EXPLORATION UNDER HETEROGENEOUS FEEDBACK: EXACT GEOMETRY OF ORDERED ADMISSION BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient exploration depends on where informative feedback is available, not just which actions offer high utility. We study ordered admission bandits with increasing known utility scales, decreasing unknown survival probabilities, and a strictly unimodal product. Under heterogeneous feedback, the optimum's neighbors need not be the best exploration sources. We reduce the structured-bandit allocation program exactly to a left-prefix constraint and a right-neighbor constraint. The resulting closed form selects the left source by its regret-to-information ratio, recovering neighbor exploration under homogeneous releases and proving an unbounded penalty for enforcing locality under heterogeneous releases. For regular single-candidate count/randomized-response feedback, we develop Certified Oracle Tracking (COT) and prove that suitable fixed tunings approach the lower-bound constant arbitrarily closely. Experiments with 10, 20, and 50 arms demonstrate source migration and competition. On frozen real-data gates, changing only the release channel isolates the value of information location: COT reduces weak-feedback regret by 73.5% against a matched neighbor-oracle ablation, while their full-count trajectories coincide. These results characterize when exploration should move beyond neighbors and provide a certified policy for using that information.

1 INTRODUCTION

Efficient sequential learning depends on both action utility and feedback quality. In ordered admission problems, actions have increasing utility scales and decreasing probabilities of passing a fixed gate; we assume their product is strictly unimodal. Under homogeneous feedback, classical unimodal bandit results make the optimum's neighbors sufficient for efficient exploration (Combes & Proutiere, 2014; Combes et al., 2014). This local geometry provides a natural basis for learning from admission observations.

However, feedback quality can vary across actions, and a promising action may provide little information about its optimality. A distant action with more precise feedback can instead exclude a competing optimum: survival order couples the prefix, allowing observations from one arm to constrain others without observing them directly. General structured-bandit methods optimize exploration over alternative parameters (Combes et al., 2017; Degenne et al., 2020), but do not directly resolve which actions are worth querying in this ordered, heterogeneous setting.

We study ordered admission bandits with known, arm-dependent release channels. The problem presents three challenges. First, survival order couples the alternatives that observations must exclude. Second, heterogeneous releases make information quality compete with the regret cost of querying an arm. Third, the learner must identify the useful information source before exploiting the resulting allocation. Our goal is to characterize this exploration geometry and turn it into an executable learning rule.

We reduce the structured allocation program to two information constraints and obtain a closed-form oracle. For regular single-candidate Bernoulli releases, we develop Certified Oracle Tracking (COT), combining oracle quotas, structured estimation, and a likelihood certificate for exploitation. Synthetic families and controlled release interventions on frozen gates connect the predicted allocation to observed learning behavior. Our main contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- We characterize the exact allocation through a left-prefix constraint and a right-neighbor constraint. An explicit regret-to-information ratio selects the left source, recovering neighbor exploration under homogeneous feedback and establishing an unbounded locality penalty under heterogeneous feedback.
- We construct COT and prove that, under regularity conditions, suitable fixed tunings approach the lower-bound constant arbitrarily closely.
- We test source migration in 10-, 20-, and 50-arm families and on frozen real-data gates. Under weak feedback, COT reduces regret by 73.5% against a matched neighbor-oracle ablation; under full counts, their trajectories coincide.

2 Related Work

Ordered rewards and locality. Rate adaptation couples increasing transmission scales with decreasing success probabilities (Combes et al., 2014); OSUB exploits unimodal rewards (Combes & Proutiere, 2014), as does Thompson sampling in Bernoulli rank-one problems (Trinh et al., 2020). Fixed-confidence unimodal identification also finds sparse allocations around the optimum (Poiani et al., 2025), building on identification tracking (Garivier & Kaufmann, 2016). Our homogeneous corollary recovers neighbor exploration. Arm-dependent observation laws change that result: survival order lets a remote action exclude a poorly observed competitor, and its regret cost determines whether that information is worth collecting.

Structured exploration. OSSB and structure-adaptive bandits optimize allocations over constrained alternatives (Combes et al., 2017; Degenne et al., 2020); DUSA gives a dual approach for convex distributional constraints (Van Parys & Golrezaei, 2024). Informative actions need not have promising rewards: optimism can be arbitrarily suboptimal in linear bandits (Lattimore & Szepesvári, 2017), CROP queries actions that eliminate optimistic hypotheses (Jun & Zhang, 2020), and information-directed sampling balances regret against information (Kirschner et al., 2021). Our result resolves this tradeoff in closed form: a single regret-to-information ratio selects the left source, independently of the right constraint. COT then uses that explicit oracle online.

Released observations. Corrupt bandits allow known arm-dependent feedback maps (Gajane et al., 2018); their independent-arm requirements do not capture the coupled survival prefix. Threshold bandits vary censoring and payoffs through a supplied threshold (Abernethy et al., 2016), whereas our gate is fixed and its receipt varies. Side-observation models measure other arms (Atsidakou et al., 2022); here each observation concerns its selected arm and order carries information to competitors. Local-privacy analyses quantify information contraction (Duchi et al., 2013; Kairouz et al., 2014). Theorem 1 shows how unequal contractions relocate its cheapest source. Appendix B compares the closest assumptions; Appendix I relates the admission interface to poisoning and retraining studies.

3 Method

Our method separates the information needed to reject an alternative from the cost of collecting it. We define the ordered model, solve its allocation program, and use that solution in a certified learning policy.

3.1 Problem Statement and Threat Model

Access and capabilities. A submitter or auditor adaptively selects among K fixed candidate generators using past receipts. It knows the action order, scales r_k, batch size τ and channels H_k, and observes the selected arm's released statistic. Scores, thresholds, gate parameters and downstream losses remain private. The learner is given the ordered class below; generator-specific parametric relations are not supplied. It controls the action sequence. The gate, generators and releases are stationary, with independent candidates conditional on the chosen arm. Figure 26 shows the interface.

Ordered objective. The private gate's score and acceptance threshold stay fixed. Arm $k \in [K]$ generates independent candidates with admission indicator $Y \sim \mathrm{Bernoulli}(p_k)$. For the ordered

theory we assume $0 < r_1 < \cdots < r_K$ and $1 > p_1 \geq \cdots \geq p_K > 0$, with $\mu_k = r_k p_k$ *strictly increasing* up to a unique maximizer s and strictly decreasing afterwards. Write Θ_r for this class and $\Delta_k = \mu_s - \mu_k$. Unimodality is an assumption on the product. The scales r are supplied before online learning. They can represent specified utilities or externally calibrated conditional impacts; the latter interpretation is tested separately in Section 4.4.

Released law. One decision submits τ candidates from one known arm. The count is $C \sim \mathrm{Binomial}(\tau, p_k)$; a known, possibly arm-dependent channel H_k releases $Z \sim Q^H_{k,p_k}$. For B decisions and $N_k(B)$ selections of arm k,

$$\mathrm{Reg}(B) = \tau \sum_k \Delta_k \mathbb{E}_p N_k(B), \qquad d_k(p,q) = \mathrm{KL}(Q^H_{k,p} \| Q^H_{k,q}). \tag{1}$$

This is candidate-level pseudo-regret for arm-resolved feedback.

Established lower-bound framework. Let $\mathrm{Alt}(p) = \{q \in \Theta_r : s(q) \neq s(p),\, Q^H_{s,p_s} = Q^H_{s,q_s}\}$. The structured change-of-measure bound (Lai & Robbins, 1985; Burnetas & Katehakis, 1996; Graves & Lai, 1997; Combes et al., 2017) gives, for every policy with $\mathrm{Reg}(B) = o(B^\alpha)$ on every instance and every $\alpha > 0$,

$$\liminf_{B\to\infty} \frac{\mathrm{Reg}(B)}{\log B} \geq C_H(p), \qquad C_H(p) = \inf_{\eta \geq 0} \left\{ \tau \sum_{k \neq s} \Delta_k \eta_k : \inf_{q \in \mathrm{Alt}(p)} \sum_{k \neq s} \eta_k d_k(p_k, q_k) \geq 1 \right\}. \tag{2}$$

Here η_k is a batch allocation per $\log B$. Appendix A.1 derives this bound for the released observation law. The program specifies the information needed for learning; the next subsection solves its constraints.

3.2 Exact exploration geometry

The following channel condition lets us identify exactly which observations satisfy the allocation constraints. It covers the count and monotone releases used in our experiments.

Assumption 1 (Order-preserving divergence)**.** *For every arm, $p \mapsto Q^H_{k,p}$ is injective; $d_k(p,q)$ is finite and continuous on $(0,1)^2$, zero only at $p = q$, and nondecreasing in q for $q \geq p$.*

This includes full counts, counts after binary randomized response with known $\epsilon_k > 0$, an any-survivor bit, and a randomly selected candidate's admission bit. Channels may differ across arms. Injectivity forces $q_s = p_s$ in the confusing alternatives.

For $s > 1$, put $t_- = \mu_s / r_{s-1}$. A left alternative exists exactly when $t_- < 1$. For such a left side define

$$a_i = d_i\big(p_i, \max\{p_i, t_-\}\big), \quad i < s. \tag{3}$$

For $s < K$ put $t_+ = \mu_s / r_{s+1}$ and $b_+ = d_{s+1}(p_{s+1}, t_+)$. The right alternative always exists when $s < K$, because $p_{s+1} < t_+ < p_s$.

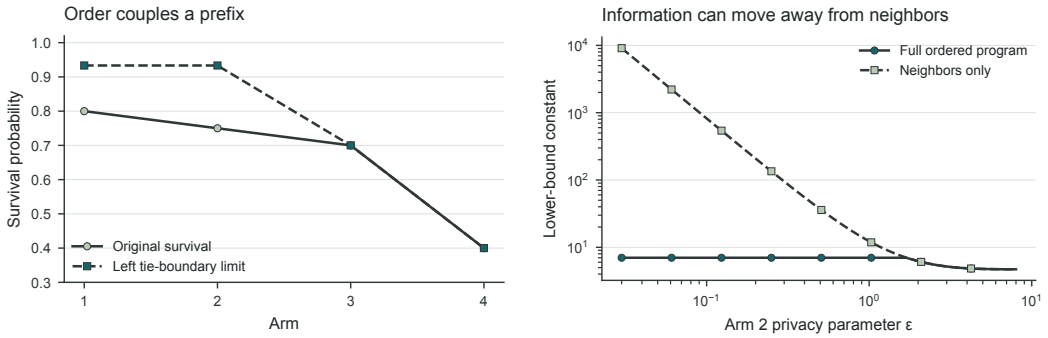
Theorem 1 (Two information constraints)**.** *Under Assumption 1, the feasible allocations in Equation* (2) *are exactly those satisfying*

$$\sum_{i<s} a_i \eta_i \geq 1 \quad \textit{if } s > 1,\ t_- < 1; \qquad b_+ \eta_{s+1} \geq 1 \quad \textit{if } s < K. \tag{4}$$

Consequently, with an absent side contributing zero,

$$C_H(p) = \min_{i<s : a_i > 0} \frac{\tau \Delta_i}{a_i} + \frac{\tau \Delta_{s+1}}{b_+}. \tag{5}$$

The first term is present only when the left alternative exists. One minimizing allocation samples a single minimizing left-prefix arm and the right neighbor.



(a) **(b)**

Figure 1: **Why useful exploration can move away from a neighbor.** (a) A left-optimum alternative must raise the survival prefix; the dashed line is the least-informative tie boundary. (b) Weakening arm 2's RR channel separates the exact and neighbor-restricted constants while all utilities remain fixed. The four-arm construction is specified in the text.

Why the prefix is sufficient. If an alternative optimum lies left of s, strict unimodality requires $r_{s-1}q_{s-1} > \mu_s$, hence $q_i > t_-$ for every $i < s$. The coordinatewise least informative limit is $q_i^- = \max(p_i, t_-)$ for $i < s$, with all remaining coordinates fixed. It is approached by valid alternatives with optimum $s-1$. The corresponding right limit changes only q_{s+1} to t_+. Every alternative is on one of these two sides. Continuity gives necessity; monotonicity of divergence gives sufficiency. The constraints have disjoint supports, so the linear program allocates the left-side information to the smallest cost-to-information ratio. Appendix A.2 gives the full proof, including feasibility of the boundary approximations.

Corollary 1 (When neighbors remain sufficient)**.** *Suppose additionally that $d_i = d$ for every arm and $d(p,t)$ is nonincreasing in $p < t$. Then the minimizing left-prefix arm in Equation* (5) *is $s-1$. Homogeneous full-count, randomized-count, any-survivor and sampled-bit channels satisfy this condition.*

Earlier left arms have both a larger utility gap and no more information about t_-. Thus the familiar neighbor constant is recovered even when an isolated left-neighbor perturbation violates survival order. This agrees with the count-feedback setting of rate adaptation (Combes et al., 2014); it also identifies the condition needed to extend that conclusion to released feedback.

An unbounded price of neighbor-restricted exploration. Take $r = (.5, .75, 1, 1.25)$ and $p = (.8, .75, .7, .4)$, so $s = 3$ and $t_- = 14/15$. Let $\tau = 20$, release full counts on arms 1, 3 and 4, and randomized-response counts with parameter ϵ on arm 2. With homogeneous counts, $C_H = 4.7094$ and the left information comes from arm 2. At $\epsilon = .1$, the minimizing left arm is instead arm 1 and $C_H = 6.9952$. Constraining non-neighbor allocations to zero gives 822.2800. As $\epsilon \downarrow 0$, arm 2's divergence is $\Theta(\epsilon^2)$, whereas arm 1's divergence and cost stay fixed. The ratio therefore diverges. The restricted program lower-bounds every uniformly efficient policy whose non-neighbor expected counts are $o(\log B)$. Figure 1 connects the forced prefix change to this divergence in exploration cost.

3.3 From the allocation oracle to a learning policy

For regular single-candidate count/RR releases, the closed form supports an executable tracking rule. We use the compact model $\mathcal{P}$: the union of the weakly unimodal mode polytopes within $[\ell, u]^K$, with survival order imposed in each polytope. The true parameter lies strictly inside its mode. Count and RR bits have Bernoulli means $\theta_k = \beta_k + (1-2\beta_k)p_k$, where $\beta_k = 0$ for counts and $\beta_k = (1+e^{\epsilon_k})^{-1}$ for RR. The known reward map is $r_k(\theta_k - \beta_k)/(1-2\beta_k)$.

Certified Oracle Tracking. COT initializes every arm, fits an ordered unimodal estimate, and evaluates $c(\widehat{p})$ using Theorem 1. The oracle quotas are $(1+\gamma)c_i(\widehat{p})\log t$. A counter L records non-exploitation decisions: COT samples a least-played arm if $\min_i N_i \le \lambda L$, and otherwise samples the arm with smallest N_i/c_i among unmet quotas. Once quotas are met, a likelihood test must also

Algorithm 1 Certified Oracle Tracking (COT), $\tau = 1$

```
Initialize each arm once; set non-exploitation count L = 0.
for decisions t = K + 1, K + 2, ... do
    Fit p̂ in its strict mode; compute c(p̂) by the closed form.
    Set j = arg max_i r_i p̂_i and b = log max(3, t).
    if N_i ≥ (1 + γ) c_i b for all i and Z_j > (1 + γ/2) b then
        Play j (exploitation).
    else
        L ← L + 1; let h be a least-sampled arm.
        if N_h ≤ λL then
            Play h (estimation).
        else if some oracle quota is unmet then
            Play arg min_{i:c_i>0} N_i/c_i (allocation exploration).
        else
            Play j (certification).
        end if
    end if
    Update counts and empirical means using the released bit.
end for
```

exclude all parameters at which the estimated leader is suboptimal. If this certificate is unresolved, COT samples the leader to refine it. Two scalar convex minimizations compute the test. Specifically, with empirical released means z_i,

$$Z_j(t) = \inf_{q \in \mathcal{R}_j^- \cup \mathcal{R}_j^+} \sum_i N_i(t) D_{\mathrm{B}}(z_i, g_i(q_i)). \tag{6}$$

The left set requires $q_i \geq (r_j/r_{j-1})q_j$ for $i < j$; the right set requires $q_{j+1} \geq (r_j/r_{j+1})q_j$. These relaxations contain every wrong-leader parameter in the model. Appendix A.5 gives the estimator, scalar minimizations and all tie rules. Lemmas 1–4 separate exploitation errors, estimation, certification and regret accounting.

Corollary 2 (Near-attainment for regular Bernoulli releases)**.** *Let $\tau = 1$, with known count/RR channels and $\epsilon_k > 0$ for RR. Use $\mathcal{P}$ with fixed $0 < \ell < u < 1$ and known strict feasible points in its modes. Suppose $u > p_1 > \cdots > p_K > \ell$, utility is strictly unimodal with $1 < s < K$, $t_- < u$, and the minimizing left ratio is unique. Then the two-constraint constant on $\mathcal{P}$ remains $C_H(p)$, and for every $\delta > 0$ some fixed $0 < \lambda < 1/K$, $\gamma > 0$ satisfy*

$$\limsup_{B \to \infty} \frac{\mathrm{Reg}_{\mathrm{COT}(\lambda,\gamma)}(B)}{\log B} \leq C_H(p) + \delta. \tag{7}$$

The closed model admits boundary and tied parameters without changing either information infimum at the strict truth. Its unique cheapest prefix arm makes the normalized allocation locally continuous. The proof combines this regularity with Bernoulli concentration and the OSSB scheduling principle (Combes et al., 2017), explicitly accounting for structural estimation and the exploitation certificate. Fixed tuning adds exploration; the experiments report it alongside the exact allocation. Prefix-KL remains the separate finite-budget routing policy used for batched gate queries.

4 Experiments

We test one sequence of predictions: the oracle changes its source, a tracking policy can use that source, and release interventions expose the same geometry on actual gates. All online policies receive only known scales, channels and released observations. True probabilities, utility gaps and theoretical allocations are evaluator-only. Intervals are two-sided 95% Student-t intervals. Synthetic comparisons use paired streams; real-data intervals use split means after averaging streams and fixed thresholds.

