THE PRICE OF CALIBRATION IN LEARNING FROM COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study numerical regret when each round permits at most a fixed number of comparisons and at most B of T rounds receive numerical audits. We first classify comparison-only learnability for arbitrary families of continuous symmetric links. Sublinear regret is possible exactly when one polynomial of degree at most the label cap recovers every numerical half-gap up to an unknown positive scale bounded away from zero. The minimax rate is then square-root; otherwise it is linear. Thus known-link reconstruction, shared numerical interpretation, and uniform signal strength are distinct requirements. We then quantify the cost of missing calibration. For a smooth one-parameter family, three labels give uniformly bounded unbiased gap estimation when calibration is known. For fixed K actions and unknown parameter in a radius-r interval, the minimax regret becomes $\Theta_K(\sqrt{T} + T\min\{r, (B+1)^{-1/2}\})$. The lower bound allows audits selected after complete current transcripts. A fixed latent comparison distribution extends the same budget law to a quintic family and every fixed cap $m \geq 5$, with m-dependent constants. In recorded LLM-judgment and human-score replays, a direct-judge baseline outperforms the audited methods on the principal training task; removing its rule does not reverse the reported training differences. These observations separate a worst-case need for numerical information from its finite-task value.

1 INTRODUCTION

A language-model judge can rank two answers without assigning their numerical value. When the goal is to maximize a cumulative rubric score, these are different tasks: many small wins need not compensate for a few large losses. Ranking-feedback bandits study this numerical objective (Maran et al., 2024); instantaneous rankings can make sublinear numerical regret impossible (Liu et al., 2026). Known-link algorithms convert relative judgments into numerical gap estimates when the permitted comparison experiment admits an appropriate statistic.

We study what happens when this link belongs to a known family but its calibration parameter is unknown. Suppose every known member admits a uniformly bounded, exactly unbiased gap statistic. Numerical information is then recoverable at bounded estimation cost once the parameter is given. Can a learner attain the same regret guarantee without that parameter?

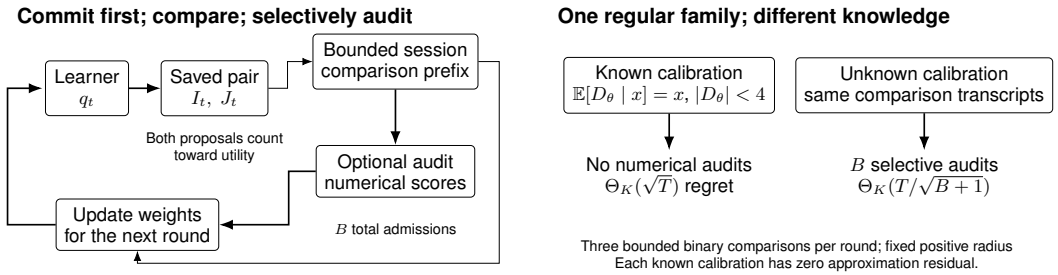
A general classification answers when this transfer is possible: one polynomial of the comparison-sign mean must recover the half-gap under every link, up to a uniformly positive scale. Theorem 3.1 proves necessity as well as sufficiency for every continuous symmetric link family. Nonlinear shape uncertainty can violate this condition even when each individual inverse is a bounded polynomial.

We quantify the resulting audit cost in a smooth one-parameter family. Three binary labels per round suffice for square-root regret under known calibration. Under unknown calibration, the additional regret depends on the declared uncertainty radius r and the budget of B numerically audited rounds. For fixed K actions, the minimax rate is

$$\Theta_K\left(\sqrt{T} + T\min\{r, (B+1)^{-1/2}\}\right).$$

The square-root term is the baseline online-learning regret. The second is the additional worst-case regret due to unknown calibration. A nominal-calibration update and an audit-only update attain the rate by choosing the smaller guarantee before the run.

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.



(a) The comparison-and-audit protocol.

(b) Known and unknown calibration.

