2}e^{-M\kappa_{\rho\Delta}}.$$

Finally, $\kappa_x = 4x \operatorname{artanh}(2x) \leq 16x^2$ for $x \leq 1/4$. The maximum regret is at least the two-point average, which proves the converse. $\square$

A.11 Proof of Proposition 6

Index environments by leaves $h \in \{0, 1\}^d$ and put $\theta_h = (-1)^{\sum_i h_i}$. Besides decision actions $\{+1, -1\}$, include one diagnostic q_v for every internal binary-tree node. Decision outcomes have law $\text{Bern}(1/2 + a\theta_h\Delta)$.

Matching the feedback summaries. Fix distinct $x_1, \dots, x_{4p+5} \in (0, 1/2)$ and map each point into $\mathbb{R}^{2p+1}$ by

$$\mathbf{F}(x) = (x, \psi_1(x), \dots, \psi_p(x), \psi_1(1-x), \dots, \psi_p(1-x)).$$

Tverberg's theorem partitions these $4p+5 = 2\{(2p+1)+1\}+1$ points into three nonempty sets I_0, I_1, I_2 whose convex hulls intersect [5]. Hence there are probability weights $w_{j,i}$ supported on I_j with $\sum_{i \in I_j} w_{j,i} \mathbf{F}(x_i)$ independent of j . Define

$$Q_j = \frac{1}{2} \sum_{i \in I_j} w_{j,i} (\delta_{x_i} + \delta_{1-x_i}).$$

The three laws have pairwise disjoint supports, common mean $1/2$, and identical expectations of every ψ_r : the latter is half the matched $\psi_r(x)$ coordinate plus half the matched $\psi_r(1-x)$ coordinate. Their combined support has at most $8p+10$ points. Choosing $\psi_r(y) = y^r$ recovers finite raw-moment matching; the equal-weight Prouhet construction gives a separate explicit instance [9].

Equal targets and different information. In the coded dictionary, q_v returns $Q_{h_{|v|+1}}$ when v is a prefix of h , and Q_2 otherwise. In the uninformative dictionary every diagnostic returns Q_0 . Their common target is $f_h(a) = 1/2 + a\theta_h\Delta$ on decisions and $f_h(q_v) = 1/2$ on diagnostics. With $\phi(+1) = 1$, $\phi(-1) = -1$, and $\phi(q_v) = 0$, the centered target is $\theta_h\Delta\phi$ under the rank-one kernel $k(x, x') = \phi(x)\phi(x')$.

For the uninformative dictionary, fix opposite-parity leaves and let P_+, P_- be their transcript laws. Only audited decision actions distinguish them. The adaptive KL chain rule and the pathwise budget give

$$\text{KL}(P_+ \| P_-) \leq M\kappa_\Delta, \quad \kappa_\Delta = \text{kl}\{\text{Bern}(1/2 + \Delta), \text{Bern}(1/2 - \Delta)\}.$$

Achievable regret in the informative dictionary. The coded policy starts at the root. At depth r , it audits q_v for the prefix v already recovered. The resulting sample belongs to the disjoint support of Q_0, Q_1 , or Q_2 . Because the queried node is on the true path, only Q_0 or Q_1 can occur, and its support identifies bit h_{r+1} from one sample. After d audits the leaf and its parity are known. Each diagnostic action has target value $1/2$ while the optimal decision has value $1/2 + \Delta$, so these d rounds incur exactly $d\Delta$ regret. The policy then plays the optimal sign.

Lower bound in the uninformative dictionary. For the uninformative dictionary, map decision actions to their signs and map every diagnostic action to an independent fair sign. Under either environment, the expected instantaneous regret is exactly 2Δ times the probability that this mapped sign is wrong. For each t , apply Bretagnolle–Huber to the mapped sign based on the pre-action transcript. Data processing and the preceding KL bound give

$$P_+\{\hat{\theta}_t = -1\} + P_-\{\hat{\theta}_t = +1\} \geq \frac{1}{2}e^{-M\kappa_\Delta}.$$

Consequently,

$$\frac{\mathbb{E}_+[R_T] + \mathbb{E}_-[R_T]}{2} = \Delta \sum_{t=1}^T [P_+\{\hat{\theta}_t = -1\} + P_-\{\hat{\theta}_t = +1\}] \geq \frac{\Delta T}{2} e^{-M\kappa_\Delta}.$$

Thus one of the two leaves attains the stated lower bound. Finally, $\kappa_\Delta = 4\Delta \text{artanh}(2\Delta) \leq 16\Delta^2$ for $\Delta \leq 1/4$. Under $M = d$, $T = K$, and $\Delta = 1/(4\sqrt{K})$, we have $M\kappa_\Delta \leq d/K \leq 1/2$. Since $e^{-1/2}/2 \geq 1/4$, the uninformative regret is at least $\Delta K/4$, whereas the coded regret is $d\Delta$. Their ratio is at least $K/(4d) = K/(4\log_2 K)$. $\square$

A.12 Proof of Theorem 4

For the deterministic class, let $z_{s,j} \geq 0$ be a feasible fractional hitting weight on $\mathcal{S} \times [d]$. Fixing j gives $\sum_{s \in C_h} z_{s,j} \geq 1$ for every h , so $\sum_s z_{s,j} \geq \tau^*$. Summing over j gives mass at least $d\tau^*$. Conversely, copying an optimizer of (1) into every coordinate slice gives feasible mass $d\tau^*$. Thus the product set system has fractional hitting number exactly $d\tau^*$.

Normalize these weights to a distribution on (s, j) and draw M pairs independently, retaining their set of distinct values. Every $C_h \times \{J\}$ receives probability mass at least $1/(d\tau^*)$, so its hit probability is at least

$$1 - \left(1 - \frac{1}{d\tau^*}\right)^M.$$

For the reverse bound, linear-program duality supplies a distribution on (h, J) such that every single pair (s, j) belongs to $C_h \times \{J\}$ with probability at most $1/(d\tau^*)$. For any fixed set Q of at most M pairs, the union bound gives hit probability at most $M/(d\tau^*)$. Averaging over randomized Q and truncating at one proves (9).

Deterministic lower bound. For the regret lower bound, fix a policy and run it on the null transcript $Y_t = 1/2$. Its audited state-coordinate support Q is a random set of size at most M . The finite zero-sum game in (8) has a dual prior on (h, J) under which every distribution of Q misses $C_h \times \{J\}$ with probability at least $1 - c_M^{(d)}$. Draw θ independently and uniformly. On the miss event, every audited outcome equals $1/2$, so the true transcript agrees pathwise with the null transcript. Conditional on that transcript, the policy cannot depend on θ . Averaging over θ , every played action has target value $1/2$, while the optimal action has value $1/2 + \gamma\rho$. Therefore

$$\mathbb{E}[R_T] \geq \gamma\rho T \Pr\{Q \cap (C_h \times \{J\}) = \emptyset\} \geq \gamma\rho T(1 - c_M^{(d)}).$$

Matching deterministic policy. For the upper bound, draw a support Q from an optimizer of (8). Test its pairs in any fixed order. To test (s, i) , repeatedly play coordinate i with an independent uniform sign until state s appears, then audit. The deterministic residual is nonzero exactly when $s \in C_h$ and $i = J$, in which case its sign together with the played sign reveals θ . The policy then plays the optimal action forever. If every test is uninformative, it continues with independent uniform signs on an arbitrary coordinate. Before identification, uniform signs have expected target value $1/2$ and incur $\gamma\rho$ regret per round.