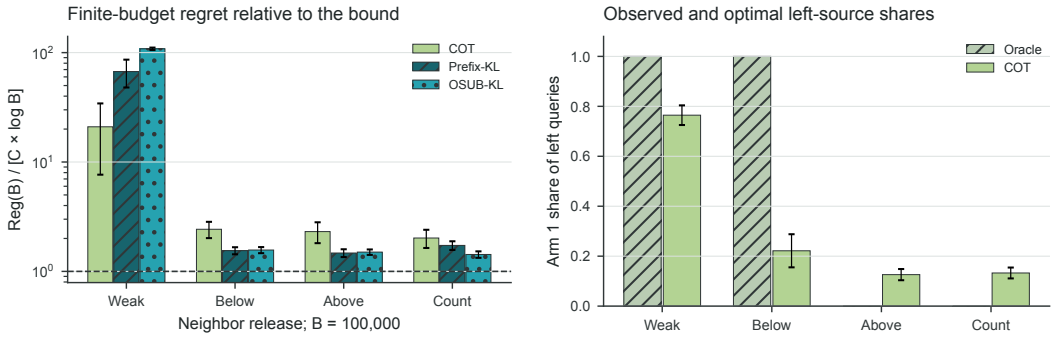


Figure 2: **The closed-form oracle enters online learning.** (a) Regret normalized by $C_H \log B$ at $B = 100{,}000$, 48 streams. The dashed line is the asymptotic benchmark. (b) Oracle and COT arm-1 shares among left queries. Error bars are 95% intervals. COT uses fixed $(K\lambda, \gamma) = (.10, .25)$: quotas are $1.25c_i \log B$, with additional estimation samples.

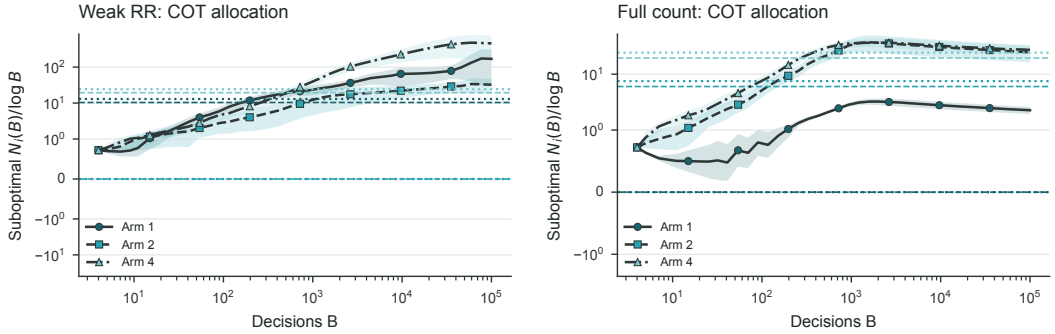


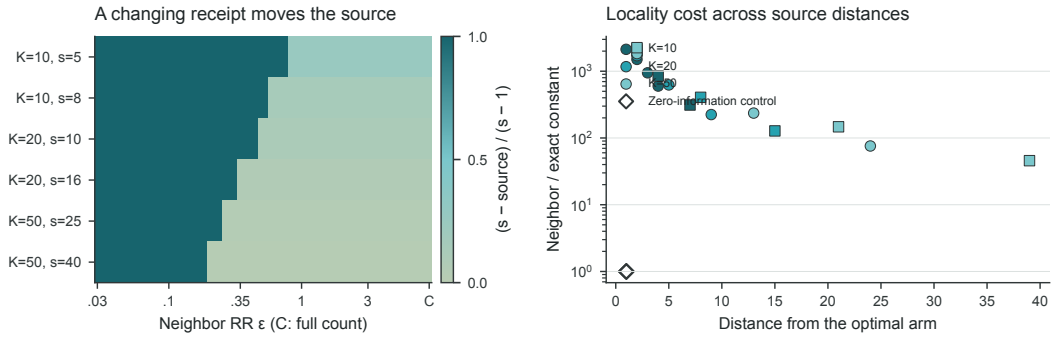
Figure 3: **Tracking counts and oracle quotas have distinct finite-budget behavior.** COT suboptimal $N_i(B)/\log B$ under (a) weak RR and (b) full counts, 48 streams with 95% intervals. Dashed lines show exact c_i; dotted lines show $1.25c_i$. The scale is linear near zero and logarithmic above one. Estimation and early source errors leave samples on arms with zero oracle allocation.

4.1 Oracle-driven tracking under single-candidate releases

We use the four-arm profile of Section 3.2 with $\tau = 1$. The source switches at $\epsilon_2^\star = 1.70058$. Four fixed conditions cover weak feedback (.1), both sides of the switch ($.85\epsilon_2^\star$, $1.15\epsilon_2^\star$), and full counts. Each has 48 paired streams to $B = 100{,}000$, recorded at 40 logarithmically spaced horizons. The default COT tuning is $(K\lambda, \gamma) = (.10, .25)$; the fixed alternatives $(.04, .10)$ and $(.20, .50)$ are reported in full.

Under weak feedback, COT reduces mean regret to 1690.3 ± 1073.5, versus 5405.8 ± 1539.5 for Prefix-KL and 8729.2 ± 225.5 for OSUB-KL. With full counts, the respective values are 109.5 ± 20.7, 93.7 ± 8.5, and 77.4 ± 5.3. Figure 2 relates these costs to the exact constant and realized source shares. At weak feedback, COT's $\mathrm{Reg}(B)/\log B = 146.8$ remains well above $C_H = 7.00$; learning the poorly observed coordinates dominates the transient cost. The strongest fixed exploration setting reduces that transient here, while the default is better under full counts. All tunings, horizon curves and suboptimal $N_i(B)/\log B$ are in Appendix C.

The allocation trajectories explain the difference (Figure 3). Weak neighbor receipts make its survival estimate slow to stabilize, so early oracle choices can allocate many samples there before the remote source dominates. Under counts, the estimated geometry is easier to resolve, but estimation still samples the remote arm. The extra count is visible against its zero oracle target. Algorithm 1 makes these two costs explicit: learning the allocation and satisfying it are separate operations.



(a) **(b)**

Figure 4: **Source migration extends beyond four arms.** (a) Optimal source distance, normalized by prefix length, across six far-probe profiles; only neighbor feedback varies. C denotes full count. (b) Neighbor/exact constants against source distance at $\epsilon_{s-1} = .03$, across all informative near/middle/far profiles. Circles and squares indicate the two peak positions. Zero-information controls remain at ratio one.

4.2 Information geometry with 10, 20 and 50 arms

The mechanism persists with more arms and competing sources. We generate 24 profiles from three arm counts, two peak positions $s \in \{\text{round}(.5K), \text{round}(.8K)\}$, and near, middle, far or zero-information probes. An explicit survival/scale rule guarantees strict order and a single peak. Two prefix probes have better channels than the remaining left arms; sweeping only the neighbor's RR parameter gives 1,584 exact allocation calculations. No profile is selected by a policy outcome. Appendix D gives the generator and all grids.

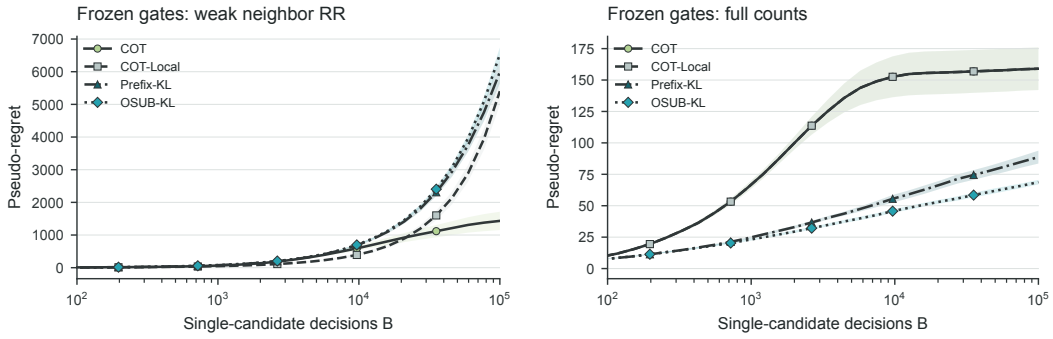
Improving the neighbor's receipt moves the source toward the optimum (Figure 4); distance and feedback quality jointly determine the locality penalty. In the negative controls, every non-neighbor prefix coefficient is zero, so even full counts cannot replace the neighbor. Six online representatives use far probes, $s = \text{round}(.8K)$, and weak/count neighbor feedback for each K, with 24 streams to 50,000 decisions. COT remains in non-exploitation phases at $K = 20, 50$. For the $K = 50$ weak case, the true $1.25c_i \log B$ quotas already total 193,626 samples, exceeding the budget before estimation costs enter. Appendix D retains every representative's regret and allocation curves, including this unresolved transient (Figure 20).

4.3 Controlled release interventions on actual gates

We use Adult, Bank Marketing, Spambase and Phishing Websites (Becker & Kohavi, 1996; Moro et al., 2014; Hopkins et al., 1999; Mohammad & McCluskey, 2012). Each of five splits assigns 50% of records to gate training, 15% to threshold calibration, 20% to a candidate pool and 15% to an unused test partition. A logistic classifier defines submitted-label log loss; thresholds are the 70th, 85th and 95th percentiles of clean calibration scores. This gives 60 frozen gates. Arm a flips a freshly sampled candidate's label with probability a. Each submission uses the gate decision on that feature row and submitted label. The evaluator uses the finite-pool admission rates $p_{\text{clean}}, p_{\text{flip}}$ to compute regret and the theoretical allocation via

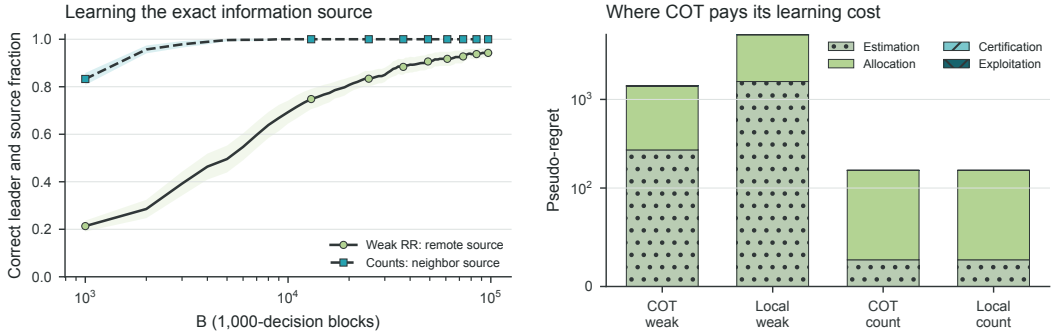
$$p(a) = (1 - a)p_{\text{clean}} + ap_{\text{flip}}. \tag{8}$$

A controlled geometry intervention. We fix four intensities $a = (.35, .40, .45, 1)$ and specified utilities $r = (.5, .75, 1, 1.25)$ across every gate. They are designed to expose the prefix constraint; r is not a measured poisoning impact. All 60 resulting profiles have strictly decreasing survival and strictly unimodal utility with optimum $s = 3$. Arms 1, 3 and 4 release counts. Only arm 2's channel changes: RR with $\epsilon_2 \in \{.1, 1, 2\}$, or full count. Thus the candidate laws, gate and utility optimum are fixed across release interventions. The release intervention isolates where information is available while keeping the admission problem unchanged. We first test COT at $\tau = 1$, then use batched queries to examine Prefix-KL's finite-budget routing.



(a) **(b)**

Figure 5: **The guaranteed policy uses remote information on frozen gates.** Single-candidate regret under (a) weak neighbor RR and (b) full counts. COT-Local changes only the left oracle support. Curves average 20 gate means, each based on 20 paired streams; shading shows 95% intervals across gates. COT and COT-Local coincide under counts.



(a) **(b)**

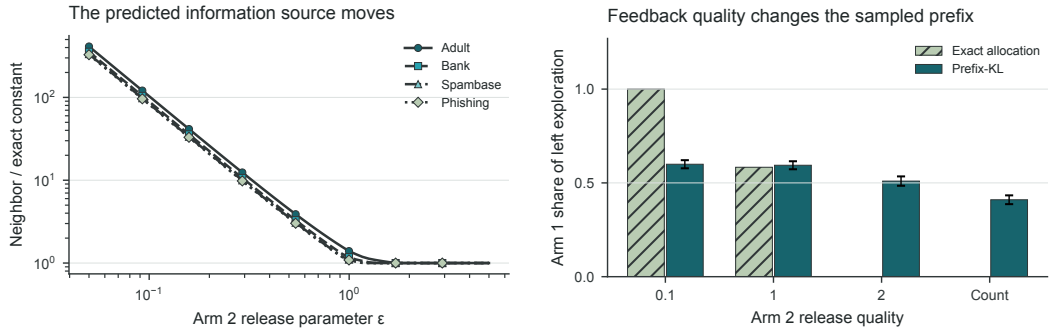
Figure 6: **Identifying the source separates learning cost from routing cost.** (a) Fraction of COT decisions with the correct leader and exact source in successive 1,000-decision blocks. (b) Regret by COT phase at $B = 100{,}000$; Local denotes COT-Local. Initialization contributes less than one and is omitted. The scale is linear below 100 and logarithmic above it.

4.3.1 COT ON THE SAME FROZEN GATES

COT uses the same frozen records and action ladder at $\tau = 1$, the setting of Corollary 2. We retain all four datasets and five splits at the middle threshold (85th percentile), giving 20 gates. Weak RR ($\epsilon_2 = .1$) and full counts each use 20 paired streams per policy to 100,000 decisions, with the unchanged default COT tuning. COT-Local is a matched ablation: it restricts only the left oracle to the neighbor; estimation, certification and scheduling remain identical. The comparison therefore isolates the value of the oracle's source choice.

Under weak RR (Figure 5), COT reduces mean regret from COT-Local's 5403.1 ± 422.8 to 1432.5 ± 280.5, a 73.5% reduction; the paired difference is 3970.6 ± 528.2. Prefix-KL and OSUB-KL incur 5991.4 ± 350.0 and 6536.8 ± 207.2. The source is learned during the run: in the final 1,000 decisions, COT identifies both the optimum and its exact remote source on 94.2% of decisions and uses certified exploitation on 93.9% (Figure 6). The prespecified sustained-operation criterion is met by 375/400 weak-RR runs, with median onset 10,001 among those runs; 25 are horizon-censored. Appendix F defines this diagnostic and reports phase counts and allocation trajectories.

The phase decomposition explains the gain: remote routing cuts allocation regret from 3759.0 to 1132.0 and estimation regret from 1592.3 to 269.2, because fewer non-exploitation decisions also lower the estimation floor. Under full counts both oracles select the neighbor and produce identical trajectories. Their regret is 159.0 ± 17.0, versus 88.7 ± 5.1 for Prefix-KL and 68.8 ± 1.8 for OSUB-KL. Here the allocation and estimation overhead outweighs any routing gain.



(a) **(b)**

Figure 7: **Predicted and observed information location on 60 actual gates.** (a) Neighbor-restricted/exact lower-bound constants, averaged over five splits and three thresholds per dataset. (b) Arm 1's share of left exploration: $\eta_1/(\eta_1 + \eta_2)$ for the exact program and $N_1/(N_1 + N_2)$ for Prefix-KL at 20,000 candidates, including forced exploration. Bars average 20 dataset–split units; intervals describe their variability.

4.3.2 Batched information routing

Prefix-KL propagates survival confidence bounds through the order and queries the prefix arm supplying the decisive bound. We compare global KL-UCB (Garivier & Cappé, 2011; Cappé et al., 2013), line OSUB-KL and a propagation-only ablation on all 60 gates: $\tau = 20$, 20 paired streams and 20,000 candidates, for 19,200 runs. Appendix E specifies the policies and retains the six synthetic batched conditions, including strong-feedback costs.

Predicted source and realized allocation. The exact program selects arm 1 in all 60 environments at $\epsilon_2 = .1$, in 35 at $\epsilon_2 = 1$, and in none at $\epsilon_2 = 2$ or full count. At .1, the neighbor-restricted constant is 68.5–132.9 times the full constant (Figure 7). The executed Prefix-KL policy changes its mean arm-1 share of left queries from .599 to .410 between weak RR and full count (Figure 7). It therefore responds to channel quality but does not reproduce the exact limiting allocation: under full count, the program assigns all left exploration to arm 2.

Regret follows the value of remote information. At $\epsilon_2 = .1$, mean regrets over the 20 dataset–split units are 489.7 for Prefix-KL and 1407.9 for OSUB-KL. The four dataset-specific paired tests have Holm-adjusted $p < .001$ over all 16 quality–dataset contrasts. At $\epsilon_2 = 2$, the respective regrets are 101.0 and 71.0; with full counts they are 79.4 and 52.0. Residual remote sampling adds cost once the neighbor is informative. Figure 25 and Table 6 retain every release condition and baseline.

4.4 Information and downstream calibration

Source selection requires identifiable alternatives: even-batch parity can hide different optima, and identifiable one-bit releases can differ exponentially in information (Appendix A.3). Calibration tests the final link from admission utility to retraining effects. Matching the intervention size improves held-out damage-ranking correlation by $.600 \pm .290$ at equal audit budget, while calibrated policies have no general downstream advantage over random allocation (Appendix I). Thus learning where to explore and calibrating what the resulting utility predicts are distinct tasks.

5 Conclusion

Heterogeneous feedback can move useful exploration beyond neighbors. Two exact information constraints yield a closed-form allocation, an unbounded locality penalty, and an oracle for certified tracking. Multi-arm experiments demonstrate source migration; a matched frozen-gate ablation isolates its regret benefit, which disappears under homogeneous counts. Feedback interfaces should thus be evaluated by where they distinguish competing alternatives and the cost of learning to exploit that information.

ETHICS STATEMENT

The work studies admission information that may be useful to a strategic submitter or to a defender designing a monitoring interface. All experiments use public tabular datasets and local, frozen models; no deployed service is targeted. We report calibration failures and random-allocation comparisons alongside successful information-learning experiments.

REPRODUCIBILITY STATEMENT

The supplement includes source data, per-stream and per-split results, complete proofs, statistical aggregation, executable policies, and vector/300-dpi figure-generation code. The README specifies the environment and commands. Appendices state every split, likelihood, clock, calibration cost, and stopping rule used in the new experiments.

AI USE STATEMENT

Generative AI tools assisted with experiment execution, related technical tasks, and language polishing. Numerical results were generated by the supplied executable experiments. The authors are responsible for the accuracy and integrity of all results and the final manuscript.

REFERENCES

Jacob D. Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandits, with and without censored feedback. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Alexia Atsidakou, Orestis Papadigenopoulos, Constantine Caramanis, Sujay Sanghavi, and Sanjay Shakkottai. Asymptotically-optimal Gaussian bandits with side observations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1057–1077, 2022. URL `https://proceedings.mlr.press/v162/atsidakou22a.html`.

Barry Becker and Ronny Kohavi. Adult, 1996.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 521–529, 2014.

Richard Combes, Alexandre Proutiere, Donggyu Yun, Jungseul Ok, and Yung Yi. Optimal rate sampling in 802.11 systems. In *IEEE INFOCOM*, pp. 2760–2767, 2014. doi: 10.1109/INFOCOM.2014.6848225.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2443–2452, 2020.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *IEEE Symposium on Foundations of Computer Science*, pp. 429–438, 2013.