Figure 1: **Known and unknown calibration under the same protocol.** In (a), both sampled actions count toward utility; the learner compares them before choosing whether to audit. In (b), three labels give a bounded unbiased half-gap statistic when calibration is known. With unknown calibration and a fixed positive radius, identical comparison transcripts can correspond to different best numerical actions.

The hard environments have fixed calibration and independent utilities, with equal current comparison laws but different best numerical actions. Conditional score information controls audits chosen after those comparisons. Figure 1(a) shows the timing; Figure 1(b) contrasts the knowledge states. Holding the entire latent comparison distribution fixed extends the audit mechanism to a quintic inverse and every fixed cap $m \geq 5$.

Partial monitoring studies what a specified signal system can identify (Bartók et al., 2014; Kirschner et al., 2023); label-efficient learning quantifies the cost of scarce numerical observations (Cesa-Bianchi et al., 2005). Prediction-powered methods use measured outcomes to correct predictions (Angelopoulos et al., 2023; Zrnic & Candes, 2024), and surrogate-reward bandits exploit stochastic structure (Ji et al., 2025). Here we isolate the cost of unknown calibration even when every known member permits bounded exact reconstruction. Recorded-data replays address a complementary question: whether extra audits improve decisions on a particular task. A direct-judge baseline outperforms the audited methods on the principal replay, and removing that judge's own rule does not reverse the reported training differences.

2 Problem Setup

Actions and numerical regret. There are $K \geq 2$ actions. At round t, the learner announces $q_t \in \Delta_K$. The environment fixes $u_t \in [-1, 1]^K$ from the past and q_t, before the learner's fresh draws. The learner samples I_t, J_t independently from q_t and accrues their unobserved average utility. Its external regret is

$$R_T = \max_{i \in [K]} \sum_{t=1}^{T} u_t(i) - \frac{1}{2} \sum_{t=1}^{T} [u_t(I_t) + u_t(J_t)]. \tag{2.1}$$

We bound $\mathbb{E}R_T$, with the hindsight maximum inside the expectation. Both proposals count toward payoff. This timing allows arbitrary predictable utility variation.

Comparisons and numerical audits. Write $x_t = [u_t(I_t) - u_t(J_t)]/2$. Conditional on the entering history and current half-gap, a binary sign W is $+1$ with probability $p_\theta(x_t)$. The comparison-link family is known; θ is fixed for the run. Each round permits at most m conditionally independent labels of the same pair, including early stopping and reversed presentation. Theorem 3.1 treats any fixed $m \geq 1$; the cubic separation uses $m = 3$, and Section 4.3 treats regular shape changes and a quintic instance. A comparison transcript includes the labels and the allowed query choices.

After the current comparisons, the learner may admit the round for a numerical audit, revealing both selected scores. At most $B \in \{0, \ldots, T\}$ rounds are admitted on every path. An admission may require two distinct scalar measurements; these counts are reported separately. The lower bound

grants exact scores. The audit-only update also permits bounded, conditionally unbiased references under Appendix A's conditions. The budgets count observations, not response time.

Known and unknown calibration. Let $V_T^{\mathrm{known}}(\theta, B)$ be the infimum, over learners given θ, of worst-case $\mathbb{E}R_T$ over predictable utility sequences. Let $V_T^{\mathrm{unk}}(r, B)$ instead give the learner only $|\theta| \le r$ and take the supremum over both the parameter and utility sequence. The radius is prior information supplied to the learner. The audit-only upper bound uses neither that radius nor a comparison link.

3 CALIBRATION–AUDIT TRADE-OFF

3.1 WHEN COMPARISONS SUFFICE

Fix a cap $m \ge 1$ and a nonempty family $\mathcal{F}$ of continuous symmetric links $p : [-1, 1] \to [0, 1]$, where $p(-x) = 1 - p(x)$. Let $V_T^{\mathrm{cmp}}(\mathcal{F}, m)$ be the minimax value under Section 2's protocol with a fixed unknown $p \in \mathcal{F}$ and $B = 0$.