Let L_Q be the total number of rounds required to complete all tests. Each waiting time has expectation at most $1/\mu_{\min}$, so $\mathbb{E}[L_Q] \leq M/\mu_{\min}$. If Q hits $C_h \times \{J\}$, all regret is charged within these test rounds. If it misses, at most $\gamma\rho T$ additional regret is charged. Hence, uniformly over (h, J, θ) ,

$$\mathbb{E}[R_T] \leq \gamma\rho \mathbb{E}[L_Q] + \gamma\rho T \Pr\{Q \cap (C_h \times \{J\}) = \emptyset\} \leq \frac{\gamma\rho M}{\mu_{\min}} + \gamma\rho T(1 - c_M^{(d)}).$$

Dividing both bounds by T and letting $T \rightarrow \infty$ proves the limit. $\square$

A.13 Proof of Proposition 7

Proof. The definitions give $0 \leq \alpha_{a,s} \leq 1$. For one stationary design round,

$$\Pr\{X_t = a, S_t = s, B_t = 1\} = q_a \mu_a(s) \alpha_{a,s} = \frac{\nu(a, s)}{\Lambda_{\text{ad}}(\boldsymbol{\nu})}.$$

Summing over (a, s) proves both the acceptance probability and the conditional accepted law. Independent fresh rounds therefore give i.i.d. accepted pairs.

For the causal lower bound, let $\mathcal{F}_{t-1}$ contain the history before X_t is selected and set $N_a = \sum_{t \geq 1} \mathbf{1}\{\tau \geq t, X_t = a\}$. Conditional on $\mathcal{F}_{t-1}$ and $X_t = a$, the state law is μ_a . Tonelli's theorem therefore gives

$$\mathbb{E} \left[\sum_{t \geq 1} \mathbf{1}\{\tau \geq t, X_t = a, S_t = s\} \right] = \mu_a(s) \mathbb{E}[N_a].$$

The audited occurrence (X_τ, S_τ) is one term in this sum and has probability $\nu(a, s)$. Hence $\nu(a, s) \leq \mu_a(s) \mathbb{E}[N_a]$, so $\mathbb{E}[N_a] \geq m_a$. Summing over a proves the waiting-time bound, and weighting by $\Delta_e(a)$ proves the sampling-regret bound. Under the stationary compiler, $\mathbb{E}[\tau] = \Lambda_{\text{ad}}(\nu)$ and $\mathbb{E}[N_a] = q_a \mathbb{E}[\tau] = m_a$, proving equality. Conditioning on the accepted history gives the stagewise predictable statement. $\square$

B Supplementary theoretical results

This appendix gives the deterministic coverage specialization, the finite-summary counterexample, and the acquisition rule for known action-dependent diagnostic laws. These results refine the main guarantees without changing the action-independent assumptions used in the primary experiments. Their proofs are collected in Appendix A.

B.1 Deterministic coverage and audit-only comparison

The deterministic coordinate model has a sharper coefficient. Let $\mathcal{Q}_M = \{Q \subseteq \mathcal{S} \times [d] : |Q| \leq M\}$. Define

$$c_M^{(d)} = \max_{\omega \in \Delta(\mathcal{Q}_M)} \min_{h, J} \Pr_{Q \sim \omega} \{Q \cap (C_h \times \{J\}) \neq \emptyset\}. \quad (8)$$

The product dictionary has fractional hitting number $d\tau^*$ and

$$1 - \left(1 - \frac{1}{d\tau^*}\right)^M \leq c_M^{(d)} \leq \min \left\{1, \frac{M}{d\tau^*}\right\}. \quad (9)$$

Theorem 4 (Exact deterministic specialization). *Write the coordinate action as $X_t = (I_t, A_t) \in [d] \times \{\pm 1\}$. For $0 < \gamma \leq 1/2$, replacing (2) by $Y_t = \frac{1}{2} + \gamma\theta A_t \mathbf{1}\{I_t = J, S_t \in C_h\}$ gives*

$$\gamma\rho T(1 - c_M^{(d)}) \leq R_T^* \leq \gamma\rho T(1 - c_M^{(d)}) + \frac{\gamma\rho M}{\mu_{\min}}.$$

For fixed M , μ , and reliability class, $\lim_{T \rightarrow \infty} R_T^*/T = \gamma\rho(1 - c_M^{(d)})$.

See Appendix A.12 for the mixed-coverage duality argument and the matching audit policy.

For deterministic outcome contamination bounded by $\varepsilon < \gamma/2$, threshold decoding remains valid and target comparison adds at most $2\varepsilon T$. This does not cover overlapping informative and uninformative ranges. Sequential audit-only policies remain a necessary comparison because they use trusted outcomes without a reliability model [15]. The pooled policy in Theorem 3 gives one such guarantee, and Section 6 compares the conditional policies with marginal-dictionary and GP controls.

Standard GP information gain cannot determine the sharper class-dependent frontier. With the same rank- d coordinate kernel, a known active coordinate gives $\mathfrak{D}_+ = 1/\Delta^2$, while an unknown coordinate gives d/Δ^2 ; the kernel information-gain curve is unchanged. Extending the result to a full RKHS ball requires a separate packing or localization argument.

B.2 Information absent from finite feedback summaries

The next result allows bounded outcomes $Y \in [0, 1]$ and a singleton diagnostic state. It therefore concerns general conditional feedback laws rather than only the Bernoulli family in (2).

Proposition 6 (Finite-statistic distributional barrier). *Fix any integer $p \geq 1$ and bounded measurable statistics $\psi_1, \dots, \psi_p : [0, 1] \rightarrow \mathbb{R}$. For every integer $d \geq 1$, $K = 2^d$, $T \geq d$, and $0 < \Delta \leq 1/4$, there are two finite verified-feedback dictionaries with identical target functions, the same rank-one centered target kernel, zero target misspecification, and identical conditional expectations of all $\psi_i(Y)$ for every environment and action. One admits a causal $M = d$ policy with $R_T = d\Delta$. In the other, every causal policy using at most M audits satisfies*

$$\sup_e \mathbb{E}_e[R_T] \geq \frac{\Delta T}{2} e^{-M\kappa_\Delta}, \quad \kappa_\Delta = 2\Delta \log \frac{1+2\Delta}{1-2\Delta} \leq 16\Delta^2.$$

At $M = d$, $T = K$, and $\Delta = 1/(4\sqrt{K})$ with $K \geq 2d$, the regret ratio is at least $K/(4\log_2 K)$.

The construction and testing reduction are proved in Appendix A.11.

The construction combines Tverberg’s convex-partition theorem [5] with active testing. It shows why misspecifying the conditional feedback law cannot be covered by the $2\varepsilon T$ target-comparison term or by any fixed finite summary vector on an unrestricted law class: full-law distinguishability changes while every displayed statistic and target quantity remains fixed. It does not apply when the statistics identify a restricted conditional-law family; full laws or a characteristic representation can still distinguish the construction.