Yue Fu, Qingqing Ye, Rong Du, and Haibo Hu. Interactive trimming against evasive online data manipulation attacks: A game-theoretic approach. In *IEEE International Conference on Data Engineering*, pp. 3587–3599, 2024.

Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits for preserving local privacy. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pp. 387–412, 2018. URL https://proceedings.mlr.press/v83/gajane18a.html.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 359–376, 2011.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Jonas Geiping, Liam Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021.

Todd L. Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997. doi: 10.1137/S0363012994275440.

Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Spambase, 1999.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *IEEE Symposium on Security and Privacy*, pp. 19–35, 2018. doi: 10.1109/SP.2018.00057.

Kwang-Sung Jun and Chicheng Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/46489c17893dfdcf028883202cefd6d1-Abstract.html.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2777–2821, 2021. URL https://proceedings.mlr.press/v134/kirschner21a.html.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017.

Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, 111:1–47, 2022.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 728–737, 2017. URL https://proceedings.mlr.press/v54/lattimore17a.html.

Rami Mohammad and Lee McCluskey. Phishing websites, 2012.

Sérgio Moro, Paulo Rita, and Paulo Cortez. Bank marketing, 2014.

Riccardo Poiani, Marc Jourdan, Emilie Kaufmann, and Rémy Degenne. Best-arm identification in unimodal bandits. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2233–2241, 2025. URL https://proceedings.mlr.press/v258/poiani25a.html.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P. Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9389–9398, 2021.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Cindy Trinh, Emilie Kaufmann, Claire Vernade, and Richard Combes. Solving bernoulli rank-one bandits with unimodal thompson sampling. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 862–889, 2020.

Bart Van Parys and Negin Golrezaei. Optimal learning for structured bandits. *Management Science*, 70(6):3951–3998, 2024. doi: 10.1287/mnsc.2020.02108.

A PROOFS AND ASSUMPTIONS

A.1 THE ESTABLISHED RELEASED-LAW LOWER BOUND

Fix $p \in \Theta_r$ and $q \in \mathrm{Alt}(p)$. Let P_p^B, P_q^B be the laws of the adaptive history through decision B. Because a policy uses the same conditional arm-selection kernel in both worlds, the chain rule gives

$$\mathrm{KL}(P_p^B \| P_q^B) = \sum_k \mathbb{E}_p N_k(B) d_k(p_k, q_k). \tag{9}$$

Let $E_B = \{N_s(B) > B/2\}$. Uniform efficiency and positive suboptimality gaps imply $P_p(E_B) \to 1$ and $P_q(E_B) = o(B^{\alpha-1})$ for every $\alpha > 0$. Data processing to this event gives

$$\liminf_{B\to\infty} \frac{1}{\log B} \sum_k \mathbb{E}_p N_k(B) d_k(p_k, q_k) \geq 1. \tag{10}$$

The optimal-arm term is zero by the definition of $\mathrm{Alt}(p)$. On a subsequence realizing a finite limit inferior of normalized regret, every normalized suboptimal-arm count is bounded because its gap is positive. Extract a coordinatewise convergent subsequence. Its allocation vector satisfies every fixed alternative constraint and hence their infimum constraint. Minimizing the regret objective proves Equation (2). An infinite limit inferior is immediate; an empty alternative set imposes no constraint. This is the standard structured-bandit argument (Graves & Lai, 1997; Combes et al., 2017). If a class contains two observationally identical worlds with different optima, no uniformly efficient policy exists on that full class; we use the explicit linear-regret argument below rather than treating the conditional lower bound as an existence claim.

A.2 PROOF OF THE EXACT TWO-CONSTRAINT REDUCTION

Write $m = s - 1$ and $j = s + 1$ for existing neighbors. Assumption 1 implies $q_s = p_s$ in every confusing alternative, so the mean at s remains μ_s.

Left alternatives: necessary information. If $s(q) < s$, strict unimodality of $r_k q_k$ implies $r_m q_m > r_s q_s = \mu_s$. Thus $q_m > t_-$ and monotone survival gives $q_i \geq q_m > t_-$ for every $i < s$. This is impossible when $t_- \geq 1$. When $p_i < t_-$, monotonicity of $d_i(p_i, \cdot)$ above p_i gives $d_i(p_i, q_i) \geq d_i(p_i, t_-)$. When $p_i \geq t_-$, nonnegativity suffices and $a_i = 0$. Consequently every left alternative has information at least $I_-(\eta) = \sum_{i<s} \eta_i a_i$.

Left alternatives: attainability of the information infimum. For $t_- < 1$ and any sufficiently small $u > 0$, define

$$q_i^{-,u} = \begin{cases} \max(p_i, t_- + u), & i < s, \\ p_i, & i \geq s. \end{cases} \tag{11}$$

This vector is nonincreasing: a maximum with a constant preserves the order of the prefix, and $t_- > p_s$ joins it to the unchanged suffix. For $i < s$, $r_i q_i^{-,u} = \max(r_i p_i, r_i(t_- + u))$. Both arguments of this maximum are strictly increasing over the prefix; their maximum is therefore strictly increasing too. At m, $r_m q_m^{-,u} > \mu_s$, while the suffix starting at s is the original strictly decreasing sequence. Thus $q^{-,u} \in \Theta_r$ with optimum m, and $q_s^{-,u} = p_s$. By continuity its information tends to $I_-(\eta)$ as $u \downarrow 0$. Notice that the affected coordinates are a *left prefix*, not a right suffix.

Right alternatives. If $s(q) > s$, then $r_j q_j > \mu_s$, hence $q_j > t_+$, and its information is at least $I_+(\eta) = \eta_j d_j(p_j, t_+)$. Strict unimodality of the original instance gives $p_j < t_+$. Strictly increasing impacts give $t_+ < p_s$. Choose $u > 0$ with $t_+ + u < p_s$ and change only p_j to $t_+ + u$. Survival order is maintained because this raises p_j but keeps it below p_s; the utility at j becomes the unique peak, and every other strict utility comparison is unchanged. These alternatives approach I_+.

Every confusing alternative is left or right. Therefore the inner infimum is $\min(I_-, I_+)$ over the existing sides. Requiring it to be at least one is exactly Equation (4). The supports of its two constraints are disjoint. On the left, let $\gamma = \min_{i:a_i>0} \tau\Delta_i/a_i$. Then $\tau \sum_{i<s} \Delta_i \eta_i \geq \gamma \sum_{i<s} a_i \eta_i \geq \gamma$, with equality by allocating $1/a_{i^\star}$ to a minimizing coordinate. The right allocation is $1/b_+$. This proves Equation (5). Evaluating its divergences and scanning the prefix takes $O(K)$ operations. □

Homogeneous-channel corollary. For $i < s-1$, strict unimodality gives $\Delta_i > \Delta_{s-1}$, while $p_i \geq p_{s-1}$. If $p_i \geq t_-$ then $a_i = 0$ and that coordinate is uninformative. Otherwise $a_i = d(p_i, t_-) \leq d(p_{s-1}, t_-) = a_{s-1}$ by the additional monotonicity assumption. Its cost/information ratio cannot improve on the neighbor's. For Bernoulli laws with a common increasing map g, the derivative of $D_{\mathrm{B}}(z, w)$ in $z < w$ is $\log[z(1-w)/(w(1-z))] < 0$; composition with g preserves this property. Counts and randomized counts introduce only a common factor τ. This covers all homogeneous injective channels claimed in the corollary. □

Separation and numerical values. In the four-arm example, $\Delta = (.3, .1375, 0, .2)$ and $t_- = 14/15$. At $\epsilon = .1$ the optimal left allocation is on arm 1, with $\eta_1 = .5186619$, $\eta_2 = 0$; the right allocation is $\eta_4 = .9707972$. The objective is 6.9951603. With counts on all arms, $\eta_2 = .3004409$, $\eta_1 = 0$, and the objective is 4.7094014. For small ϵ, $\alpha_\epsilon = \epsilon/2 + O(\epsilon^3)$ and $\beta_\epsilon = 1/2 - \epsilon/4 + O(\epsilon^3)$, so for fixed $p \neq q$,

$$\tau D_{\mathrm{B}}(\beta_\epsilon + \alpha_\epsilon p, \beta_\epsilon + \alpha_\epsilon q) = \tfrac{1}{2}\tau\epsilon^2(p-q)^2 + O(\epsilon^3). \tag{12}$$

Neighbor-restricted left cost diverges as ϵ^{-2}, whereas the arm-1 option remains finite. The supplied code compares the closed form to a linear-program solver, verifies both boundary constructions, and checks feasibility against sampled confusing alternatives for 56 independently constructed strictly unimodal instances. These numerical checks supplement, and do not replace, the proof.

A.3 Release channels, parity, and audits

Identifiability and information quantity. Information location presumes that the released laws distinguish the relevant alternatives. Two short examples show why that premise matters, and why identifiability alone is not enough for efficient learning.

For $D_{\mathrm{B}}(p, q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$, full counts give $d_{\mathrm{count}}(p, q) = \tau D_{\mathrm{B}}(p, q)$. Randomized response maps p to $\beta_\epsilon + \alpha_\epsilon p$, where $\beta_\epsilon = (e^\epsilon + 1)^{-1}$ and $\alpha_\epsilon = 1 - 2\beta_\epsilon$, giving $d_{\mathrm{RR}}(p, q) = \tau D_{\mathrm{B}}(\beta_\epsilon + \alpha_\epsilon p, \beta_\epsilon + \alpha_\epsilon q)$. Data processing ensures that every post-processing of a count has divergence at most $\tau D_{\mathrm{B}}(p, q)$ (Duchi et al., 2013; Kairouz et al., 2014).

Proposition 1 (One bit can lose exponentially more information)**.** *For $0 < p < q < 1$, an any-survivor bit has divergence*

$$d_{\mathrm{any}}(p, q) = D_{\mathrm{B}}((1-p)^\tau, (1-q)^\tau) \leq \tau(1-p)^\tau \log\frac{1-p}{1-q}. \tag{13}$$

Releasing the survival bit of one uniformly selected candidate instead has $d_{\mathrm{sample}}(p, q) = D_{\mathrm{B}}(p, q)$, independent of τ. Both channels are injective, although their information can differ exponentially in batch size.

The sampled bit can be generated from the count alone by drawing $Z \mid C \sim \mathrm{Bernoulli}(C/\tau)$. Under homogeneous channels, Equation (5) implies $C_{\mathrm{sample}} = \tau C_{\mathrm{count}}$, while any-survivor information can collapse exponentially. The two one-bit outputs have equal communication size but different monitoring meanings: a sampled bit does not answer whether *any* candidate passed.

When feedback conceals the optimum. An optimal arm is identifiable only if all parameter vectors with identical per-arm released laws have the same optimum. For even $\tau = 10$, parity has mean $h(p) = [1 - (1-2p)^{10}]/2$. With $r = (.5, .6, 1)$, the two vectors $p^A = (.95, .9, .2)$ and $p^B = (.95, .9, .8)$ are ordered and have strictly unimodal utilities with optima 2 and 3. Yet $h(.2) = h(.8)$, so their adaptive histories have identical laws. More quantitatively, every policy satisfies

$$\max\{\mathrm{Reg}_A(B), \mathrm{Reg}_B(B)\} \geq .13\tau B. \tag{14}$$

At arm 3's utility tie $q = .54$, parity KL is positive; at the admissible alternative $q = .8$, it is zero. Assumption 1 is essential. A tagged count audit with probability $\rho > 0$ restores injectivity: releasing (audit flag, Z) has divergence $\rho\tau D_{\mathrm{B}}(p, q) + (1-\rho)d_{\mathrm{parity}}(p, q)$. The remaining bottleneck at an aliased pair is proportional to ρ.

Derivations. For a binomial count, the likelihood ratio cancels the binomial coefficient; expectation gives $\mathrm{KL}(\mathrm{Bin}(\tau,p)\|\,\mathrm{Bin}(\tau,q)) = \tau D_{\mathrm{B}}(p,q)$. A homogeneous count is therefore sufficient for its τ raw bits. This statement concerns information retained by the release, not the cost of delaying decisions until a batch ends.

A randomized survival bit has probability $\beta_\epsilon + \alpha_\epsilon p$. Independent perturbation followed by counting produces the corresponding binomial law. It is injective for $\epsilon > 0$; the channel at $\epsilon = 0$ is constant and is excluded from the injective theorem. For any increasing g, the derivative of $D_{\mathrm{B}}(g(p), g(q))$ in $q > p$ is $g'(q)(g(q) - g(p))/[g(q)(1 - g(q))] \geq 0$ whenever differentiable; the explicit maps here are differentiable in the interior.

For the any-survivor channel let $a = (1-p)^\tau$ and $b = (1-q)^\tau$. Then $a > b$, and

$$D_{\mathrm{B}}(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b)) \leq a\log(a/b). \tag{15}$$

Substitution proves Equation (13). Selecting one exchangeable candidate, or drawing a bit with conditional mean C/τ, instead yields $\mathrm{Bernoulli}(p)$. Both releases use a single bit; only the any-survivor law saturates exponentially at high survival probabilities. The bound concerns fixed $p < q$ as τ varies.

For even τ, the binomial generating function gives $\Pr(C \text{ odd}) = [1-(1-2p)^\tau]/2 = h(p) = h(1-p)$. The two worlds in the main text have utility vectors $(.475, .54, .2)$ and $(.475, .54, .8)$. Each satisfies all order and strict unimodality conditions. All released laws coincide, so induction over adaptive choices makes the expected number of pulls of each arm identical across worlds. The sums of the per-candidate gaps are .39, .26 and .34 for arms 1, 2 and 3. Therefore $\mathrm{Reg}_A(B) + \mathrm{Reg}_B(B) \geq .26\tau B$, proving Equation (14). A positive KL at the utility tie alone cannot rule out a zero-information, more distant alternative.

If an independent audit flag is released together with a full count with probability ρ and parity otherwise, the two flag values have disjoint observable supports. Conditioning on the flag proves

$$d_\rho(p,q) = \rho\tau D_{\mathrm{B}}(p,q) + (1-\rho)d_{\mathrm{parity}}(p,q). \tag{16}$$

For $\rho > 0$ the first component is injective. At the aliased pair its information is exactly $\rho\tau D_{\mathrm{B}}(p,q)$. If the flag is hidden, this divergence identity need not hold, and we do not use it for that model.

A.4 Impact nonidentifiability and target error

Fix p and choose two impact vectors r, r'; conditional on any selected arm, all feedback laws are identical. Thus an observation-only estimator cannot distinguish the scales. For example, $p = (.8, .4)$ paired with strictly increasing impacts $(1, 1.1)$ or $(1, 3)$ gives different optima. External calibration is additional information.

For Equation (32), let $\widehat{\mu}_k = \widehat{r}_k p_k$ and $k^\star$ maximize the target mean m_k. For any action k,

$$m_{k^\star} - m_k \leq \widehat{\mu}_{k^\star} - \widehat{\mu}_k + 2\delta \leq \max_j \widehat{\mu}_j - \widehat{\mu}_k + 2\delta. \tag{17}$$

Summation and expectation give the bound. The target here must have stationary additive mean effects. A final retraining loss is generally nonadditive; our empirical impact then describes a particular intervention distribution and size. We do not apply the regret inequality to that loss without the required approximation. If admitted conditional scales transfer with error at most e, then $|m_k - \widehat{r}_k p_k| \leq e$ because $p_k \leq 1$; a true gap larger than $2e$ preserves the identity of the optimum.

A.5 Closed model and certified allocation tracking

We first establish the allocation result on a compact model, then specify the estimator and every decision rule used by Certified Oracle Tracking (COT). The proof separates the information allocation from the statistical certificate needed to exploit an estimated leader.

A closed model with a strict interior truth. For fixed $0 < \ell < u < 1$, define the mode polytope

$$\begin{aligned}\mathcal{P}_m = \{q \in [\ell,u]^K : q_i \geq q_{i+1}, \quad & r_i q_i \leq r_{i+1}q_{i+1}\ (i < m),\\ & r_i q_i \geq r_{i+1}q_{i+1}\ (i \geq m)\}, \qquad \mathcal{P} = \bigcup_{m=1}^{K} \mathcal{P}_m.\end{aligned} \tag{18}$$

Each set is the intersection of finitely many closed half-spaces and a bounded box. Their finite union is compact. It includes flat survival segments, flat utility segments and tied optima. For tracking we use $\mathcal{P}$ as the known model; the true p satisfies $u > p_1 > \cdots > p_K > \ell$, strict utility unimodality, $1 < s < K$, $t_- < u$, and a unique minimizing prefix ratio. Theorem 1 continues to concern the original class Θ_r. Whenever a mode has a strict feasible point $v^{(m)}$, mixing it with any $q \in \mathcal{P}_m$ produces strict feasible points converging to q. Thus the polytopes used here are closures of their strict mode classes.

For count/RR bits let $\beta_k = 0$ for a count and $\beta_k = (1 + e^{\epsilon_k})^{-1}$ for RR, with $\alpha_k = 1 - 2\beta_k > 0$ and $g_k(v) = \beta_k + \alpha_k v$. The released mean is $\theta_k = g_k(p_k)$, and its reward map is $\mu(k, \theta) = r_k(\theta_k - \beta_k)/\alpha_k$. The affine image $g(\mathcal{P})$ is compact; its coordinates are bounded away from zero and one. The reward is continuous and the Bernoulli KL is finite and continuous on this image. These are the observation, compactness and continuity requirements in Combes et al. (2017).

Closing the model leaves the constant unchanged. Consider $q \in \mathcal{P}$ with $q_s = p_s$ and an arm strictly better than s. If it lies left of s, weak unimodality forces $r_{s-1}q_{s-1} \geq r_s q_s$, and survival order forces $q_i \geq t_-$ for all $i < s$. If it lies right, weak unimodality forces $r_{s+1}q_{s+1} \geq r_s q_s$. The same inequalities hold if a different arm ties s at the maximum. Consequently every added boundary or tie alternative satisfies one of the two information constraints. Their union adds no smaller information cost.

Conversely, the left and right perturbations in Appendix A.2 stay in $[\ell, u]^K$ for sufficiently small perturbations, because $t_- < u$ and $\ell < p_{s+1} < t_+ < p_s < u$. They remain strictly unimodal and approach both information infima. Hence the closed-model program has exactly the constant and the suboptimal allocation in Equation (5). Boundary ties may be included or excluded from the infimum without changing it. This argument establishes equality of the information programs directly, rather than inferring it from compactness.

The local allocation map. Normalize the zero-cost coordinate by $c_s = 0$. At a strict interior truth,

$$c_{i^\star}(p) = 1/a_{i^\star}(p), \qquad c_{s+1}(p) = 1/b_+(p), \qquad c_k(p) = 0 \quad \text{otherwise.} \tag{19}$$

The strict utility gaps fix s in a neighborhood. The strict prefix-ratio gap fixes $i^\star$, and the positive selected divergences have continuous reciprocals. An arm with $a_i = 0$ cannot become a better source under an arbitrarily small perturbation: its utility gap stays positive while its divergence tends to zero. Thus the normalized allocation is unique and locally continuous. This proves the required local regularity explicitly; no global uniqueness at tied parameter vectors is needed.

Relation to OSSB. Theorem 2 of Combes et al. (2017) uses empirical observation means, allocation quotas, an estimation counter, and least-normalized-count exploration. Its asymptotic factor tends to one as the two fixed tuning parameters tend to zero. A complete executable policy must also define its estimate outside the structured model and decide when a leader is statistically certified. COT uses that scheduling principle with the structured estimator and full-alternative certificate below. The guarantee proved here applies to this specified policy; the experiments therefore label it COT, separately from both OSSB and Prefix-KL.

An estimator defined for every history. Initialize each arm once. At decision t, let N_i be its count and z_i its empirical released mean. Clip $(z_i - \beta_i)/\alpha_i$ to $[\ell, u]$ to obtain v_i, and set $m = \arg\max_i r_i v_i$ (smallest-index tie rule). Project v in sup norm onto $\mathcal{P}_m$, obtaining w. Mix with a fixed, known strict point in that mode:

$$\widehat{p}(t) = (1 - \omega_t)w + \omega_t v^{(m)}, \qquad \omega_t = \min\{.1, (t+1)^{-1/2}\}. \tag{20}$$

The reference depends only on r, ℓ, u, m. In the experiments its utilities are $r_i v_i^{(m)} = h \exp(-\zeta|i - m|)$, with $\zeta = \frac{1}{4}\min_i \log(r_{i+1}/r_i)$ and h the midpoint of the interval that puts all coordinates strictly inside the box. The code reduces ζ if that interval is empty. A strict reference for each mode is part of the tracking model specification.

The projection has a simple exact feasibility test. At candidate radius d, start with intervals $[L_i, U_i] = [v_i - d, v_i + d] \cap [\ell, u]$. Write $R_i = r_{i+1}/r_i$. Before the mode, propagate $L_{i+1} \leftarrow \max(L_{i+1}, L_i/R_i)$ and $U_{i+1} \leftarrow \min(U_{i+1}, U_i)$; after the mode propagate $U_{i+1} \leftarrow \min(U_{i+1}, U_i/R_i)$. An empty

interval means infeasibility. Otherwise a backward pass selects the feasible coordinates closest to v_i. Bisection finds the minimum radius; its numerical tolerance decreases as $O(t^{-2})$. Mixing gives strict survival and strict utility comparisons, so the oracle is defined at every estimated mode. Equal prefix ratios use the smallest index; a numerically unresolved divergence triggers least-count estimation.

This is a consistent structural estimate, not an independent-arm MLE. If $\|z-\theta\|_\infty \le d$, clipping gives $\|v-p\|_\infty \le d/\alpha_{\min}$. For sufficiently small d, the selected mode is s and $p \in \mathcal{P}_s$ is a feasible projection candidate. Therefore

$$\|\widehat{p}-p\|_\infty \le 2d/\alpha_{\min} + \omega_t(u-\ell) + O(t^{-2}). \tag{21}$$

All likelihood certificates use the original z, preserving the Bernoulli concentration law despite the structural projection.

A computable certificate for exploitation. For an estimated leader j, define two relaxed alternative sets

$$\begin{aligned}\mathcal{R}_j^- &= \{q \in [\ell,u]^K : q_i \ge (r_j/r_{j-1})q_j \quad (i<j)\},\\ \mathcal{R}_j^+ &= \{q \in [\ell,u]^K : q_{j+1} \ge (r_j/r_{j+1})q_j\},\end{aligned} \tag{22}$$

omitting nonexistent sides. Any parameter in $\mathcal{P}$ at which j is suboptimal belongs to their union. Set

$$Z_j(t) = \inf_{q \in \mathcal{R}_j^- \cup \mathcal{R}_j^+} \sum_i N_i D_{\mathrm{B}}(z_i, g_i(q_i)). \tag{23}$$

Each side reduces to one convex scalar minimization over q_j. Conditional on q_j, set every affected coordinate to $\max\{v_i, (r_j/r_{j\mp1})q_j\}$ and every free coordinate to v_i. The left scalar domain is $[\ell, ur_{j-1}/r_j]$ and the right domain is $[\ell,u]$. An empty domain has value $+\infty$. The code bisects the monotone derivative and uses a tangent lower bound for the certificate. Both the bracket width and the relative subtraction margin decrease as $O(t^{-2})$. On the fixed interior observation box the objective and its one-sided derivatives are $O(t)$, so the scalar-solver error is $O(t^{-1})$. The lower bound makes numerical acceptance conservative.

Complete decision rule. Fix $0<\lambda<1/K$ and $\gamma>0$. Let L count non-exploitation decisions after initialization. Evaluate $c = c(\widehat{p})$ with Equation (19), using u as the upper boundary when testing whether a left alternative exists. Set $j = \arg\max_i r_i\widehat{p}_i$. If all $N_i \ge (1+\gamma)c_i \log\max(3,t)$ and $Z_j(t) > (1+\gamma/2)\log\max(3,t)$, play j and leave L unchanged. Otherwise increment L and use the following order:

1. If $\min_i N_i \le \lambda L$, play a least-sampled arm (estimation).
2. If an oracle quota is unmet, minimize N_i/c_i over $c_i>0$ (allocation exploration).
3. If all quotas are met, play j to tighten its certificate (certification).

Initialization and all three non-exploitation steps are counted in the reported allocations. True probabilities, optima, gaps and allocations are used only by the separate evaluator. Fixed tuning inflates oracle quotas to $(1+\gamma)c_i\log t$; estimation adds a further λL floor. Neither target is an assertion that the actual finite-budget counts equal $c_i \log t$.

Regret guarantee for the specified policy. At the regular truth of Corollary 2, for every fixed $\gamma>0$ there is a finite constant $A_\gamma(p)$, independent of λ, such that

$$\limsup_{B\to\infty} \frac{\mathrm{Reg}_{\mathrm{COT}(\lambda,\gamma)}(B)}{\log B} \le (1+\gamma)C_H(p) + \frac{\lambda A_\gamma(p)}{1-K\lambda}. \tag{24}$$

We prove the bound through four lemmas. Counts and empirical means below are measured immediately before the current decision. Couple each arm to an infinite independent Bernoulli sequence; adaptive sampling reveals its next unused bit. This convention makes the concentration arguments apply to the actual online histories.

Lemma 1 (Summable erroneous exploitation)**.** *Let W_B count exploitation decisions with leader $j \ne s$ through B. For each fixed $\gamma>0$, $\sup_B \mathbb{E}_p W_B < \infty$.*

Proof. If $j \neq s$, the true parameter belongs to $\mathcal{R}_j^- \cup \mathcal{R}_j^+$: weak unimodality puts a neighbor of j at least as high as j, and survival order propagates the left comparison over its prefix. Thus

$$Z_j(t) \leq S(t) := \sum_i N_i(t) D_{\mathrm{B}}(z_i(t), \theta_i).$$

Exploitation requires $Z_j(t) > (1+\gamma/2)\log\max(3,t)$. The adaptive Bernoulli concentration bound in Combes et al. (2017, Supplement, Lemma 1) gives

$$\sum_{t\geq 3} \Pr_p\{S(t) \geq (1+\gamma/2)\log t\} \leq G(\gamma/2, K) < \infty.$$

This bound concerns the raw empirical released means, which is why Equation (23) uses z rather than its structural projection. Summing the implication above proves the claim, including histories at which the estimate has the wrong mode. □

Lemma 2 (Estimation on the non-exploitation clock)**.** *Before non-exploitation decision L, every arm has at least $\lambda L/2 - K - 2$ observations. For any fixed neighborhood U of the regular truth, the expected total number of non-exploitation decisions with $\widehat{p} \notin U$ or $\|z-\theta\|_\infty > d$ is finite, provided $d > 0$ is chosen sufficiently small for U.*

Proof. Consider the last $\lceil L/2 \rceil$ non-exploitation decisions. If the final minimum count were below $\lambda L/2 - 2$, then throughout this interval it would be below the current λ-floor. Every such decision would consequently sample a least-count arm. Each K of these choices increases the minimum by at least one; intervening exploitation can only increase counts. This contradicts $\lambda < 1/K$, up to the stated integer and initialization terms. Applying the argument at $L-1$ gives the pre-decision bound.

Let $h_L = \max\{1, \lceil \lambda L/2 - K - 2 \rceil\}$ and let $\bar{Z}_{i,n}$ denote the first n observations in arm i's sequence. At a non-exploitation history indexed by L, regardless of its calendar time,

$$\begin{aligned}\Pr_p\{\|z-\theta\|_\infty > d \text{ at index } L\} &\leq \sum_{i=1}^{K} \sum_{n \geq h_L} \Pr_p\{|\bar{Z}_{i,n} - \theta_i| > d\} \\ &\leq \frac{2Ke^{-2d^2 h_L}}{1-e^{-2d^2}}. \end{aligned} \tag{25}$$

There is at most one such history for each L. The bound is summable in L, without asserting independence between counts and observations. Equation (21) puts $\widehat{p}$ in U whenever z is sufficiently accurate and t is sufficiently large. The vanishing reference mixture and projection tolerance add only a finite initial period. Shrinking U also fixes the true leader, the cheapest prefix arm, and arbitrarily small coordinatewise errors in $c(\widehat{p})$. □

Lemma 3 (A uniform finite certification threshold)**.** *For every $\gamma > 0$ there exist $M_\gamma(p) < \infty$, a neighborhood U_γ of p, and $d_\gamma > 0$ such that, whenever $\widehat{p} \in U_\gamma$, $\|z-\theta\|_\infty \leq d_\gamma$ and*

$$N_i \geq (1+\gamma)c_i(\widehat{p})\log t \quad \textit{for all } i, \qquad N_s \geq M_\gamma(p)\log t,$$

the leader is s and $Z_s(t) > (1+\gamma/2)\log t$. The constants and neighborhood can be chosen independently of λ.

Proof. Write $h = i^\star(p)$, $j = s+1$, $h_- = h$ and $h_+ = j$. On each compact relaxed alternative set define the retained objective

$$F_\pm(M, z, v) = \min_{q \in \mathcal{R}_s^\pm} \left\{ M D_{\mathrm{B}}(z_s, g_s(q_s)) + (1+\gamma)c_{h_\pm}(v) D_{\mathrm{B}}(z_{h_\pm}, g_{h_\pm}(q_{h_\pm})) \right\}.$$

At $(z, v) = (\theta, p)$, these minima increase with M and are bounded above by $1+\gamma$: take $q_s = p_s$ and the selected coordinate at its tie threshold. Along any sequence $M \to \infty$, compactness gives a convergent subsequence of minimizers. Bounded cost and injectivity force its q_s limit to be p_s. The left constraint then forces $q_h \geq t_-$; the right forces $q_j \geq t_+$. Their minimal divergences are a_h and b_+, respectively. Since $c_h a_h = c_j b_+ = 1$, both limits equal $1+\gamma$.

Choose a finite M so that both minima exceed $1 + 3\gamma/4$ at (θ, p). Each $F_\pm$ is continuous in (z, v) locally: its feasible set is fixed and compact, its Bernoulli objective is continuous, and the selected

allocation coordinates are locally continuous. Hence one common neighborhood makes both minima exceed $1 + 5\gamma/8$. Dividing the full certificate by $\log t$, retaining these two terms on each side, and using the assumed count lower bounds proves the result. The positive slack also allows a scalar-solver error $o(\log t)$. Neither the compactness argument nor the selected M uses λ. □

Lemma 4 (Phase counts and regret)**.** *At the regular truth, COT satisfies Equation* (24) *with*

$$A_\gamma(p) = \left[(1+\gamma)\sum_i c_i(p) + M_\gamma(p)\right]\sum_i \Delta_i.$$

Proof. Fix $e > 0$ and intersect the neighborhood in Lemma 3 with one where the leader is s and $|c_i(\widehat{p}) - c_i(p)| \le e$. Call all other non-exploitation histories inaccurate and count them in J_B. Lemma 2 gives $\sup_B \mathbb{E}J_B < \infty$.

On accurate histories, allocation exploration of arm i requires an unmet quota for that arm: the least normalized count is below $1 + \gamma$ times $\log t$ whenever any quota is unmet. Its total number is therefore at most $(1 + \gamma)(c_i(p) + e) \log B + 1$. Estimation selects an arm whose count is at most λL_B, so it contributes at most $\lambda L_B + 1$ choices per arm. Certification selects s; by Lemma 3, it stops once N_s reaches $M_\gamma(p) \log B$. All other selections of s help reach this threshold. Counting the three phases, with each inaccurate decision counted once, gives

$$(1 - K\lambda)L_B \le \left[(1+\gamma)\left(\sum_i c_i(p) + Ke\right) + M_\gamma(p)\right]\log B + J_B + O(K). \tag{26}$$

Accurate certification and correct exploitation incur zero regret. Initialization, inaccurate non-exploitation and erroneous exploitation contribute $O(1) + \Delta_{\max}\mathbb{E}(J_B + W_B)$, uniformly bounded by the first two lemmas. The remaining expected regret is at most

$$(1+\gamma)\sum_i \Delta_i(c_i(p) + e)\log B + \lambda\Big(\sum_i \Delta_i\Big)\mathbb{E}L_B + O(1).$$

Insert Equation (26), divide by $\log B$, then let $e \downarrow 0$. Since $\sum_i \Delta_i c_i(p) = C_H(p)$ for $\tau = 1$, this proves the displayed bound. The threshold M_γ stays fixed as e decreases; only the neighborhood and finite transient count change. □

Proof of Corollary 2. The closed-model argument above establishes the same two information infima as Theorem 1. Lemma 4 supplies the upper bound for the specified single-candidate policy. Given $\delta > 0$, choose γ so that $\gamma C_H(p) < \delta/2$, then choose $0 < \lambda < 1/K$ so that $\lambda A_\gamma(p)/(1 - K\lambda) < \delta/2$. Both tunings are fixed throughout the run. This proves Equation (7). □

B Comparison with the closest bandit models

The ordered model combines survival constraints with arm-dependent releases. Table 1 identifies which ingredients the closest frameworks handle and which allocation constraints the closed form resolves. The finite-budget and asymptotic questions are complementary: an exact allocation identifies the limiting information bottleneck, while learning which allocation applies incurs an additional transient cost (Jun & Zhang, 2020).

C Oracle-driven tracking experiments

All experiments in this section use $\tau = 1$, model bounds $[.02, .98]$, and the exact COT policy from Appendix A.5. The four-arm profile has $r = (.5, .75, 1, 1.25)$ and $p = (.8, .75, .7, .4)$. Arms 1, 3 and 4 release counts; arm 2 releases RR or a count. The crossing of the two left ratios is computed by scalar root finding, giving $\epsilon_2^\star = 1.7005786632$. Evaluation conditions are fixed at .1, $.85\epsilon_2^\star$, $1.15\epsilon_2^\star$, and full count. The two ratios are unequal in all four conditions.

The default tuning is $(K\lambda, \gamma) = (.10, .25)$. Sensitivity settings $(.04, .10)$ and $(.20, .50)$ are fixed across all four instances. Each method/setting uses 48 streams to 100,000 decisions. Forty distinct,

Table 1: Relationship to the closest frameworks. Arm-dependent observation maps alone are established; the new calculation combines them with the ordered-survival and unimodal-utility constraints.

Work	Observation and structure	Relation to the present result
Rate adaptation / OSUB (Combes et al., 2014; Combes & Proutiere, 2014)	Success observations with scaled or unimodal rewards.	Neighbor geometry; recovered by Corollary 1.
Corrupt bandits (Gajane et al., 2018)	Known arm-dependent corruption; monotone maps support KL-UCB-CF.	Independent-arm information requirements do not include the coupled left-prefix alternative in Equation (4).
OSSB / structure-adaptive bandits (Combes et al., 2017; Degenne et al., 2020)	General constrained alternatives and allocation-based exploration.	Supply the lower-bound and tracking machinery; our theorem solves the ordered released-law allocation oracle.
DUSA (Van Parys & Golrezaei, 2024)	Convex compact constraints on reward distributions and a dual allocation program.	General optimization approach; the unknown-mode unimodal class here is a union of mode-specific sets. Our reduction uses its particular order geometry.
Gaussian side observations (Atsidakou et al., 2022)	A known matrix specifies noisy observations about other arms.	Here observations concern the selected arm only; cross-arm information comes from parameter constraints.
This paper	Ordered survival, unimodal scaled utility, known arm-dependent releases.	Two constraints, an $O(K)$ closed form, and an unbounded locality penalty.

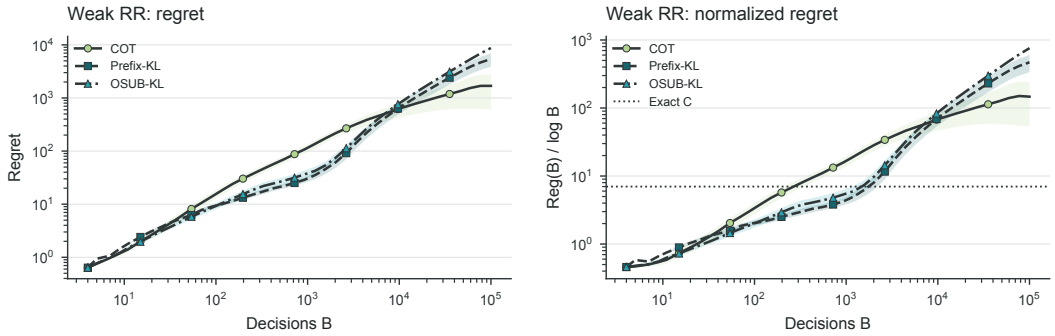
logarithmically spaced checkpoints include initialization and the final horizon. COT, Prefix-KL and OSUB-KL consume matched uniform random streams within an instance. Instance j, starting at zero, has seed $20260927 + 10000j$; stream h adds $1000003h$. The evaluator samples a Bernoulli bit with mean $g_k(p_k)$ after the policy chooses its arm. Only that bit enters the policy state. Table 2 compares final regret and the exact information source.