B.3 Action-dependent diagnostic laws

The structured model in Section 3 uses a common state law so that the reliable-set mass and the target reward separate. The rejection-sampling compiler has a broader scope. Suppose instead that, conditional on choosing action a and on the past, the next state has a known law μ_a that is common across candidate environments. Let $\nu \in \Delta(\mathcal{A} \times \mathcal{S})$ be supported only on cells with $\mu_a(s) > 0$, and define

$$m_a = \max_{s: \mu_a(s) > 0} \frac{\nu(a, s)}{\mu_a(s)}, \quad \Lambda_{\text{ad}}(\nu) = \sum_a m_a, \quad q_a = \frac{m_a}{\Lambda_{\text{ad}}(\nu)}, \quad \alpha_{a,s} = \frac{\nu(a, s)}{m_a \mu_a(s)},$$

where $\alpha_{a,s}$ is defined by the formula when $m_a \mu_a(s) > 0$ and set to zero otherwise. Zero rows are never selected.

Proposition 7 (Action-dependent compiler). *Selecting $X_t \sim \mathbf{q}$ and auditing after state s with probability $\alpha_{X_t, s}$ produces accepted action–state pairs that are i.i.d. ν , with acceptance probability $1/\Lambda_{\text{ad}}(\nu)$ per round. Moreover, every causal rule whose first audited pair has law ν satisfies*

$$\mathbb{E}[N_a] \geq m_a, \quad \mathbb{E}[\tau] \geq \Lambda_{\text{ad}}(\nu), \quad \mathbb{E}_e \left[\sum_{t=1}^{\tau} \Delta_e(X_t) \right] \geq C_e^{\text{ad}}(\nu) := \sum_a \Delta_e(a) m_a,$$

where N_a counts plays of action a through the first audit time τ . The stationary compiler attains all three bounds. The identities also hold conditionally for a predictable sequence of accepted-query laws, each held fixed until the next audit.

See Appendix A.13 for the construction and causal lower bound.

This proposition changes only the acquisition compiler. If a separate exact finite dictionary supplies outcome laws $P_e^{a,s}$ and action-dependent targets $f_e(a) = \sum_s \mu_a(s) \mathbb{E}_e[Y | a, s]$, the accepted-history Hellinger argument in Proposition 4 carries over after replacing Λ and C_e by Λ_{ad} and C_e^{ad} . Proposition 7 does not extend the reliable-set product factorization or adaptive converse, and it does not cover estimated state laws or a learned outcome dictionary.

C Experimental details and additional evidence

We first specify the synthetic comparisons and uncertainty calculations, then examine conditional-model errors and AI evaluation. Each study retains its full comparison family. The AI results distinguish offline identification, fitted-table policy evaluation, and independently tested synthetic policies. Numerical checks of the mathematical identities appear at the end of this appendix.

C.1 Synthetic environments, policies, and statistical analysis

Policy choices were fixed using separate development seeds. Implementation checks enforce action-before-state timing, posterior normalization, and the pathwise audit cap (Appendix C.8). Each cell uses 256 paired seeds and horizon 8,192. The primary grid crosses three action geometries (rank-two and rank-eight signed coordinates and a periodic RBF kernel on 16 actions), reliability dictionary sizes one and four, known cold-state mass zero and .75, and budgets 128, 512, and 2,048. This gives 36 cells, seven methods, and 64,512 trajectories. All cells are retained. Reported cumulative regret is target regret over calendar time; offline simple regret is evaluated from the terminal recommendation. We also record audits, cheap state observations, completion time, and runtime.

Pooled TS uses the exact state-marginal signal dictionary and purchases every early outcome until the cap. **Conditional TS** maintains the joint posterior over (h, z) and purchases every outcome in the union of candidate reliable sets until the cap; the audit rule does not learn. Both sample a target model for each action while budget remains, then exploit the maximum-likelihood model. **Direct UCB** purchases early outcomes and uses empirical-best exploitation after exhaustion. Joint and pooled likelihood controls use the same frozen action LP as Algorithm 1, but audit every early round. The known- h design is an oracle. States are generated by the environment, never chosen by these controls.

The paired intervals are descriptive, unadjusted per-setting comparisons across 256 shared environment/noise streams. They do not give simultaneous 95% coverage over the grid. Aggregate figure intervals use 2,000 paired bootstrap resamples within each cell, conditional on the fixed grid; cells are not resampled as random tasks. **Conditional TS** versus **pooled TS** changes both likelihood and audit gate; this is not an isolated kernel-sharing ablation. They do not establish the performance of learned audit allocation. The offline application replays evaluate correction estimators, while the JudgeBench and 32-judge policy studies below execute fitted finite-dictionary policies.

The overlapping suite replaces singleton reliable states by cyclic blocks of sizes two and four on eight states. It crosses two reward geometries, two state-arrival regimes, and three budgets (24 cells; 30,720 trajectories). Thirty-eight construction checks verify every target mean, overlap mass, posterior likelihood, and absence of unaudited outcomes. Paired 95% intervals favor the conditional policy in 19 settings, are unresolved in five, and favor the pooled policy in none; its budget-wise geometric regret ratios are .904, .522, .392.