Table 2: Single-candidate tracking at $B = 100{,}000$. Mean regret $\pm$ 95% interval over 48 streams. $i^\star$ is the exact left information source; COT uses the fixed default tuning. Lower regret is better.

Neighbor receipt	C_H	$i^\star$	COT	Prefix-KL	OSUB-KL
Weak RR	7.00	1	1690.3 ± 1073.5	5405.8 ± 1539.5	8729.2 ± 225.5
Below switch	7.00	1	195.5 ± 33.2	124.7 ± 9.3	126.3 ± 8.1
Above switch	6.32	2	168.3 ± 36.4	107.2 ± 8.8	109.3 ± 6.3
Full count	4.71	2	109.5 ± 20.7	93.7 ± 8.5	77.4 ± 5.3

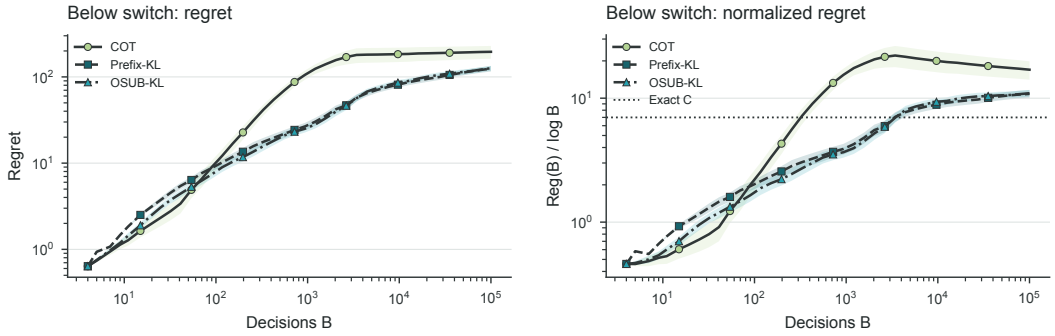
What the allocation plots measure. At each checkpoint we retain every arm count, the fitted oracle, the estimated source and leader, and the counts of initialization, estimation, allocation exploration, certification and exploitation decisions. Pseudo-regret is reconstructed independently as $\sum_i \Delta_i N_i$. The allocation panels plot suboptimal $N_i / \log B$, the exact c_i, and the nominal fixed-tuning quota $1.25c_i$. The estimation floor λL is additional and depends on the realized history. These quantities have different meanings, even when plotted on the same axes.

With weak RR, the mean COT arm-1 share among left queries is .765; with full counts it is .133. Near the switch, sampling is less decisive: the corresponding below/above shares are .222 and .126 despite a change in the unique oracle source. At 100,000 decisions the default COT normalized regrets are 146.8, 17.0, 14.6 and 9.5, while exact constants are 7.00, 7.00, 6.32 and 4.71. The large weak-RR interval reflects variation in the time needed to estimate the noisy neighbor. Across low, default and high tunings, weak-RR regrets are 2131.0, 1690.3 and 937.2; full-count regrets are 185.6, 109.5 and 146.5. Figures 8–11 give the four regret trajectories; Figure 12 resolves the switch allocations. Figure 13 and the raw files retain all tunings.



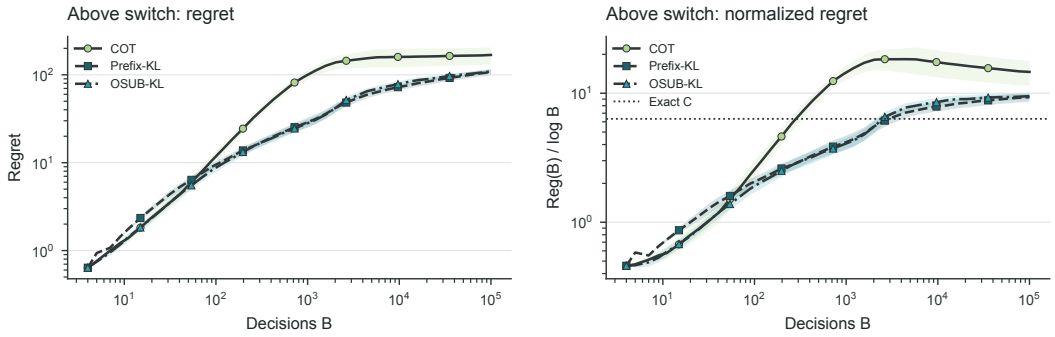
(a) **(b)**

Figure 8: Weak neighbor RR: (a) regret and (b) regret divided by $\log B$. Curves compare the three policies over 48 streams with 95% mean intervals. The horizontal line in (b) is C_H. COT uses the default tuning.



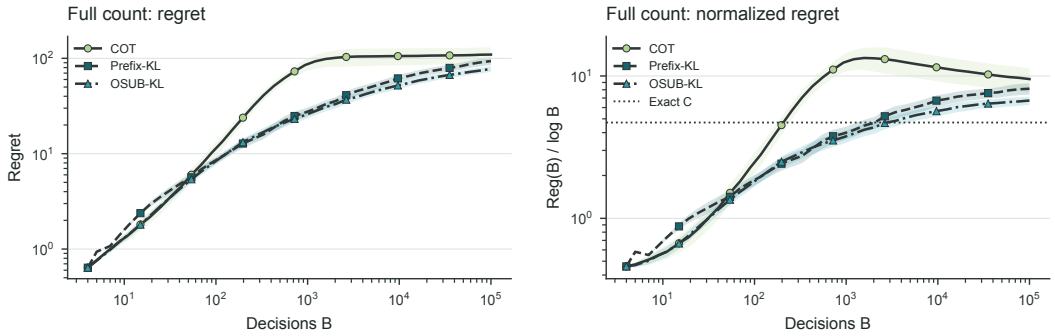
(a) **(b)**

Figure 9: Below the source switch: (a) regret and (b) regret divided by $\log B$. Curves compare the three policies over 48 streams with 95% mean intervals. The horizontal line in (b) is C_H. COT uses the default tuning.



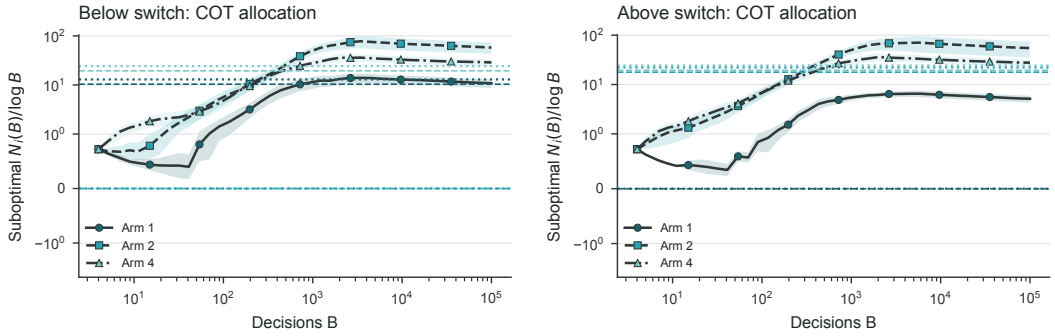
(a) **(b)**

Figure 10: Above the source switch: (a) regret and (b) regret divided by $\log B$. Curves compare the three policies over 48 streams with 95% mean intervals. The horizontal line in (b) is C_H. COT uses the default tuning.



(a) **(b)**

Figure 11: Full counts: (a) regret and (b) regret divided by $\log B$. Curves compare the three policies over 48 streams with 95% mean intervals. The horizontal line in (b) is C_H. COT uses the default tuning.



(a) **(b)**

Figure 12: COT suboptimal allocations (a) below and (b) above the source switch. Curves show $N_i/\log B$ with 95% intervals across 48 streams; dashed and dotted lines show c_i and $1.25c_i$. The weak-RR and full-count counterparts appear in Figure 3.

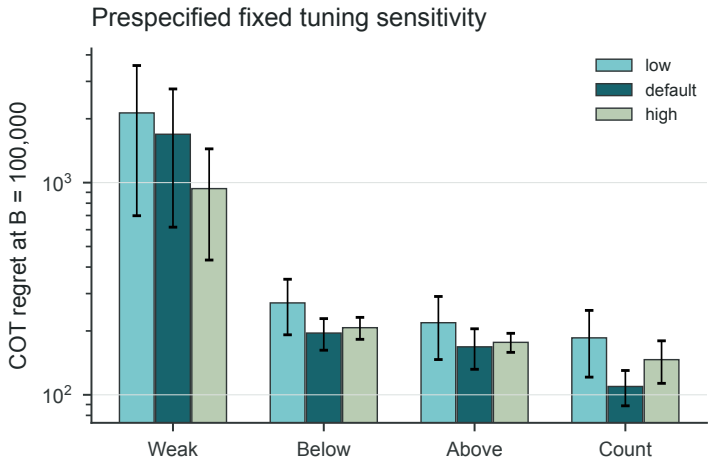


Figure 13: All prespecified COT tunings at 100,000 decisions, with 95% intervals over 48 streams. Low, default and high denote $(K\lambda, \gamma) = (.04, .10), (.10, .25), (.20, .50)$, respectively. More exploration helps the weak-RR transient but adds cost under counts.

D MULTI-ARM INFORMATION GEOMETRY AND REPRESENTATIVE LEARNING RUNS

Deterministic instance generator. Use $K = 10, 20, 50$ arms and place the peak at either $s = \mathrm{round}(.5K)$ or $s = \mathrm{round}(.8K)$. For $i \le s$ put $x_i = (i-1)/(s-1)$ and

$$r_i = .45 + .55x_i, \qquad p_i = .66 + \frac{.15}{s-1}(1 - x_i). \tag{27}$$

For $i > s$ put $y_i = (i-s)/(K-s)$ and set $r_i = 1 + .25y_i$, $p_i = .66e^{-1.2y_i}$. Survival decreases strictly. On the left, differentiating the product with respect to x gives $.55(.66 + .15/(s-1)) - .45(.15/(s-1)) - 1.1(.15/(s-1))x > 0$ for all selected $s \ge 5$. On the right its derivative is $.66e^{-1.2y}[.25 - 1.2(1 + .25y)] < 0$. Thus each profile has a unique strict peak at s. Its entire prefix lies below t_-, making each prefix arm a possible information source.

Near, middle and far count probes are placed at distance $d = 2, \max\{3, \mathrm{round}(.55(s-1))\}$, and $s - 1$ from s. A second probe at the next index toward s (or the preceding index if that would be the neighbor) releases RR with $\epsilon = 1.5$. Other left arms start at $\epsilon = .08$; the optimum and all right arms release counts. We sweep only the immediate left neighbor through 65 geometrically spaced values from .03 to 8, followed by a full count. The known scales, probabilities and all other channels remain fixed.

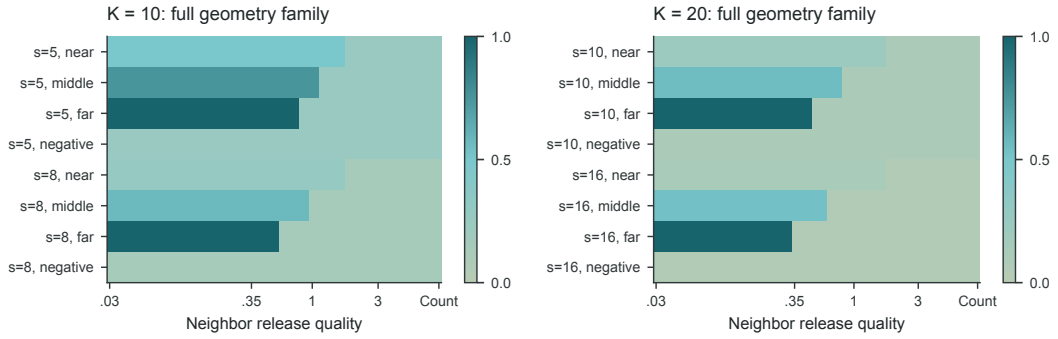
The negative-control profile replaces the left survival rule by $p_i = .90 - .24x_i$, keeping the same increasing scales and right suffix. Its left utility derivative is $.387 - .264x > 0$. For every selected s, the non-neighbor prefix probabilities are at least t_-, so their a_i are zero. A full-count far probe therefore cannot replace the neighbor. These four profile types, two peak positions and three arm counts give 24 profiles and $24 \times 66 = 1584$ geometry points.

Representative online runs. For each K, we select the far-probe profile with $s = \mathrm{round}(.8K)$ and neighbor feedback .08 or full count. Each of the six instances has 24 paired streams per policy to 50,000 decisions, with 40 logarithmic checkpoints. Only the fixed default COT tuning is used; the full sensitivity analysis is the four-arm experiment. No other combinations are selected for online evaluation.

Table 3: All six prespecified multi-arm representatives, 50,000 decisions. Mean regret ± 95% intervals over 24 streams. Strong means a full-count neighbor; other left channels retain their prescribed quality.

Instance	C_H	$i^\star$	COT	Prefix-KL	OSUB-KL
K10, weak	123.50	1	2443.8 ± 440.0	2575.5 ± 255.8	1987.7 ± 138.4
K10, strong	9.76	7	3704.7 ± 738.8	3496.5 ± 927.8	1548.3 ± 987.3
K20, weak	686.11	1	1649.7 ± 347.6	2312.5 ± 1139.9	1934.7 ± 651.7
K20, strong	21.13	15	2223.4 ± 149.7	3108.1 ± 899.4	2095.0 ± 781.2
K50, weak	5130.90	1	1375.5 ± 162.3	961.9 ± 601.5	3904.8 ± 1232.4
K50, strong	55.13	39	1721.9 ± 105.9	1964.0 ± 1104.0	3138.5 ± 1146.3

The regret ranking varies with K and channel quality (Table 3). COT makes no certified exploitation decisions in either $K = 20$ or $K = 50$ condition at this budget; its other phases can still select the optimal arm. At $K = 50$, even the true nominal suboptimal quotas $1.25 \sum_i c_i \log(50000)$ total approximately 193,626 decisions under weak neighbor feedback and 70,755 under full count, both exceeding the available horizon. This quantifies why the asymptotic allocation cannot yet describe these finite-budget counts. A finite-budget $\mathrm{Reg}(B)/\log B$ below C_H is consistent with an asymptotic lower bound. Figures 14–15 show the complete grids; Figures 16–21 show all six online representatives.



(a) **(b)**

Figure 14: Complete information-source grids for (a) $K = 10$ and (b) $K = 20$. Color is $(s - i^\star)/(s - 1)$, with near/middle/far probes and a zero-information control at each of two peak positions. Only the neighbor channel changes along a row.

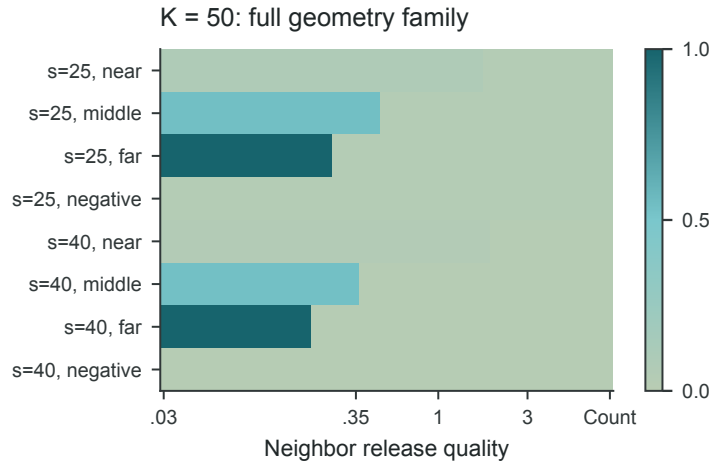
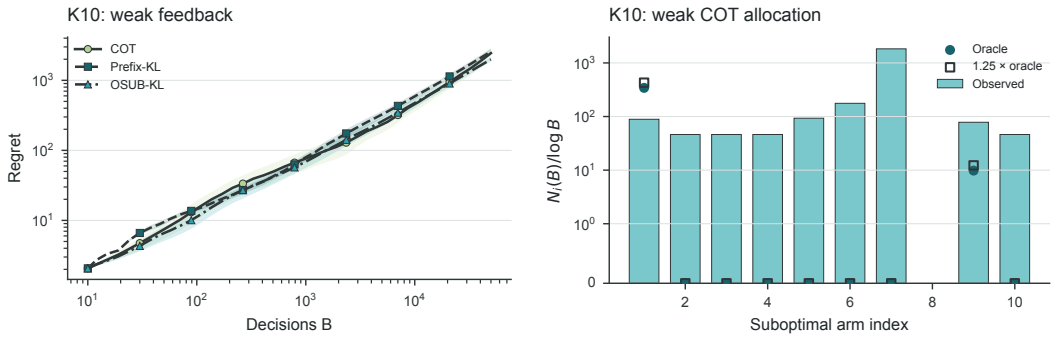
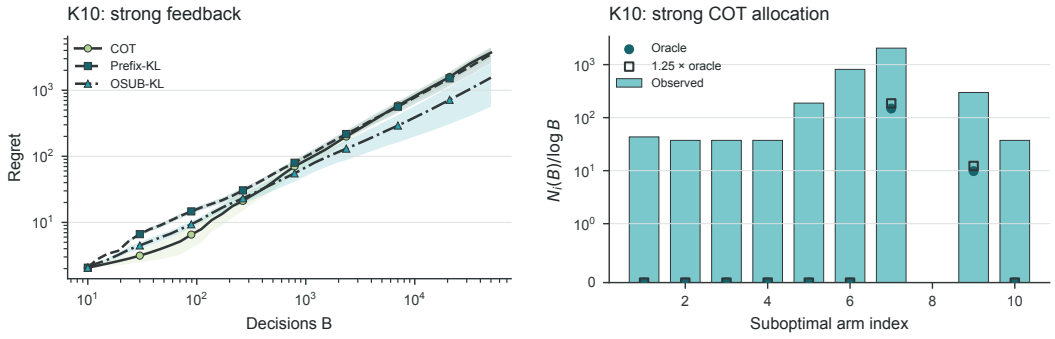


Figure 15: The complete $K = 50$ source grid, using the same color scale and construction as Figure 14.



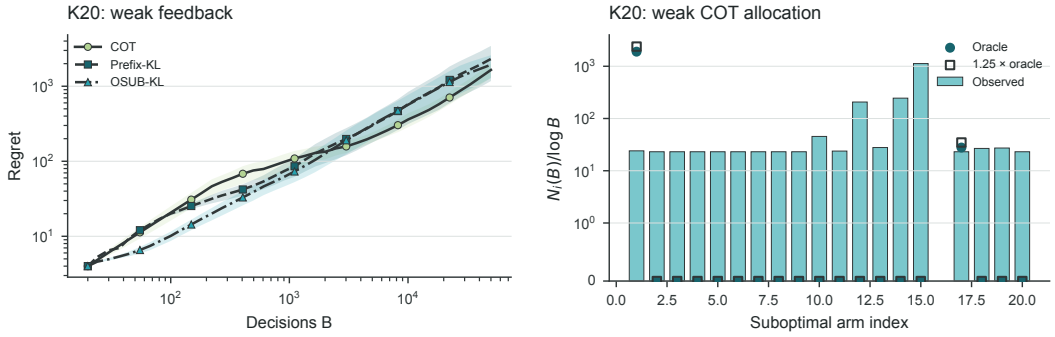
(a) **(b)**

Figure 16: $K = 10$, weak neighbor feedback: (a) regret trajectories, 24 streams with 95% intervals; (b) final COT suboptimal counts divided by $\log B$, compared with the exact allocation and the $1.25c_i$ quota. The optimal arm is omitted from (b). Strong changes only the neighbor to full count.



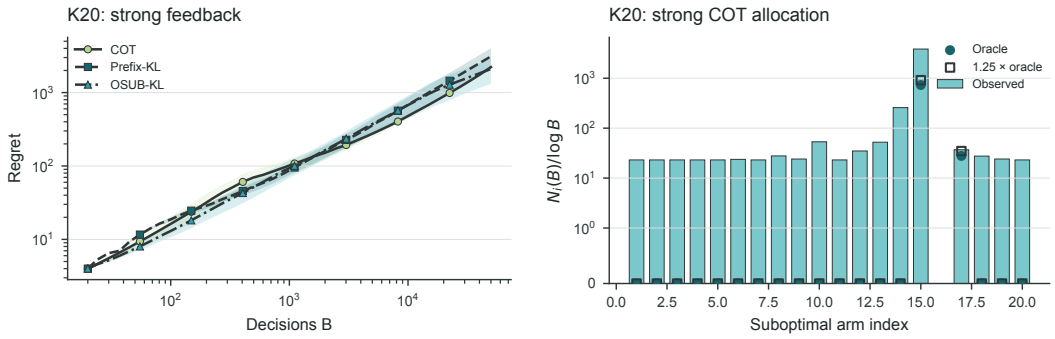
(a) **(b)**

Figure 17: $K = 10$, strong neighbor feedback: (a) regret trajectories, 24 streams with 95% intervals; (b) final COT suboptimal counts divided by $\log B$, compared with the exact allocation and the $1.25c_i$ quota. The optimal arm is omitted from (b). Strong changes only the neighbor to full count.



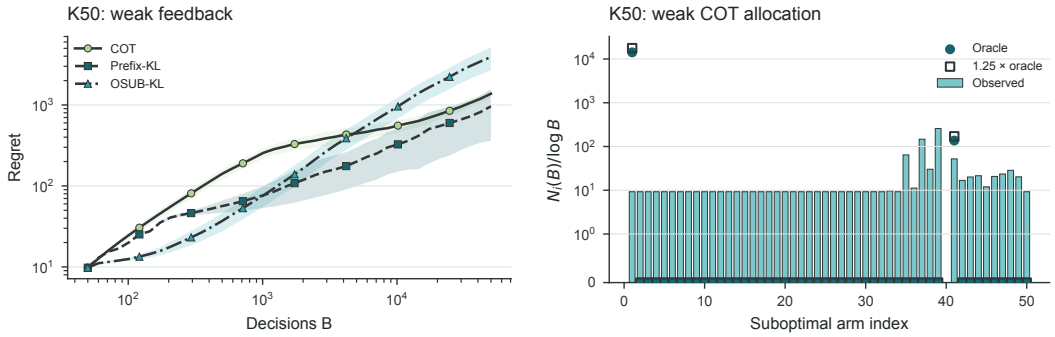
(a) **(b)**

Figure 18: $K = 20$, weak neighbor feedback: (a) regret trajectories, 24 streams with 95% intervals; (b) final COT suboptimal counts divided by $\log B$, compared with the exact allocation and the $1.25c_i$ quota. The optimal arm is omitted from (b). Strong changes only the neighbor to full count.



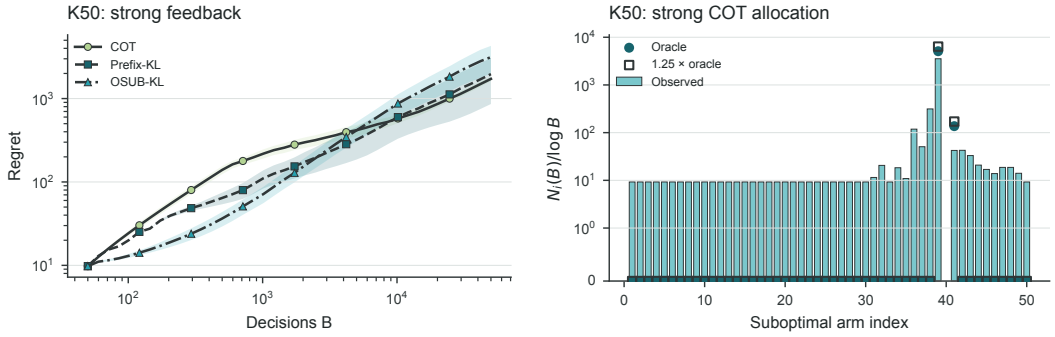
(a) **(b)**

Figure 19: $K = 20$, strong neighbor feedback: (a) regret trajectories, 24 streams with 95% intervals; (b) final COT suboptimal counts divided by $\log B$, compared with the exact allocation and the $1.25c_i$ quota. The optimal arm is omitted from (b). Strong changes only the neighbor to full count.



(a) **(b)**

Figure 20: $K = 50$, weak neighbor feedback: (a) regret trajectories, 24 streams with 95% intervals; (b) final COT suboptimal counts divided by $\log B$, compared with the exact allocation and the $1.25c_i$ quota. The optimal arm is omitted from (b). Strong changes only the neighbor to full count.



(a) **(b)**

Figure 21: $K = 50$, strong neighbor feedback: (a) regret trajectories, 24 streams with 95% intervals; (b) final COT suboptimal counts divided by $\log B$, compared with the exact allocation and the $1.25c_i$ quota. The optimal arm is omitted from (b). Strong changes only the neighbor to full count.

E EXECUTABLE POLICIES AND LIKELIHOODS

The batched policies use `revision_experiments.py`. An arm is initialized with one batch. Let S_k, n_k be the sum and number of released Bernoulli sufficient-statistic trials: $n_k = \tau N_k$ for counts and randomized counts, and $n_k = N_k$ for any-survivor and sampled-bit releases. Let $z_k = S_k/n_k$ and g_k be the released success map. A constrained MLE $\widehat{p}_k$ inverts g_k and clips to $[0, 1]$ where necessary. Define

$$\Lambda_k(q) = n_k\{D_{\mathrm{B}}(z_k, g_k(q)) - D_{\mathrm{B}}(z_k, g_k(\widehat{p}_k))\}, \qquad U_k = \sup\{q \in [\widehat{p}_k, 1] : \Lambda_k(q) \le b\}. \quad (28)$$

For any-survivor releases we evaluate this expression in the rare-failure parameter $(1 - p)^\tau$ to avoid cancellation near one. Bisection uses 20 iterations. Full-count feedback has zero subtracted deviance. Randomized response may have nonzero deviance when z_k is outside the feasible released interval. The subtraction is the exact likelihood-ratio normalization, not a heuristic clipping of the raw empirical statistic.

Global KL-UCB uses $b = \log \max(3, b_{\mathrm{round}}) + 3 \log\log \max(3, b_{\mathrm{round}})$ and maximizes $r_k U_k$. Line OSUB-KL uses the same budget with the current leader's occurrence count, forces the leader on counts congruent to one modulo three, and otherwise maximizes the index on the leader and its immediate neighbors. The leader maximizes the independent channel-MLE utility.

Ordered-KL uses the global clock and upper bounds $\overline{U}_j = \min_{i \le j} U_i$, selecting $\arg\max_j r_j \overline{U}_j$. Prefix-KL uses the following precisely specified variant. Compute a weighted least-squares

decreasing-isotonic projection of the independent estimates, with weights

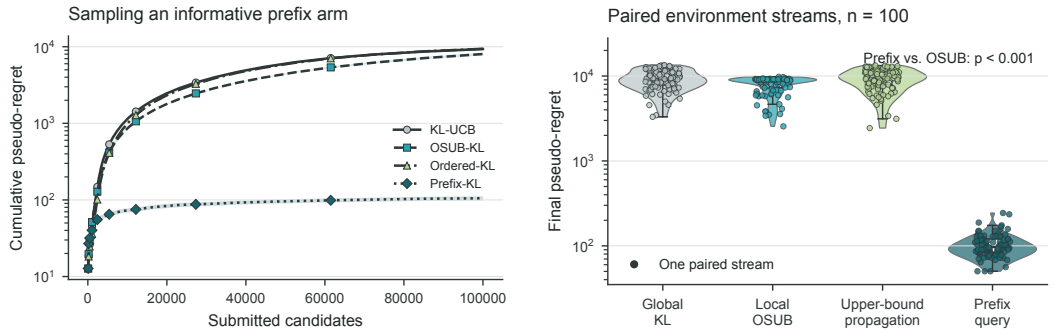
$$w_i = \frac{n_i \alpha_i^2}{\max\{z_i(1-z_i), .01\}}, \qquad \alpha_i = 1 - 2\beta_i, \tag{29}$$

where count arms have $\beta_i = 0$. This projection is used only to pick a leader, not to modify any measured environment or confidence bound. Let ℓ maximize projected utility and j maximize $r_j \overline{U}_j$. If $j = \ell$, pull j. Otherwise find $i \in \arg\min_{i \le j} U_i$ and pull i only if $r_j \widetilde{p}_i < r_\ell \widetilde{p}_\ell$, where $\widetilde{p}$ is the projected estimate; if this condition fails, pull j itself. The condition avoids repeatedly querying a prefix arm whose estimated probability cannot exclude the optimistic competitor. Before this choice, if any arm has fewer than $\lceil \sqrt{\log b_{\text{round}}} \rceil$ batches, pull an arm with smallest batch count. Ties use the smallest arm index. The weighted projection is implemented by pool-adjacent-violators. This is a diagnostic allocation rule; we give no regret upper bound or assertion of optimality for it.

Table 4: Pseudo-regret (lower is better), mean ± 95% interval over 100 streams per condition at 100,000 candidates. The first survival probability varies while $(p_2, p_3, p_4) = (.75, .7, .4)$. Only arm 2 is privatized. Prefix queries help most when its feedback is weak; they can cost more under strong feedback.

p_1	ϵ_2	Global KL	OSUB-KL	Ordered-KL	Prefix-KL
0.80	0.1	9332.8 ± 473.7	8032.1 ± 341.7	9401.6 ± 491.9	105.0 ± 6.6
0.80	0.5	491.8 ± 28.4	509.4 ± 29.2	492.5 ± 29.2	105.0 ± 6.7
0.80	2.0	95.3 ± 4.2	94.5 ± 4.6	95.0 ± 4.4	99.9 ± 5.1
0.88	0.1	9332.0 ± 473.7	8056.9 ± 339.7	9190.9 ± 501.1	1716.9 ± 655.1
0.88	0.5	491.0 ± 28.4	509.0 ± 29.0	484.2 ± 27.5	256.1 ± 29.2
0.88	2.0	94.5 ± 4.2	93.7 ± 4.6	95.2 ± 4.5	158.3 ± 12.3

Table 4 retains all six conditions. Its seed is 20260925, with separate streams for latent survival, privacy coins, and posterior sampling. A preliminary implementation-development run used seed 8917; it is not part of the reported evaluation. The six conditions vary $p_1 \in \{.8, .88\}$ and $\epsilon_2 \in \{.1, .5, 2\}$. The supplied final script generates all conditions, including those where Prefix-KL is inferior. No true p is passed to a policy index or leader calculation. Figure 22 shows the weak-RR trajectories and their final distribution.



(a) **(b)**

Figure 22: **Information routing changes learning on the four-arm instance.** At $\epsilon_2 = .1$, (a) regret trajectories and (b) final regret distributions compare 100 paired streams. Shading gives 95% mean intervals; points show individual streams, with median and interquartile boxes. Propagating confidence bounds alone does not recover the benefit of querying the informative prefix arm.

F SINGLE-CANDIDATE COT ON FROZEN GATES

Fixed selection and observations. We reuse the four datasets, five splits and fitted gates from Section 4.3, selecting the 85th-percentile threshold in every split. This selects the middle of the three existing thresholds by a fixed rule, yielding 20 gates without changing the action ladder or fitting new candidate generators. The frozen clean-label and flipped-label admission indicators are saved

for every pool row. Their means reproduce the archived rates to absolute tolerance 10^{-12}. At each decision the evaluator draws one pool index uniformly, flips its label with probability a_k, looks up that row's deterministic gate outcome, and applies the known RR coin. This is an actual frozen-row query, with one released bit passed to the policy. The evaluator uses pool means solely to compute regret, exact allocations and source accuracy. The policy receives neither these rates nor Equation (8).

All profiles satisfy $.02 < p_4 < \cdots < p_1 < .98$, strict utility unimodality with $s = 3$, and $t_- < .98$. At weak RR the unique exact source is arm 1 in every gate; under full counts it is arm 2. Thus the experiments instantiate the regular model of Corollary 2 on the same records as the batched gate intervention.

Matched oracle ablation and execution. COT, COT-Local, Prefix-KL and OSUB-KL use $\tau = 1$ and the two fixed neighbor releases $\epsilon_2 = .1$ and full count. Other arms release counts. COT retains $(\lambda, \gamma) = (.025, .25)$ from the four-arm experiment. COT-Local changes only the left minimization to $i = s - 1$ at the estimated parameter. Its estimator, likelihood certificate, phase priorities and numerical tolerances are identical. In particular the ablation continues to estimate every arm, so it isolates oracle routing within the same learning architecture. Under homogeneous counts the two choices agree at every structured estimate; their complete trajectories coincide.

Each gate–release–policy cell has 20 streams to $B = 100{,}000$, for 3,200 runs and 128,000 logarithmic checkpoints. Every decision consumes three uniforms in fixed order (pool index, label flip, release flip), including a release coin under full counts. The seed is $20260928 + 10000\,\mathrm{env} + 1000003\,\mathrm{stream}$, with the archived gate index and zero-based stream index. All policies and both releases share these streams. The configuration is saved before execution; short runs use a separate output directory. Gate selection, horizons, seeds and tunings do not depend on regret outcomes.

Metrics and statistical units. The trace records every arm count, the fitted oracle, initialization and all four COT phases, and their separate pseudo-regret. Counts reconstruct total regret as $\sum_i \Delta_i N_i$; phase costs and 1,000-decision block costs independently sum to the same total. Phase names describe COT's schedule; baseline leader choices are not likelihood-certified exploitation. For the main comparison, streams are averaged within each gate, then means and 95% Student-t intervals describe the 20 dataset–split units. COT-Local minus COT uses paired differences of those same unit means. Table 5 reports each dataset over its five split means.

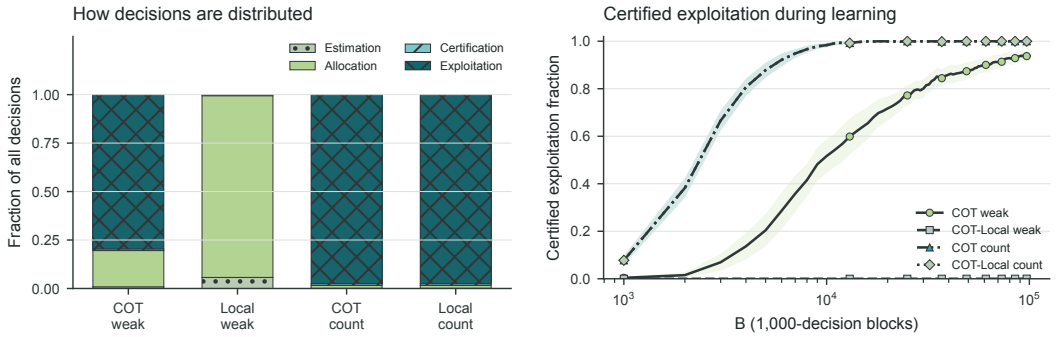
Table 5: Single-candidate frozen-gate regret at 100,000 decisions. Each cell is a mean ± 95% interval over five split means, each averaging 20 streams. The threshold is fixed at the 85th percentile.

Dataset	ϵ_2	COT	COT-Local	Prefix-KL	OSUB-KL
Adult	.1	797.9 ± 325.4	5952.4 ± 713.6	5879.7 ± 944.2	7259.6 ± 94.6
Adult	Count	137.7 ± 42.6	137.7 ± 42.6	87.3 ± 4.1	71.8 ± 4.1
Bank	.1	1506.8 ± 503.2	4981.9 ± 913.6	6031.1 ± 997.2	6405.4 ± 113.5
Bank	Count	159.1 ± 20.9	159.1 ± 20.9	89.6 ± 13.0	69.2 ± 5.8
Spambase	.1	1696.1 ± 861.5	5341.3 ± 1768.4	6067.3 ± 1223.8	6167.8 ± 40.7
Spambase	Count	188.9 ± 28.1	188.9 ± 28.1	82.4 ± 8.4	66.0 ± 2.1
Phishing	.1	1729.3 ± 662.7	5336.7 ± 753.4	5987.6 ± 821.3	6314.3 ± 116.7
Phishing	Count	150.3 ± 61.8	150.3 ± 61.8	95.3 ± 21.5	68.1 ± 4.8

When a run enters sustained certified operation. For COT, source accuracy requires both the correct leader and its exact cheapest prefix source. We evaluate disjoint blocks of 1,000 decisions. A block qualifies when source accuracy is at least .90, leader accuracy at least .95, and certified exploitation fraction at least .80. A run meets the sustained criterion if its final uninterrupted sequence of qualifying blocks has length at least five. The start reported is the first decision of this final sequence, measured retrospectively; this criterion diagnoses finite-budget operation, rather than defining a stopping rule or asserting asymptotic convergence.