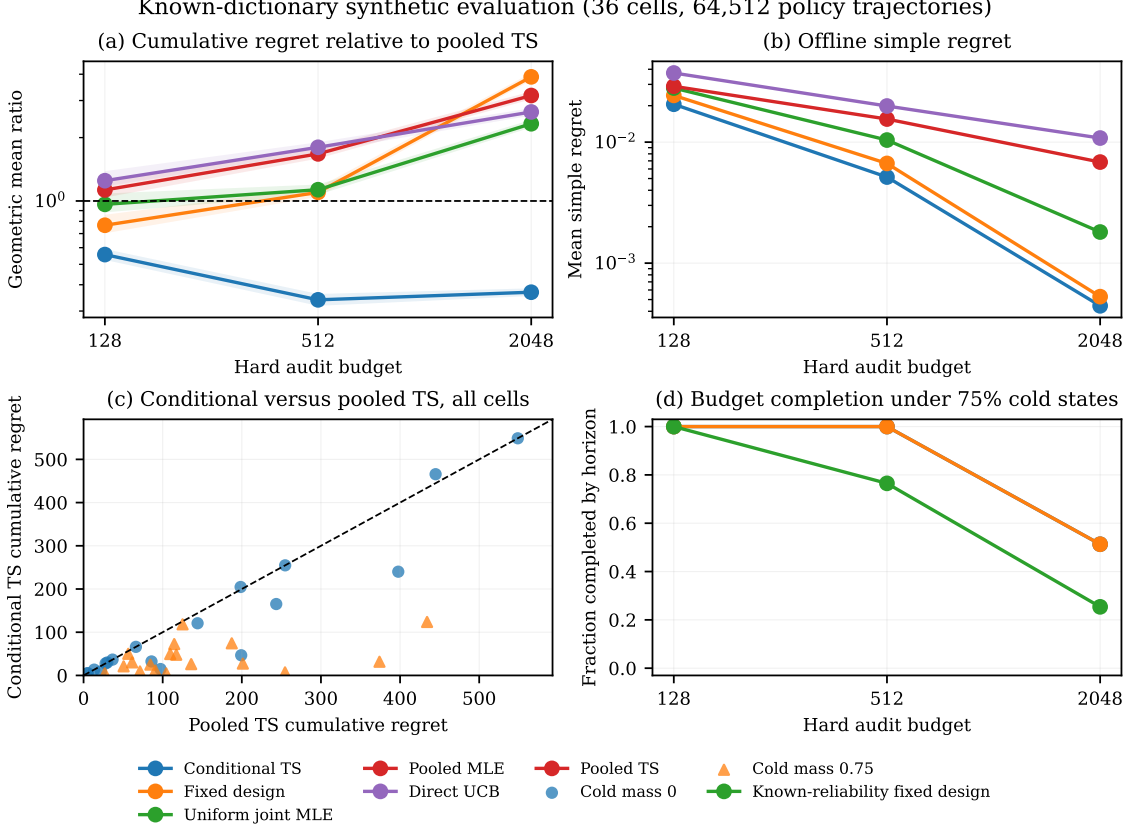


Figure 3: Complete primary synthetic comparison. Panels (a) and (b) show cumulative-regret ratios and terminal simple regret. Panel (c) compares all 36 settings with **pooled TS**. Circles and triangles indicate candidate-state probabilities 1 and .25, and points below the diagonal favor **conditional TS**. Panel (d) reports the fraction reaching the audit cap by the horizon when 75% of states are uninformative. There are 256 paired replications per setting. Shading in (a) is a paired-bootstrap 95% interval. The red curve denotes pooled maximum likelihood in (a) and pooled TS in (b).

C.2 Adaptive verification: regret and acquisition tradeoffs

The theory-aligned adaptive suite uses eight actions, four squared-exponential kernel sections, and three prespecified reliability families: four singleton sets, six cyclic adjacent pairs, and four disjoint adjacent pairs. It crosses two horizon-cap pairs (512, 32) and (2048, 64), giving 24 settings. The floor $\varepsilon = .25$, Chernoff alternative rule, analysis, and 512 paired replications were selected on a separate 24-system development study before the fresh suite was measured. Every environment, setting, and outcome is retained.

The fresh suite contains 36,864 trajectories. Safe adaptive versus the fixed robust joint design has geometric regret ratio .933 and nine lower-regret, 15 unresolved, and no higher-regret paired comparisons; the three reliability-family ratios are .937, .934, .929. Pure adaptive allocation has ratio .915 and 13 lower-regret and 11 unresolved comparisons, but does not preserve the robust information floor. One safe-adaptive point estimate is adverse, with ratio 1.021, and remains unresolved. Independent reconstruction from every-round action counts and accepted success/failure counts reproduces total regret and every final likelihood for all trajectories; the maximum errors are 6.14×10^{-12} and 8.53×10^{-14} .

A supplementary study was specified after the preceding outcomes and then run on new paired streams. It retains all 24 settings and matches the exact dictionary, zero-label calibration endowment, horizon, cap, and maximum-likelihood commit rule. Every trajectory has exactly T action deployments and M trusted labels. Regret charges all T actions, including unaudited waiting rounds. Table 3 reports all comparators and separately measures the round on which the cap is reached. The audit-only controls either sample an environment from the exact posterior or choose the posterior-mean-best action, and purchase every early trusted outcome until the cap.

Comparator	Regret ratio	Time ratio	Error (pp)	Lower	Unclear	Higher
Fixed robust joint design	.916	2.554	−7.06	13	11	0
Pure adaptive design	1.006	.827	+0.66	0	24	0
Product design	.916	4.764	−6.89	12	12	0
Uniform action and verification	.907	4.764	−7.71	11	13	0
Optimized action, uniform verification	.921	4.764	−6.96	11	12	1
Audit-only dictionary TS	.943	4.764	−4.92	6	18	0
Audit-only posterior greedy	.966	4.764	−4.21	3	21	0

Table 3: Supplementary certified-adaptive comparisons on 512 paired replications per setting. Regret and time are safe/comparator geometric ratios across all 24 settings; time is the round of the M th audit. Error is the mean difference in final wrong-decision rate, in percentage points, so negative values favor safe adaptive. Lower, unclear, and higher count settings in which the paired 95% interval places safe-adaptive regret below, across, or above the comparator. The intervals are unadjusted descriptive comparisons, so unclear does not establish equivalence. Comparator choices followed the fresh suite and are not independent confirmation.

The one resolved adverse comparison is against optimized-action uniform verification for the disjoint-pair family at $(T, M) = (512, 32)$ and the smoothest signal: ratio 1.078 and paired regret difference .738 with interval [.026, 1.451]. At the level of sample means, safe adaptive has higher regret in 2 settings against fixed robust design, 12 against pure adaptive, 1 against each product or uniform design, 3 against audit-only dictionary TS, and 5 against audit-only posterior greedy. Safe adaptive takes $4.764\times$ as many rounds geometrically as every-round controls (setting range 3.238–6.803), so this is a trusted-label comparison rather than a general cost advantage. The supplementary study comprises 98,304 trajectories. Its acquisition accounting uses every action and every purchased label. Proposition 4’s raw union-bound terms range from 14.47 to 22.82, so clipping makes the guaranteed error upper bound one in all 24 settings. The result proves a positive information floor but does not quantitatively explain the observed finite-budget gain. These exact-model experiments do not establish learned-dictionary validity, calibration coverage, a kernel effect, monetary savings, or real-task benefit.

C.3 Gaussian-process and factorial controls

A development screen for the Gaussian-process controls was specified before execution. It crossed coordinate and periodic-RBF actions, one and four reliability groups, cold-state masses zero and .75, budgets 64 and 256, horizon 1,024, and 24 repetitions. GP and identity controls received the same prior-scale grid, and UCB controls received the same exploration-width grid. Each method minimized mean cumulative regret over the complete development grid. The selected settings and the defaults were then frozen before an independently seeded measured comparison.

The measured grid adds budget 1,024, uses horizon 4,096 and 128 paired repetitions, and contains 24 cells, 13 methods, and 39,936 trajectories. GP-UCB and GP-TS use the target kernel and a

Gaussian working likelihood with variance $1/4$ for bounded Bernoulli outcomes. **Identity-UCB** and **identity-TS** use the same updates and hyperparameters without cross-action covariance. The prior mean is $1/2$ and covariance is $(s_0\rho U)^2\mathbf{K}$, with $s_0 \in \{.5, 1, 2\}$ in the screen. The UCB multiplier is chosen from $\{1, 2, 3\}$. Defaults use $s_0 = 1$ and multiplier two. Selection gives $(s_0, \beta) = (2, 3)$ for GP-UCB, $s_0 = 2$ for GP-TS, $(.5, 3)$ for **identity-UCB**, and $s_0 = 2$ for **identity-TS**. Both types audit every round until the cap and then exploit a posterior-mean maximizer. **Candidate-gated pooled TS** uses audited mean $1/2 + u_z(a)/g$, where g is the number of candidate reliability groups. **All-round pooled TS** instead uses $1/2 + \rho u_z(a)$. These controls are computational GP baselines, not exact Bayesian inference for the finite signal dictionary. Every method observes 4,096 cheap states and respects the hard audit cap.