Under weak RR, 375/400 runs qualify; among those runs, the median start is 10,001. The other 25 are horizon-censored. Under counts, all 400 qualify, with median start 2,001. The final-block source accuracies are .9425 and 1.0000, and certified-exploitation fractions are .9390 and .9999. The median first certified exploitation decision among runs reaching one is 6,903 under weak RR

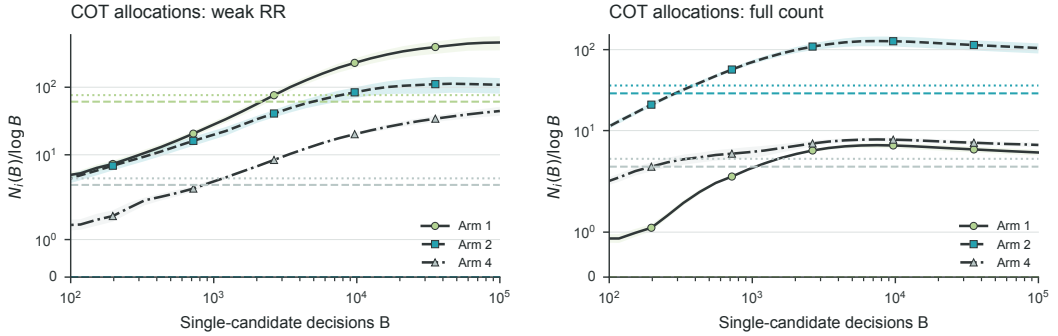
and 1,432 under counts. First certification and sustained operation measure different points in the transient. Figure 23 displays phase counts and certification over time.



(a) **(b)**

Figure 23: COT's learning phases on frozen gates. (a) Mean fraction of the 100,000 decisions spent in each phase; initialization is omitted. (b) Certified-exploitation fractions in successive 1,000-decision blocks, with 95% intervals across 20 gate means. COT-Local remains almost entirely in non-exploitation under weak RR; under counts its curve coincides with COT.

Allocation and remaining transient cost. At the final horizon COT's arm-1 share of left queries is .793 under weak RR and .075 under counts, compared with exact shares one and zero. The mean weak-RR regret decomposes into 269.2 estimation, 1132.0 allocation exploration and 30.6 certification; erroneous exploitation contributes .0053. COT-Local spends 1592.3, 3759.0 and 51.1 in the first three phases. Remote routing therefore saves both direct allocation cost and the additional estimation driven by the non-exploitation clock. Under counts, COT's allocation exploration costs 131.1 and estimation costs 27.1; these account for its higher regret than OSUB-KL. The total weak-RR regret still exceeds the oracle benchmark: identifying the poorly observed neighbor leaves a visible transient even when the source is correct near the end of the run. Figure 24 compares the resulting counts with both allocation targets.



(a) **(b)**

Figure 24: COT suboptimal counts on frozen gates under (a) weak neighbor RR and (b) full counts. Curves are $N_i/\log B$, with 95% intervals across 20 gate means. Dashed lines average the exact c_i; dotted lines average the fixed-tuning $1.25c_i$ quotas over the same gates. The estimation floor and early source changes add counts beyond these targets.

G CONTROLLED HETEROGENEOUS RELEASES ON FROZEN GATES

A setting in which relocation cannot occur. The original 11-arm ladder uses $a \in \{0, .1, \ldots, 1\}$ and $r(a) = .05 + .95a$. In all 60 environments it has $a_i = 0$ for every $i < s - 1$: these arms already have survival above the left tie boundary. No improvement to their feedback quality can make them informative for that constraint. This explains why adding heterogeneity to that benchmark alone

would miss the main effect. Its homogeneous count/RR/sampled-bit/any-survivor comparison instead tests information quantity: channel ordering persists on all four datasets, with all 12 count contrasts significant after Holm correction at $p < .001$ (Table 7). The confidence-certificate experiment makes the distinction concrete: count and RR certify the optimum in every run, while any-survivor bits certify in none within 50,000 candidates (Appendix H).

The file `run_gate_geometry.py` refits the same 60 frozen gates as the homogeneous release study below. It verifies their clean/flipped pool admission rates against the archived values to absolute tolerance 10^{-12}. All policies receive the fixed scales $r = (.5, .75, 1, 1.25)$ and known releases; true p enters only the evaluator. For every batch, fresh pool indices and label-flip uniforms generate the submitted rows, and the frozen gate indicators determine admission. Reusing precomputed deterministic clean/flipped indicators for a row is exactly equivalent to reevaluating the frozen score on that row and label. Latent candidates and flip uniforms use seed 20260926, privacy coins use the offset 191, and all methods and channel conditions reuse these streams. There are $4 \times 5 \times 3 \times 20 \times 4 \times 4 = 19{,}200$ runs, with 1,000 batches of 20 candidates each.

Why this action ladder exposes the prefix. Writing $u = p_{\text{clean}}, v = p_{\text{flip}}$, the four survivals are $(.65u + .35v, .60u + .40v, .55u + .45v, v)$. When $s = 3$, the left boundary is $t_- = (.55u + .45v)/.75$, so

$$t_- - p(a_1) = u/12 + v/4 > 0, \qquad t_- - p(a_2) = 2u/15 + v/5 > 0. \tag{30}$$

Thus both left arms can supply information. The fixed utility scales make the first three utilities strictly increasing; $\mu_3 > \mu_4$ holds when $v/u < 11/16$. The script verifies this inequality, $t_- < 1$, and all ordered model conditions for every gate. These are designed, controlled admission utilities. The learner is given the order constraints, not the evaluator's two-parameter affine formula for $p(a)$; a learner additionally given that formula would face a smaller alternative class and a different lower bound. The original 11-arm linear-scale ladder is retained as a negative geometry control: every non-neighbor left coefficient is zero in its 60 profiles.

Exact versus empirical allocation. The theoretical record includes all four η_k, both one-sided costs, the cheapest left arm, and the neighbor-restricted cost. It evaluates the four executed release conditions and a 61-point logarithmic sweep of $\epsilon_2 \in [.05, 5]$. These calculations use fixed finite-pool admission rates, with no selection by policy performance. Empirical allocations count all submissions, including initialization and forced exploration. For each stream we compute $N_1/(N_1 + N_2)$, then average streams and threshold conditions within split. The theoretical ratio $\eta_1/(\eta_1 + \eta_2)$ is averaged over the same environments. The two quantities compare a finite-horizon policy with a limiting allocation; their discrepancy is retained.

Statistical units. Each cell in Table 6 uses five split means, each averaging 20 streams and three threshold conditions. Paired policy tests use Prefix-KL minus OSUB-KL within those five splits. Holm correction covers all 16 dataset–quality comparisons. The arm-1 share in the main figure averages the 20 dataset–split units of the four fixed datasets. Its 95% interval is descriptive of those units, not of a population of datasets. Under weak feedback, the means favor prefix routing on each dataset; under full count, its extra remote exploration has a finite-budget cost on each.

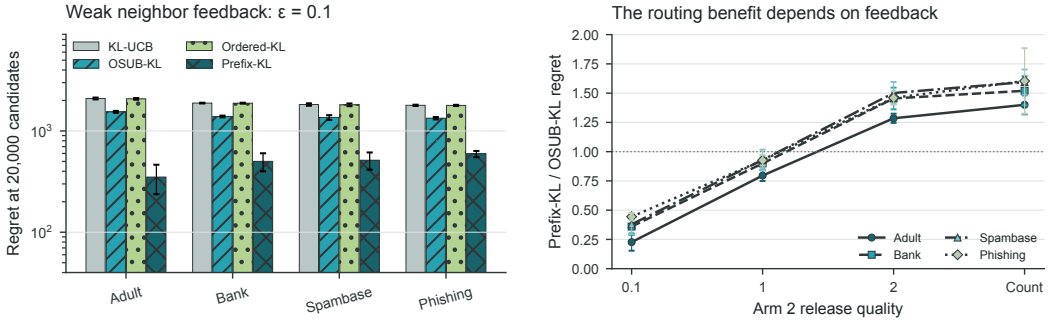
H DIRECT FROZEN-GATE EXPERIMENTS

Adult combines the provided train and test files, giving 48,842 records. Bank Marketing uses the 45,211 records in `bank-full.csv`. Spambase has 4,601 records and Phishing Websites has 11,055. Source archives and extracted data are included. Categorical columns are one-hot encoded without targets. Numerical standardization uses only the gate-training partition; an intercept is added. The gate is an L2-regularized logistic model optimized by L-BFGS. The objective is summed logistic loss plus $\lambda\|w_{1:}\|_2^2/2$, divided by sample size; $\lambda = 2 \times 10^{-4}$ for the gate and one for surrogate/victim models. The intercept is unpenalized. The files contain the exact iteration limits and stopping settings; all stochastic choices are separately seeded.

For split seed $20270 + j$, $j = 0, \ldots, 4$, a permutation makes the disjoint 50/15/20/15 partitions described in the main text. Three clean-score quantiles are calibrated independently from the same gate-calibration partition. Candidates are sampled with replacement from the held-out candidate pool.

Table 6: Actual gate queries with arm-dependent feedback, 20,000 candidates. Only arm 2's release changes; other arms always release counts. Means ± 95% intervals use five split means. All methods use identical candidate and privacy streams. Count denotes full-count feedback on arm 2 as well.

Dataset	ϵ_2	Global KL	OSUB-KL	Ordered-KL	Prefix-KL
Adult	0.1	2093.3 ± 43.8	1545.0 ± 33.5	2074.6 ± 37.9	350.6 ± 113.3
Adult	1	146.6 ± 9.7	143.6 ± 4.9	145.7 ± 5.6	114.2 ± 4.1
Adult	2	74.1 ± 3.0	71.9 ± 2.6	73.9 ± 3.2	92.4 ± 5.2
Adult	Count	57.1 ± 1.5	54.7 ± 1.7	56.2 ± 1.5	76.6 ± 5.1
Bank	0.1	1884.8 ± 15.3	1390.7 ± 27.6	1879.8 ± 19.6	500.9 ± 101.2
Bank	1	151.7 ± 5.8	153.8 ± 7.0	153.5 ± 4.6	138.2 ± 5.8
Bank	2	74.8 ± 3.2	70.9 ± 2.8	74.1 ± 2.2	103.1 ± 5.9
Bank	Count	52.8 ± 3.4	51.0 ± 1.7	52.7 ± 2.4	77.5 ± 5.1
Spambase	0.1	1824.4 ± 49.8	1359.5 ± 71.6	1811.5 ± 56.1	514.0 ± 99.4
Spambase	1	155.1 ± 14.8	153.1 ± 11.1	150.4 ± 11.6	141.8 ± 7.1
Spambase	2	73.2 ± 2.1	69.7 ± 4.8	74.4 ± 2.9	104.5 ± 1.8
Spambase	Count	52.3 ± 1.8	50.8 ± 0.9	52.5 ± 2.0	80.9 ± 4.7
Phishing	0.1	1791.5 ± 25.4	1336.5 ± 35.7	1787.9 ± 23.8	593.3 ± 41.4
Phishing	1	153.3 ± 6.6	153.2 ± 3.1	152.1 ± 5.5	141.8 ± 4.9
Phishing	2	72.7 ± 4.8	71.3 ± 2.6	71.4 ± 4.1	104.1 ± 7.2
Phishing	Count	52.3 ± 2.5	51.6 ± 2.2	52.4 ± 3.0	82.6 ± 13.1



(a) **(b)**

Figure 25: **When remote queries reduce regret.** (a) Regret under weak neighbor feedback ($\epsilon_2 = .1$), comparing all four policies at 20,000 submissions. (b) Prefix-KL/OSUB-KL regret ratios across release qualities; values below one favor Prefix-KL. Means and 95% intervals use five split means per dataset after averaging three thresholds and 20 paired streams. Strong feedback exposes the cost of residual remote exploration.

For each online submission, an independent uniform variable decides whether to flip its true binary label under the selected intensity. The released count is formed from the actual gate decisions on these records. For the sampled bit we use the first candidate in the iid homogeneous batch, distributionally equivalent to a uniformly selected position. Common candidate and flip streams are reused across channels; independent privacy coins are also fixed.

The evaluator computes $p(a)$ exactly for this finite-pool, independent-flip generator using the mean gate indicators for clean and flipped labels. It computes cumulative regret against the best fixed specified utility. Online policies receive neither these means nor test losses. Unlike the earlier single-split benchmark, this experiment does not replace gate evaluations with Bernoulli simulations using fitted arm probabilities. There are 60 frozen-gate environments and 20 paired streams per environment and channel, for 4,800 runs in total. Stream seed 49831 starts the common generator, and the code records all mappings from stream to dataset, split and threshold.

The split-level statistic first averages 20 streams and then the three fixed threshold conditions. Student-t intervals use five such independent means per dataset, with four degrees of freedom. Each channel-versus-count test is a paired t test on those five differences. Holm correction is applied jointly to the 12 comparisons (four datasets times three channels). We do not treat threshold variants

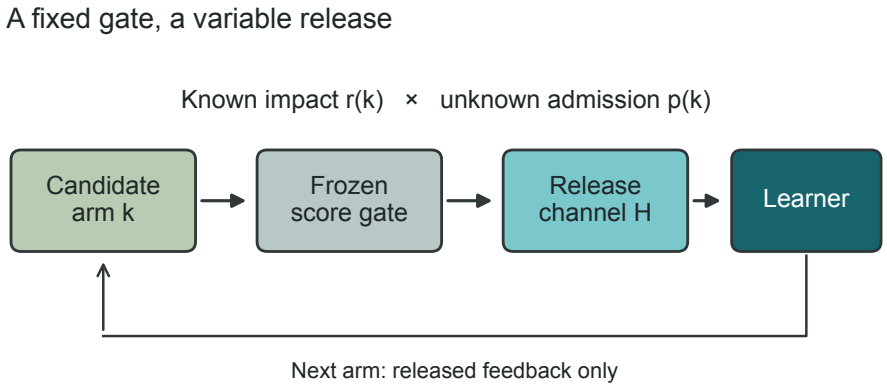


Figure 26: A fixed sanitizer exposes a learning interface through its released receipt. The learner adapts its next generator using only this feedback; admission decisions and utility scales remain fixed when the release channel changes.

or repeated streams from the same fitted gate as new data splits. Figure 27 shows the fixed admission composition; Figures 28–29 show learning under each release on all four datasets.

Table 7: Direct gate-query pseudo-regret at 20,000 submissions. Means ± 95% CI across five independent data splits, after averaging three thresholds and 20 paired streams within each split.

Dataset	Count	RR ($\epsilon = 1$)	Sampled bit	Any survivor
Adult	100.4 ± 4.9	165.0 ± 5.0	243.8 ± 22.4	585.8 ± 34.2
Bank	98.4 ± 7.8	170.2 ± 13.2	253.6 ± 27.4	613.9 ± 48.8
Phishing	91.1 ± 3.1	168.0 ± 13.1	241.5 ± 12.6	556.5 ± 30.1
Spambase	91.5 ± 4.7	164.5 ± 7.3	232.4 ± 19.3	573.1 ± 37.2

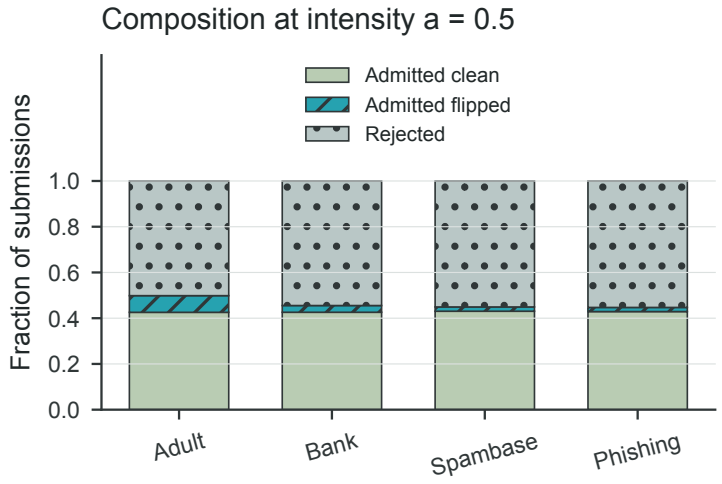


Figure 27: Admitted clean, admitted flipped, and rejected fractions at label-flip probability .5. Bars average five splits at the 85th-percentile clean-score threshold. This composition is unchanged by the release channel.

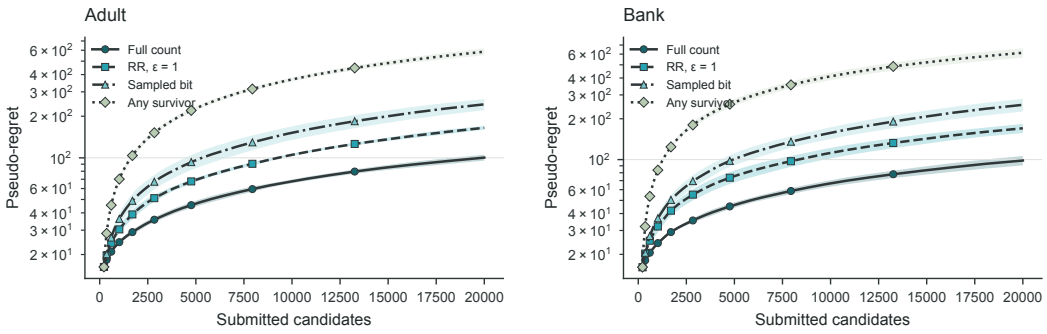


Figure 28: Homogeneous-release learning curves on (a) Adult and (b) Bank. Points average three thresholds and 20 streams within each split; shading gives 95% intervals over five split means. The gate and generators are fixed across channels.

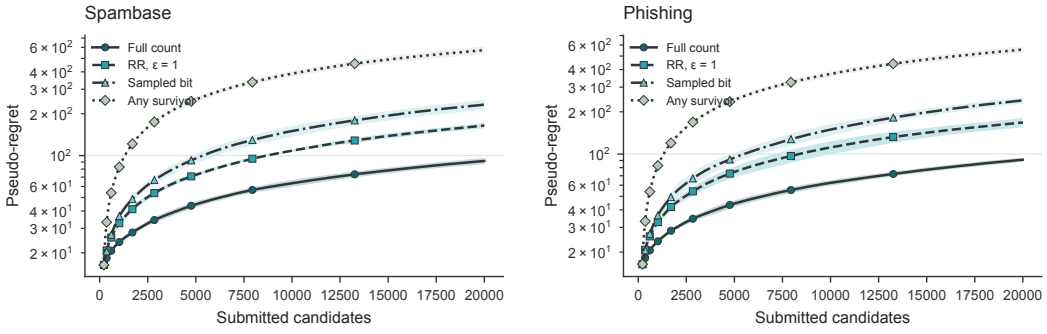


Figure 29: Homogeneous-release learning curves on (a) Spambase and (b) Phishing, using the same protocol as Figure 28. Shading gives 95% intervals over five split means.