Policy	Cumulative regret ↓	Standard error	Simple regret ↓	Audits
Conditional TS	92.24	2.58	.0198	446.14
Candidate-gated pooled TS	99.92	2.58	.0219	446.14
All-round joint TS	106.09	2.62	.0241	448.00
GP-UCB, default	119.27	2.64	.0275	448.00
GP-UCB, selected	122.34	2.65	.0276	448.00
Identity-UCB, selected	123.17	2.72	.0284	448.00
All-round pooled TS	124.30	2.66	.0287	448.00
Identity-UCB, default	124.83	2.78	.0286	448.00
GP-TS, default	130.40	2.65	.0292	448.00
GP-TS, selected	136.68	2.70	.0310	448.00
Reliability-Kernel-Design	138.46	2.74	.0226	446.14
Identity-TS, default	144.58	2.76	.0325	448.00
Identity-TS, selected	149.37	2.76	.0337	448.00

Table 4: Independent Gaussian-process control suite. Means average the 24 fixed settings equally. SE is $[\sum_{c=1}^{24} s_c^2/128]^{1/2}/24$, where s_c^2 is the within-setting sample variance. Bold denotes the lowest mean regret. “Selected” denotes the setting chosen on the separate development grid.

Against default GP-UCB, **conditional TS** has 15 lower-regret, seven unresolved, and two higher-regret paired comparisons and geometric-mean regret ratio .663. Against default GP-TS 15 comparisons favor the conditional policy and nine are unresolved, with ratio .612. Both GP-UCB losses occur in coordinate cells. The selected GP settings have higher aggregate regret than their defaults and are retained in Table 4.

Matched comparisons separate the ingredients. Kernel versus identity covariance favors GP-UCB in three settings and GP-TS in six, with respective regret ratios .925 and .886. Joint versus pooled likelihood favors the joint model in six settings under the candidate gate and nine under all-round auditing, with ratios .901 and .800. Candidate-state auditing is favored in ten settings under either likelihood, with ratios .820 for the joint model and .728 for the pooled model. All remaining comparisons are unresolved, and none favor the comparator. Thus the complete conditional policy has a nominal advantage in these families, while the isolated kernel effect is smaller and often unresolved. The intervals are unadjusted descriptive 95% normal intervals for paired seed differences within each cell.

C.4 Feedback-law and target misspecification

The supported robustness grid has 36 cells, five methods, and 46,080 trajectories. Write the nominal mean as $1/2 + \mathbf{1}\{s \in C_h\}u_z(a)$. The feedback perturbation uses $1/2 + [\mathbf{1}\{s \in C_h\}(1 - \alpha) + \mathbf{1}\{s \notin$

$C_h\}\alpha\rho/(1-\rho)]u_z(a)$, preserving the target exactly. Here $\rho = 1/16$, $|u_z(a)| \leq 1/4$, and the conditional-error radius is $\alpha/4$. Exact flattening requires $\alpha = 1 - \rho$. The target perturbation instead adds a state-independent cosine residual $r(a)$, of radius .05 or .10, to both the conditional outcome mean and the scoring target. This tests target/dictionary error and also shifts the conditional means; it does not leave the full feedback model correct. On the finite full-rank RBF grid, it is not evidence that a specified RKHS norm bound is violated.

All reported perturbation experiments use the outcome laws above, with target rewards obtained by marginalizing those same laws. Policies, audit budgets, horizons, kernels, and perturbation magnitudes were specified before evaluation. Exact checks verified the marginal targets, bounded Bernoulli probabilities, and state independence in the flattened model. Table 5 reports every specified perturbation, including settings in which conditional inference increases regret.

Scenario	Conditional regret	Pooled regret	Geometric ratio
Nominal	54.8	88.1	.572
Feedback error .10	78.0	88.4	.875
Feedback error .20	90.8	89.2	1.027
Feedback flattened	92.4	87.4	1.068
Target perturbation .05	263.1	199.1	1.398
Target perturbation .10	350.7	204.5	1.799

Table 5: Robustness suite. Means are final cumulative target regret; ratios are geometric means over six cells per scenario.

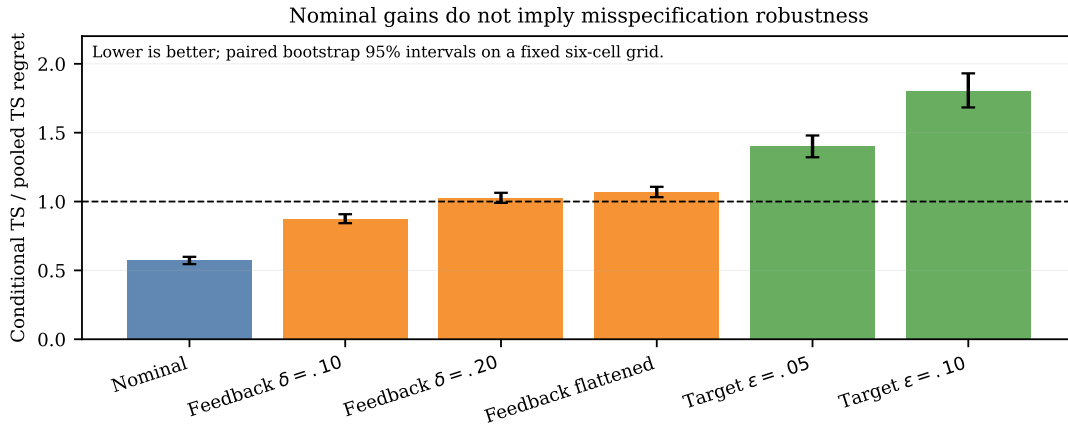


Figure 4: Misspecification defeats the nominal advantage. Feedback perturbations preserve target rewards; target perturbations shift both rewards and outcome means. Bars show geometric mean ratios with paired-bootstrap 95% intervals conditional on six cells.

C.5 Offline AI-system identification

SAID RAG Evaluation contains 24 configurations and 15 requests per dataset, with three human ratings per output and four archived judge dimensions. The RAG method uses two-fold multivariate cross-fitting and a frozen additive generator-pipeline kernel. HotpotQA is development; PubMedQA remains unopened. AgentRewardBench development froze five binary judges and coefficient .9, the request split, audit design, 40,000-trial count, .02 simple-regret tolerance, and 90% success endpoint. The confirmation program verified the frozen protocol hash before reading targets.

Selection endpoint and evaluation costs. All selection experiments report the first audit budget attaining 90% probability of simple regret at most .02. SAID uses 10,000 fixed-seed trials and AgentRewardBench uses 40,000. Audit-only methods receive the same trusted samples as corrected methods. Kernel smoothers are frozen before confirmation. Judge calls are counted even when archived; no missing ground truth is replaced by a prediction.

RAGTruth transfer provides no gain over its strongest kernel audit-only baselines: QA ties at 48 ratings, data-to-text ties at 6, and summarization uses 18 versus 12. SummEval’s kernel baseline saturates the endpoint. No claim is made for the unevaluated PubMedQA split.

A separate 32-judge Civil Comments development screen is a negative control. Its best oracle proxy predicts only $1.07\times$ mean audit savings, and no tested method reaches the endpoint by 160 trusted outcomes per judge. Its reserved 248-item split is excluded from the reported comparison.

For AgentRewardBench, the frozen judge set averages two priced GPT-4o-mini variants and three self-hosted Llama/Qwen variants. Logged API costs cover only the priced judges. The table therefore makes no monetary-saving claim. Raw or row-level derivatives are not redistributed.

C.6 JudgeBench: conditional correction and dictionary fit

Conditional-correction policies. JudgeBench is analyzed only on the frozen calibration and development groups. The simulator draws states with calibration frequencies and scores each action by its development conditional means under that same state law. The retained data loaders skip the 106 reserved rows before decoding (see the provenance statement below). Policies, horizons, caps, repetitions, and seeds are unchanged from the frozen protocol; only the inconsistent regret estimand and acquisition ledger were repaired.

Table 6: JudgeBench conditional-correction development replay. Lower cumulative target regret is better. Calibration-informed methods share 144 offline labels and 1,584 archived judge outputs; audit-only receives neither. Negative paired differences favor state-aware audits; all 95% intervals include zero. Bold marks the lowest mean, not significance.

Method / comparison	$T = 512$		$T = 2048$	
	$M = 16$	$M = 64$	$M = 16$	$M = 64$
<i>Mean cumulative regret: shared calibration endowment</i>				
Calibration greedy	0.000	0.000	0.000	0.000
Calibration only	9.327	9.242	37.190	37.015
Uniform audits	9.090	8.252	35.885	33.249
State-aware audits	9.040	8.566	36.315	34.649
Audit only (no calibration)	97.036	82.003	382.906	322.982
<i>Paired comparison: state-aware minus uniform audits</i>				
Mean difference	−0.050	0.314	0.430	1.400
95% interval	[−0.305, 0.189]	[−0.039, 0.691]	[−0.518, 1.360]	[−0.147, 3.020]

The policy Monte Carlo intervals condition on the fixed tables. Separately, 20,000 paired item bootstraps resample development items within each state while retaining calibration state weights. Calibration greedy is best or tied in 90.625% of those resamples; its regret against the resampled best action has mean .00211 and percentile interval $[0, .03041]$. This is a finite-item sensitivity check, not population-level uncertainty or policy re-learning. Historical price, token, labeling-time, and latency fields are absent from the release.

Table 7: JudgeBench online acquisition. “Reveals” counts trusted-label requests. “Unique” counts distinct development items, and “reuse” is their difference. Values are Monte Carlo means. Offline calibration costs are additional.

T	M	State-aware audits			Uniform audits		
		Reveals	Unique	Reuse	Reveals	Unique	Reuse
512	16	14.6	13.5	1.1	14.4	13.4	1.0
512	64	61.0	44.4	16.6	61.3	45.6	15.8
2048	16	14.5	13.5	.9	14.3	13.4	.9
2048	64	60.8	44.2	16.6	60.5	45.1	15.4

The finite-dictionary online policy. We separately execute the online policy in Algorithm 1. The reliability dictionary is constructed only from the 144 calibration items: deterministic two-fold validation selects two candidate sets for each prespecified cardinality $k \in \{1, 2\}$, signals are refit on all calibration items, and the dictionary and both designs are then frozen. This construction uses 144 trusted labels and 1,584 archived judge outputs before online interaction. Dictionary learning is outside the fixed-dictionary theorem. The online policies share this endowment, observe one archived judge output per round, and obey the same pathwise trusted-label cap. All four (T, M) cells, both dictionary sizes, matched uniform-state and uniform-action ablations, calibration controls, and audit-only TS were retained.

The development Bernoulli NLL is .64475 for the pooled calibration model and .65311 for a saturated action-state calibration model, versus .68828 and .67361 for the $k = 1$ and $k = 2$ dictionaries. Thus neither selected dictionary predicts the development table as well as the pooled model. Table 8 shows the corresponding policy result.

Table 8: Non-confirmatory JudgeBench execution of Algorithm 1. RK denotes **Reliability-Kernel-Design**. Entries are mean cumulative regret across 512 repetitions. Lower is better. Both prespecified dictionary sizes and all four cells are retained.

T	M	Calibration greedy	Uniform verified TS	RK design $k = 1$	RK design $k = 2$
512	16	0.000	5.469	58.801	80.364
512	64	0.000	4.921	62.836	113.122
2048	16	0.000	21.758	232.605	288.367
2048	64	0.000	19.090	238.288	312.919

The geometric-mean regret ratios of the RK designs to uniform verified TS are 11.63 and 16.46. Against the same action design with uniform-state auditing, the $k = 1$ design is worse in all four cells; the $k = 2$ design is better in one cell and worse in three. The selective gate approximately doubles the rounds needed to collect M audits for $k = 1$ and quadruples them for $k = 2$. At $(T, M) = (2048, 64)$, the two RK designs use 30.0 and 24.6 unique development items for 64 reveal calls, versus 43.6 under uniform verification. This is adverse, non-confirmatory development evidence: the 100 development labels had been opened by earlier analyses, while retained loaders skip the 106 reserved rows before decoding.

C.7 Thirty-two LLM judges: signal clipping and reliability coverage

Task, feedback, and calibration. A fresh replay couples all 32 archived LLM judges from Li et al. [40] to ChaosNLI human-majority labels [10]. An action chooses a judge before the task family is realized; the three action-independent states are MNL-matched, SNLI, and α NLI; the

cheap output is the selected judge’s archived class; and an audit reveals the majority of 100 human annotations. The hash split contains 591 calibration and 1,248 development items. The 1,140-row reserved partition is excluded from the reported study. All ten policies, both prespecified dictionary sizes, four horizon–budget settings, and 512 repetitions were fixed before outcome decoding for this study. Calibration-informed policies share an offline endowment of 591 trusted labels and 18,912 judge outputs, observe one archived output per online round, and obey the same pathwise reveal cap. The study comprises 20,480 trajectories.

The saturated action–state model has development Bernoulli NLL .48280, compared with .52020 for the pooled model and .59460 and .53301 for the $k = 1$ and $k = 2$ dictionaries. All 64 refit signal entries for $k = 1$ and 63 of 64 for $k = 2$ equal the clipping ceiling .24. Consequently, both action designs collapse to a one-point design chosen from a clipped tie. The RK policies have lower fixed-table regret than their uniform-action and uniformly verified controls, but their regret is exactly equal to the matched uniform-state ablation in every cell and repetition. The state gate therefore changes only acquisition time: at $(T, M) = (4096, 128)$, the mean time to collect the cap is 192.5 rounds for $k = 1$ and 386.5 for $k = 2$, versus 128 under uniform-state verification. This adverse result shows that high-accuracy judge rewards can lie outside the useful range of the centered clipped dictionary. It supplies no positive evidence for state-aware auditing; historical price, token, latency, and human-labeling costs are unavailable and are not imputed.

Separating reliability coverage from action-model expansion. Expanding both the reliability class and the signal dictionary changes two parts of the policy simultaneously. A postdiagnostic study separates their effects. For each $k \in \{1, 2\}$, the reliability class contains all three k -subsets of the three task states, while the signal dictionary retains only the two baseline-offset signals selected by calibration-only cross-validation; the environment class is their Cartesian product. The reliability LP therefore gives a uniform state gate, while the action LP is unchanged from the two-signal design. Because the common fitted baseline cancels from every likelihood contrast, the product-separation guarantee applies conditional on these fitted signals. It does not cover calibration or population uncertainty.