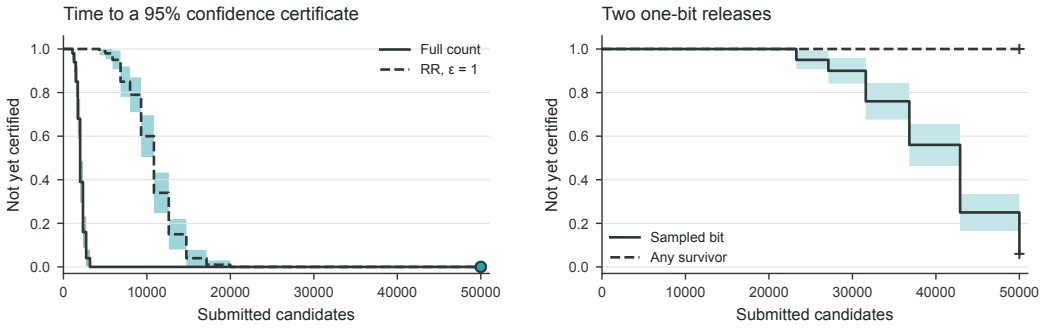
H.1 The confidence-certificate stopping experiment

We use three arms, $r = (.2, .6, 1)$ and $p = (.95, .65, .2)$, and rotate through arms uniformly in batches of 20. Each release channel has 100 streams from seed 8917. The maximum horizon is 2,500 batches. A certificate is checked on 45 logarithmically spaced batch counts (duplicates removed) including the final count. After $n \geq 1$ released Bernoulli sufficient-statistic trials at an arm with empirical mean z, use

$$w_n = \sqrt{\frac{\log(2Kn(n+1)/\delta)}{2n}}, \qquad \delta = .05. \tag{31}$$

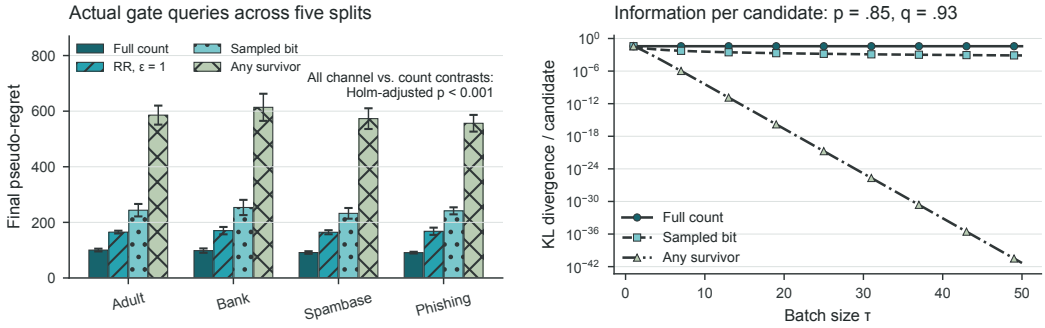
Hoeffding's inequality bounds the two-sided failure probability at (k, n) by $\delta/[Kn(n+1)]$. Summing over every arm and every integer $n \geq 1$ gives simultaneous coverage at least $1-\delta$, including adaptively sampled histories. Invert $[\max(0, z - w_n), \min(1, z + w_n)]$ through the known increasing release map and multiply by r_k. A certificate occurs when one lower utility bound is strictly above every competing upper bound. The event time is the first checked candidate count with a certificate. All unobserved events are administratively right censored at 50,000 candidates.

This experiment uses uniform allocation to isolate release information from policy design. The stopping event requires simultaneous confidence bounds, following the time-uniform coverage principle of confidence sequences (Howard et al., 2021); the explicit union bound above supplies coverage for the bounds executed here. The Kaplan–Meier estimator is $\widehat{S}(t) = \prod_{u \leq t}(1 - d_u/n_u)$, with Greenwood variance and normal pointwise bands clipped to $[0, 1]$. No log-rank independence assumption is used to compare the paired curves. Communication size is one bit for both any-survivor and sampled-bit releases, but their monitoring semantics differ. Figure 30 gives certificate times; Figure 31 connects gate regret to per-candidate information.



(a) **(b)**

Figure 30: Time to a simultaneous 95% optimal-arm certificate under uniform allocation, using 100 paired streams. (a) Count and RR. (b) Two identifiable one-bit releases. Shading gives pointwise Greenwood intervals; a plus marks administrative censoring at 50,000 candidates. Any-survivor feedback produces no certificate within this budget.



(a) **(b)**

Figure 31: Information quantity under homogeneous releases. (a) Regret on 60 frozen gates: means and 95% intervals use five split means per dataset; all 12 channel-versus-count contrasts have Holm-adjusted $p < .001$. (b) Exact per-candidate information at $p = .85, q = .93$: sampled-bit information decays inversely with batch size, while any-survivor information decays exponentially.

I External impact calibration and downstream retraining

Admission interfaces and downstream effects. Attack construction and sanitization study complementary parts of poisoning (Jagielski et al., 2018; Shafahi et al., 2018; Geiping et al., 2021; Steinhardt et al., 2017; Koh et al., 2022; Fu et al., 2024). Influence analysis links perturbations to predictions (Koh & Liang, 2017); shared benchmarks support comparable evaluation (Schwarzschild et al., 2021). Our matched-clean audits test how admission utility transfers to retraining effects.

Admission utility and retraining damage. Survival-only histories leave an unknown impact vector unidentified. For an additive per-admitted-record effect D, the scale is $r_k = \mathbb{E}[D \mid Y = 1, k]$. If target mean effects satisfy $|m_k - \widehat{r}_k p_k| \le \delta$ uniformly, every policy obeys

$$\mathrm{Reg}_{\mathrm{target}}(T) \le \mathrm{Reg}_{\widehat{r}}(T) + 2T\delta. \tag{32}$$

Thus a persistent calibration error can dominate improved admission learning.

The protocol in Figure 33 measures whether conditional impacts transfer across audit model and intervention size, using matched-clean retraining controls.

Matching the audit to the 2,000-candidate intervention improves held-out ranking from $.023 \pm .285$ to $.744 \pm .144$ (Table 8). At equal 66,000-candidate audit budgets, the paired gain is $.600 \pm .290$ ($p = .00036$), over 20 dataset–split units. Calibrated policies retain no general downstream advantage over random allocation (Table 9).

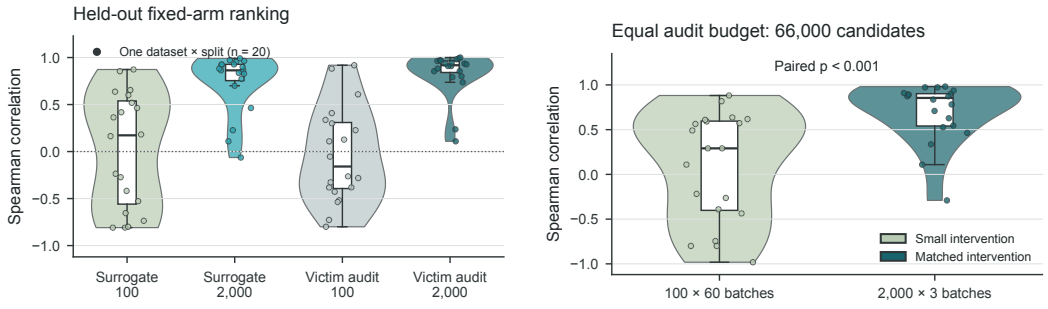


Figure 32: **What better calibration predicts.** (a) Held-out admission-utility/damage ranking correlations across 20 dataset–split units, varying audit model and intervention size. Points are units, with median and interquartile boxes. (b) Equal-total-budget control: both designs use 66,000 candidates, isolating intervention size from audit volume. Improved ranking does not yield a general downstream advantage over random allocation (Table 9).

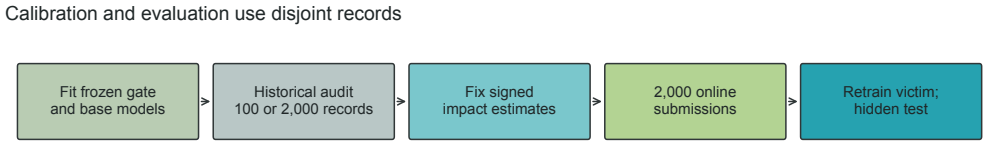


Figure 33: External calibration and evaluation use disjoint records. Historical audits fix the signed impacts before the online interaction and hidden-test evaluation.

Five new partitions per dataset use seed $60400 + j$. The train/gate-calibration/ pilot-candidate/audit-validation/online-candidate/test fractions are 40/10/10/10/15/15. Integer cuts use floors; the remaining records form the test set. The surrogate and victim train on disjoint halves of the first partition, each capped at 1,500 observations. The gate trains on that entire partition, so the gate and victim are different fixed models even when the victim is used for the stronger external-audit condition. Both base models use the same specified regularization. The gate and threshold never change online.

Conditional audit estimator. For each arm k and audit repetition j, sample M records with replacement from the pilot pool, apply the arm's flip generator and the frozen gate, and retain the admitted multiset A_{kj}. Retrain the selected audit model once with the submitted labels and once with the original labels on exactly those features, each appended to the same base training set. On the separate audit validation set, measure the NLL difference D_{kj}. Define

$$\widehat{r}_k = \frac{\sum_j D_{kj}}{\sum_j |A_{kj}|}. \tag{33}$$

The estimator is conditional on admission and is tied to an intervention size M and an audit model. All 11 arms use common candidate indices and flip uniforms within an audit repetition. Main conditions have $M \in \{100, 2000\}$ and 12 repetitions; the nominal costs are 13,200 and 264,000 submitted pilot records respectively, counting repeated draws. Pilot seed is $88400 + j$. These are external calibration costs, not free online observations.

Equal-total-record control. We use 60 surrogate pilot batches of 100 records per arm and three surrogate pilot batches of 2,000 per arm, both totaling 66,000 candidate submissions. The larger-batch condition uses the first three fixed repetitions of the main matched-budget audit. No held-out damage outcomes are used to choose these batches or the impact estimates. The paired comparison uses the 20 within-dataset/split rank-correlation differences; it is not a test treating the 11 arms as independent replicates.

Held-out ranking. Each of the 11 arms uses five held-out streams of 2,000 candidates (seeds 500–504) per dataset and split. Retraining the victim with admitted labels and exactly matched clean labels gives hidden-test NLL differences. We average within arms, compute Spearman correlation with $\widehat{r}_k p_k$ across the 11 arms, and report all 20 dataset–split correlations. Candidate-pool p_k is used only by the evaluator, never by a learned policy.

Online policy comparison. Ten additional paired streams (seeds 0–9) per split compare random allocation, nominal-impact OSUB-KL, calibrated global KL, and calibrated Thompson sampling. Environment and policy generators use distinct fixed seed offsets. Learners receive only $\widehat{r}$ and batch survivor counts, initializing each arm once. Global KL uses the clock and likelihood in Appendix E, taking the lower survival bound for negative impacts. Thompson sampling maximizes signed impact times independent $\mathrm{Beta}(S_k + 1, n_k - S_k + 1)$ draws. Calibrated signed scales can violate order, so these global policies use them without projection.

Table 8: Held-out Spearman correlation (larger is better). Means ± 95% intervals use 20 dataset–split units. Pilot budgets differ between the 100 and 2,000 conditions; Figure 32 gives the equal-budget control.

Audit model and intervention size	Spearman correlation
Surrogate, 100	0.023 ± 0.285
Surrogate, 2,000	0.744 ± 0.144
Victim audit, 100	-0.039 ± 0.237
Victim audit, 2,000	0.835 ± 0.112

Table 9: Excess test NLL relative to matched-clean retraining (% of unaugmented-victim NLL; larger means more damage). Calibrated policies use size-matched surrogate audits (2,000 candidates). Means ± 95% intervals use five split means after averaging ten streams.

Dataset	Random	Nominal OSUB	Calibrated KL	Calibrated TS
Adult	4.24 ± 0.54	3.73 ± 0.77	4.22 ± 0.52	4.45 ± 0.51
Bank	3.78 ± 1.45	3.43 ± 1.34	3.62 ± 1.20	3.64 ± 1.16
Phishing	0.89 ± 2.24	0.72 ± 2.22	1.51 ± 1.35	1.49 ± 1.22
Spambase	1.45 ± 8.84	1.19 ± 8.22	-0.77 ± 5.58	-0.78 ± 5.32

Raw outputs retain all three victim losses, accuracy differences, admission and flip counts, and allocations. Figure 34 distinguishes intervention-specific calibration from transferable impact; positive matched-clean differences need not exceed the unaugmented loss.

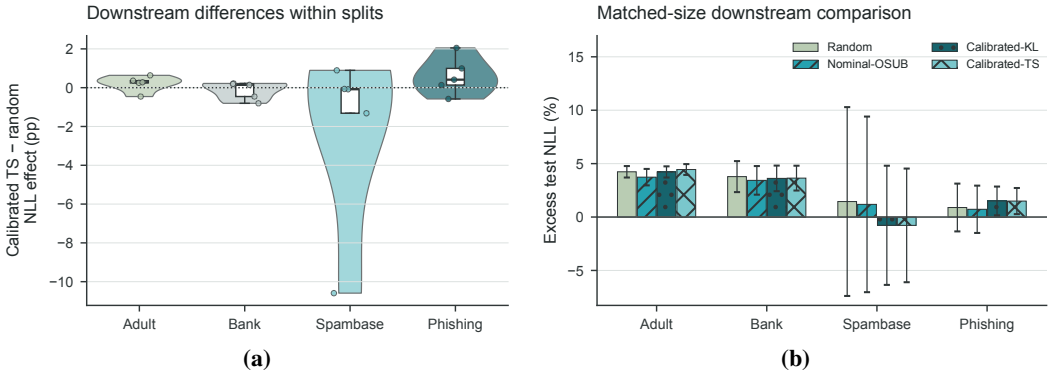


Figure 34: Downstream transfer under matched-size calibration. (a) Calibrated-TS minus random-allocation NLL: points are five split means per dataset, with median and interquartile boxes. (b) Policy comparisons against matched clean retraining, with 95% intervals over those split means. Streams are averaged within each split.

J ADDITIONAL BENCHMARKS

The single-split benchmark uses seed 2027, a logistic gate at the 85th clean-score percentile, a nested label-flip ladder, and impact $.05 + .95a$. Measured arm probabilities drive the Bernoulli bandit simulation. At 50,000 candidates, the 30-stream OSUB-KL regrets are 142.36 ± 14.50 (Adult), 153.28 ± 16.08 (Bank), 158.46 ± 18.08 (Spambase), and 140.67 ± 14.29 (Phishing). OSUB-KL improves over UCB-OSUB on all four and over Thompson sampling significantly only on Adult. Intervals describe streams within one fitted gate (Figure 35).

The 11-arm count environment uses $r = .2 + a$, $p = \text{sigmoid}(4 - 7a)$, $\tau = 20$, and 50 paired streams. At 50,000 candidates, regret is 136.75 ± 9.57 for OSUB-KL and 117.15 ± 5.96 for Thompson sampling.

The original matched-clean downstream experiment uses 30 streams, 2,000 submissions, and a 1,500-record victim base set. For OSUB-KL its excess NLL is $4.10 \pm .64\%$ on Adult, $3.35 \pm .52\%$ on Bank, $-.09 \pm 1.04\%$ on Spambase, and $2.55 \pm .49\%$ on Phishing. Compared with random allocation, the paired differences are -1.003 percentage points (95% interval $[-1.432, -.574]$), .073 ($[-.410, .555]$), -1.094 ($[-1.778, -.410]$), and $-.172$ ($[-.552, .209]$), respectively. These comparisons show no general attack advantage; Appendix I tests the transfer under controlled audit budgets. Figure 36 pairs the original utility profiles with these matched-clean outcomes.

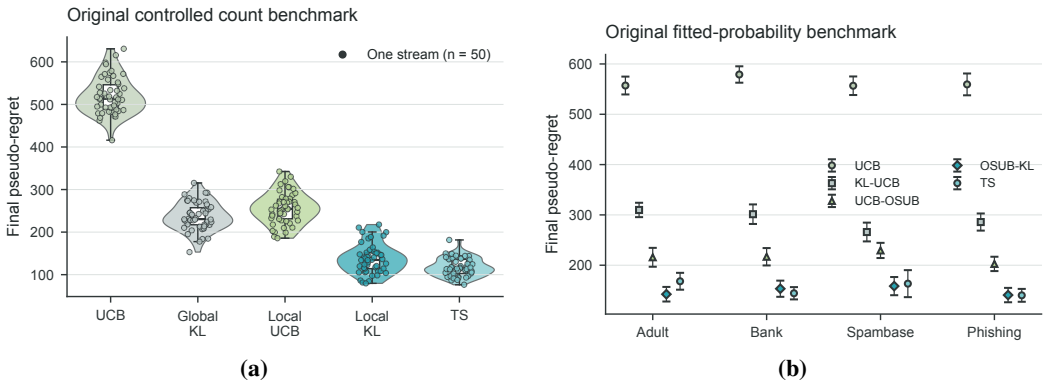


Figure 35: Additional count-feedback benchmarks. (a) The controlled instance uses 50 paired streams; points, violins and boxes show stream regrets. (b) Fitted-probability UCI benchmarks use 30 streams within one fitted gate per dataset; bars show means and 95% intervals.

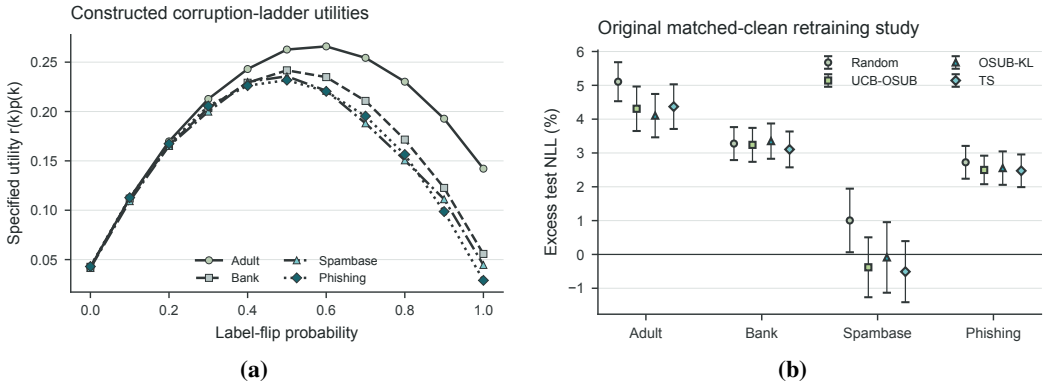


Figure 36: Utility and downstream effects in the original benchmark. (a) Utility profiles combine measured admission probabilities with specified scales. (b) Matched-clean retraining comparisons include random allocation; 95% intervals describe 30 streams within one fixed split. These results separate admission learning from attack efficacy.