Table 9: Thirty-two-judge development replay: cumulative regret, mean $\pm$ Monte Carlo standard error from 512 paired replications. Each method keeps the same two fitted signals and uniform state gate. The full-class method includes all reliability sets of size k in its decoder, while the matched control includes only the two fitted sets. Bold denotes the lower mean in each row, not a significant difference.

T	M	Full reliability class	Two fitted reliability sets
<i>One reliable state ($k = 1$)</i>			
1,024	32	32.20 $\pm$.92	31.79 $\pm$.93
1,024	128	36.09 $\pm$.82	35.65 $\pm$.83
4,096	32	127.10 $\pm$ 3.79	125.76 $\pm$ 3.79
4,096	128	140.88 $\pm$ 3.63	139.90 $\pm$ 3.64
<i>Two reliable states ($k = 2$)</i>			
1,024	32	16.00 $\pm$.24	16.32 $\pm$.24
1,024	128	20.38 $\pm$.21	20.57 $\pm$.20
4,096	32	61.38 $\pm$.98	62.60 $\pm$.97
4,096	128	69.61 $\pm$.91	70.03 $\pm$.90

Closing the reliability class changes mean regret little relative to the matched two-set model: the geometric ratios are 1.011 for $k = 1$ (all four comparisons unresolved) and .986 for $k = 2$ (one lower-regret and three unresolved comparisons). These intervals do not establish equivalence. Relative to

full closure, which also inserts every fitted signal into the action LP, the decoupled design improves all four cells for each k , with ratios .802 and .862. Thus the earlier closure penalty came from expanded action-signal constraints, not uniform state coverage. The $k = 2$ design also improves all four cells against a pure hitting gate (ratio .829), uniform verified TS (.581), and calibration greedy (.464). The adverse $k = 1$ result is retained: it is worse than uniform verified TS in all four cells (ratio 1.143) and unresolved against pure hitting (.990).

The study comprises 24,576 trajectories over 48 policy–cell groups. Independent checks verify the full reliability class, unchanged two-signal action design, hard caps, acquisition ledgers, and regret rescoring; the largest rescore error is 1.38×10^{-11} . Calibration-informed policies share 591 trusted labels and 18,912 archived judge outputs, whereas audit-only TS does not. Monetary, token, latency, and labeling-time prices are unavailable and are not imputed. Protected confirmation rows were skipped before decoding. This is development evidence selected after diagnosing the initial Panel32 failure, not fresh confirmation. It neither learns the audit gate nor establishes population-level coverage.

Calibration uncertainty. An aggregate-only calibration analysis sharpens this boundary. Simultaneous two-sided Clopper–Pearson intervals use 95% familywise coverage over all 224 pooled and state-subset action means considered by the selection family. None of the four fitted targets certifies a unique best action: their point-estimate margins are .00779, .000216, .00919, 0, whereas the corresponding lower-bound versus strongest-competitor-upper-bound margins are $-.108$, $-.156$, $-.0959$, $-.134$. A conservative nearest-candidate Hoeffding calculation, holding these point gaps fixed, requires respectively 300,053, 391,088,570, and 215,323 observations per controlled quantity; the zero-gap target cannot certify at any finite size under this rule. These are sufficient sizes, not power estimates or collection recommendations. Failure to certify does not show that a selected action is wrong, but it prevents this aggregate evidence from supplying a useful δ in Corollary 1; exact dictionary coverage is stronger still. The calculation assumes independent calibration requests within each action and subset and does not cover dictionary misspecification or population shift.

Reserved-data provenance. Reserved groups are excluded from the reported development fitting, policy selection, and performance evaluation, but their access histories differ. Retained JudgeBench loaders skip the 106 reserved rows before decoding. RAGTruth reserved labels were processed by upstream preparation and included in aggregate summaries, without use in the reported development comparisons. The LPHB confirmation rows were exposed and subsequently quarantined. SummEval reserved rows were decoded during processing but not used for development. These groups therefore do not constitute a uniformly untouched confirmation pool. No study above is reclassified as independent confirmation on the basis of those reserved groups.

C.8 Numerical checks of the theoretical identities

These checks assess the implementation of algebraic identities, optimization programs, and trajectory accounting. They supplement the proofs in Appendix A and are not an independent proof review or evidence of practical performance. Table 10 groups the checks by the mathematical claim they assess. Full generators, tolerances, and outputs are supplied with the reproducibility code.

Trajectory reconstruction. The adaptive study records every action and every accepted outcome. Independent reconstruction of all 36,864 primary trajectories reproduces cumulative regret and terminal likelihoods to 6.14×10^{-12} and 8.53×10^{-14} , respectively. The 98,304 supplementary trajectories use the same reconstruction. Timing checks ensure that actions precede the current state, audit decisions follow it, and every path respects the hard cap. These checks validate accounting rather than the accuracy of a learned outcome model.

Table 10: Numerical verification by scientific purpose. Errors are maximum absolute discrepancies unless the entry specifies a ratio or a violation count. Identity checks and synthetic theorem diagnostics are distinct from the policy comparisons reported in Section 6.

Identity or guarantee	Evaluation systems	Observed discrepancy or check
Product separation	180 random dictionaries; 40 coordinate cases	Minimum normalized margin .9999999999999989; coordinate error 1.74×10^{-18}
Deterministic coverage	1,440 LP cases from 240 set systems	Hitting error 8.88×10^{-16} ; coverage violation 2.22×10^{-16}
Bernoulli testing bounds	2,000 separation cases; 100,000 KL and 251,001 Hellinger comparisons	No inequality violation; observed separation at least twice the conservative threshold
Causal acquisition	6,300 stationary programs; 5,000 occupancy programs; 45,157 adaptive histories	Acceptance-law error 3.34×10^{-16} ; causal cost error 1.14×10^{-13}
Finite-horizon accounting	1,000 joint designs; independent forward recursion	Design-length error 4.27×10^{-14} ; regret error 2.85×10^{-14}
Adaptive information floor	144 development systems and measured adaptive laws	Minimum information divided by εI_0 equals one
Diffuse reliability	44 hitting LPs; 210,000 KL comparisons; 19,200 decoder trajectories	Hitting error 8.88×10^{-16} ; no five-standard-deviation upper-bound violation
Misspecification bounds	100,000 random systems; 10,000 flattening cases; 72,000 decoder trajectories	No constant violation; target-preservation error 1.11×10^{-16}
Finite-summary construction	Tree depths through 12; moment orders through eight	8,188 adaptive paths checked; zero requested-moment discrepancies on 29,520 points

Optimization and acquisition checks. Additional product-design comparisons verify the exact $\chi(\mathbf{p})$ cost on 1,000 cases and the prior-mixture and fixed-design dual identities on 500 cases each. The largest prior-mixture error is 9.02×10^{-11} , and the largest fixed-design inequality violation is 1.19×10^{-15} . The finite-horizon design/exploitation partition has error 1.57×10^{-13} . These independent formulations detect coding or normalization errors without replacing the analytical arguments.

D Detailed comparisons with prior work

Verified and prediction-powered bandits. PP-LUCB studies best-arm identification with LLM judges, selective audits, and residual-aware allocation [29]. PPAI estimates prediction weights under adaptive sampling [42]. Limited-verification bandits quantify the value of trusted observations under attacks [19]. These works establish that auditing and correction are not new primitives. Our objective is cumulative true-reward regret under a hard audit budget, and our new parameter is an unknown conditional-reliability class. MLA-UCB uses machine-learning-generated surrogate rewards for cumulative-regret reduction [37], while confidence-budget matching treats optional costly observations in sequential learning [15]. They provide controls for judge-assisted and audit-scarce regimes, respectively. Neither models a latent post-action reliability set coupled to kernel action geometry.

Kernel and multi-fidelity optimization. Auxiliary-feedback IGP-UCB uses parameterized correlated feedback [23]. PA-GP-UCB models prediction and reward jointly and observes both during online selection [39]. Robust multi-fidelity BO guards against unreliable low-fidelity sources [22]. Kernel experimental design and misspecified GP bandits already control target approximation error [13, 14]. ASB permits post-surrogate acquisition: its fixed-comparison theorem gives PRE/POST variance-optimal allocations, its main guarantees cover finite policy classes and a hard-budget linear construction, and its learned-refinement proposition charges residual-model error [30]. Chandak et al. optimize context-conditioned collection policies and use exact multi-proposal rejection sampling to reuse adaptively collected data [26]. These results establish post-surrogate timing, residual-based allocation, kernel modeling, and rejection sampling as existing components.

Controlled sensing and experimental design. Controlled multihypothesis testing optimizes Chernoff information over sensing controls, and adaptive causal controls can improve on open-loop designs with three or more hypotheses [1, 18]. Bayesian decision-region determination already selects tests only to resolve the action-relevant equivalence class, including correlated noisy tests [2, 4]. Transductive linear best-arm identification separates measurement and decision sets through experimental design [7]; recent hybrid-feedback GLB identification optimizes a joint query space and includes heterogeneous acquisition costs [46]. These methods supply decision-focused likelihood tests and query designs with exponentially decreasing error. Proposition 4 composes their predictable adaptive allocation principle with a fixed robust information floor and the exact action-before-state cost identity. The composition is scoped to this protocol and is not a new generic adaptive-testing principle. Our reliability-action factorization, adaptive converse, and deterministic coverage formula specialize this approach to unknown reliable-state sets; they do not characterize general adaptive optimality.

Mixture testing and the diffuse regime. Adaptive mixture testing already shows that repeated allocation can overcome rare-component dilution when the same latent item can be queried repeatedly [8]. Our diffuse construction instead makes state repetitions negligible over the decision horizon. After the negligible-collision coupling, marginalizing the reliability set gives a standard d -armed Bernoulli identification problem with gap $\rho\Delta$; classical best-arm lower bounds and fixed-budget conversions already yield the $d/(\rho^2\Delta^2)$ scale up to logarithms [3, 32]. The resulting comparison shows that rank and hitting number alone are insufficient to determine audit complexity.

Partial monitoring. Kernelized linear partial monitoring already supplies information-directed sequential policies, observability-based regret classifications, and RKHS observation operators [11]. Its contextual extension also plans against a known context distribution, so a sequential information ratio, context-weighted observability coefficient, or kernelization is not new here. A finite verified-feedback dictionary can be unfolded into a constrained partial-monitoring game over contingent audit maps; our protocol-specific remainder is the hard pathwise audit cap together with the exact action-before-state compilation cost. This is why the paper states a structured finite-family factorization rather than a new generic sequential complexity.

Nonparametric conditional feedback. Nonparametric models provide another way to represent conditional feedback laws. CME-UCB already estimates conditional mean embeddings while sequentially optimizing an unknown conditional expectation [12]; distributionally-aware kernel bandits model probability kernels through kernel mean embeddings and prove UCB guarantees and lower bounds [20]; and active cost-aware streaming labeling chooses expensive labels after inputs arrive and gives matching discrete and continuous-RKHS rates [21]. These models do not impose the action-before-state protocol with a hard audit cap. Extending their estimators to that protocol would require accounting for the acquisition costs in Proposition 3.

Post-action contexts. Post-serving-context bandits also observe informative features only after action selection; `poLinUCB` learns a pre-to-post context map and attains tight linear-contextual regret [24]. The trusted outcome here is optional and subject to a hard pathwise budget. The row-max formula is the exact domination constant obtained when classical rejection sampling is restricted to proposals of the form $q(a)\mu(s)$; it quantifies acquisition cost under this proposal restriction.

Cost-sensitive acquisition. Costly-observation bandits provide complementary cost-sensitive guarantees. Schur et al. give two-phase GP algorithms, cost-adjusted information-gain complexities, and matching finite-arm rates for observation-priced objectives [44]. Gong and Zhang establish hard free-observation-budget phase transitions for contextual bandits with paid observations [33]. Harding et al. choose a costly source before an unobserved group is revealed and obtain matching budgeted minimax rates by effective-sample-size design and post-stratification [34]; Reyzin et al. give tight cost-adaptive rates for heterogeneous expert queries [43]. Progressive-feedback bandits quantify early-surrogate value by conditional mutual information, but eventually observe the primary outcome for every play [35]. Cost-aware acquisition, hard feedback budgets, and rejection-sampled reuse are therefore prior methods. Harding et al. do not select labels after seeing the group or study sequential decision regret, while Chandak et al. do not study hard audit budgets, sampling regret, or latent reliability. The remaining scoped distinction is the finite-dictionary reliability–kernel theorem and its adaptive converse; Schur et al. explicitly leave general-kernel optimality open, so Proposition 3 is not presented as a matching kernel converse.

LLM routing with restricted feedback. Recent LLM-routing studies consider related feedback restrictions. Wei et al. give linear-contextual upper and lower bounds under a limited reward-feedback budget [45]. Sridhar et al. combine context-dependent graph feedback with ML surrogate rewards and use expert aggregation to recover a reward-only guarantee when those surrogates are misspecified [31]. Unlike both settings, trusted outcomes here are optional after observing a post-action state, and the unknown object is the state-conditional reliability law.

AI evaluation. ARES uses prediction-powered inference for RAG evaluation [25]. HbBoPs applies deep-kernel Bayesian optimization to prompt selection [28]. NAOD designs preference queries while accounting for judge nuisance [38]. Our model is narrower: fixed configurations, i.i.d. requests, bounded outcomes, and a fixed kernel. It excludes contextual routing, nonstationarity, and within-agent-step credit assignment.

Comparator	Established capability	Distinction tested here
ASB [30]	post-surrogate acquisition, learned residual error, hard budgets	latent reliability dictionary and kernel decision geometry
Multi-rejection [26]	exact reuse under multiple collection policies	hard-budget regret and latent reliability learning
Cost-aware GP [44]	optional labels and cost-adjusted kernel information gain	post-state reliability and hard pathwise audits
Controlled sensing [1, 18]	adaptive control and likelihood testing	latent reliability cost under action-before-state timing
Auxiliary IGP-UCB [23]	kernelized parameterized auxiliary feedback	audit scarcity and latent reliable-state dictionary
PP-LUCB [29]	selective audits for identification	cumulative regret and kernel action geometry
Robust / misspecified BAI [6, 16]	contaminated or structurally misspecified identification	separate target and audit-law error costs
Audit-only / CBM [15]	valid without an informative judge	benchmark that proposed feedback must improve

Table 11: Comparison of feedback protocols, guarantees, and modeling assumptions.