Theorem 3.1 (Classification without numerical audits)**.** *For fixed $K, m, \mathcal{F}$, $V_T^{\mathrm{cmp}}(\mathcal{F}, m) = o(T)$ if and only if there are one odd polynomial P of degree at most m and numbers $\lambda_p > 0$ such that*

$$P(2p(x) - 1) = \lambda_p x \quad (p \in \mathcal{F},\ |x| \le 1), \qquad \inf_{p \in \mathcal{F}} \lambda_p > 0. \tag{3.1}$$

In that case the value is $\Theta_{\mathcal{F},K,m}(\sqrt{T})$; otherwise it is $\Theta_{\mathcal{F},K,m}(T)$. The upper bound holds for predictable utilities; the lower bounds use fixed links and independent utility vectors.

The equivalence classifies the family rather than presupposing polynomial inverses. On its positive side, products of j signs estimate $(2p(x)-1)^j$. Combining them with the coefficients of P estimates $\lambda_p x$. The same update therefore optimizes a positive rescaling of numerical utility for every member, without estimating its scale.

For necessity, failure of numerical recovery leaves a component orthogonal to all degree-m comparison moments. Perturbing a symmetric population along that component preserves transcripts and reverses its numerical mean. If every inverse is a polynomial but two are not positive multiples, a common distribution on at most two latent values has opposite numerical means under the two interpretations. Finally, scales approaching zero prevent a uniform guarantee even when each fixed member is learnable. Appendix J proves these exhaustive alternatives and gives the update. For $m \le 2$, the common odd polynomial is linear, recovering Proposition B.1.

3.2 AUDIT COST UNDER SHAPE UNCERTAINTY

We now quantify the audit cost for a family that fails the shared-shape condition while retaining bounded exact estimation under every known calibration. An odd cubic inverse gives the first nonlinear instance. For $|\theta| \le 1/16$, let $z = 1/2 + \theta$ and

$$b_\theta = \frac{8\theta}{z(1 - z^2)}, \qquad g_\theta(v) = 2v + b_\theta(v^3 - v/4), \qquad p_\theta(x) = \frac{1 + g_\theta^{-1}(x)}{2}, \quad |v| \le \tfrac{1}{2}. \tag{3.2}$$

The map g_θ takes $[-1/2, 1/2]$ onto $[-1, 1]$. The links are smooth, strictly increasing, and symmetric, with $p_\theta(-x) = 1 - p_\theta(x)$. Uniformly in θ, x,

$$\tfrac{1}{4} \le p_\theta(x) \le \tfrac{3}{4}, \qquad \tfrac{2}{11} \le p'_\theta(x) \le \tfrac{2}{5}.$$

Thus comparison probabilities and slopes remain bounded away from their degenerate limits.

Exact estimation and unknown calibration. For a specified finite experiment, let $\ell_z(x)$ be a transcript likelihood and write

$$\mathcal{L} = \mathrm{span}\{\ell_z\}, \qquad \varepsilon_{\mathcal{L}} = \inf_{h \in \mathcal{L}} \|x - h(x)\|_\infty. \tag{3.3}$$

Every known member of our family has $\varepsilon_{\mathcal{L}_\theta} = 0$: with $a_\theta = 2 - b_\theta/4$,

$$D_\theta = \frac{a_\theta}{3}(W_1 + W_2 + W_3) + b_\theta W_1 W_2 W_3 \quad \text{has} \quad \mathbb{E}_\theta[D_\theta \mid x] = x, \quad |D_\theta| < 4. \tag{3.4}$$

These inverse polynomials vary in shape, violating the common-polynomial condition in Theorem 3.1 even though every known member has zero approximation residual.

Theorem 3.2 (A sharp calibration–audit separation)**.** *For the family* (3.2)*, with at most three comparisons per round and fixed $K \geq 2$:*

$$V_T^{\text{known}}(\theta, 0) = \Theta_K(\sqrt{T}) \qquad \textit{uniformly in } |\theta| \leq 1/16; \tag{3.5}$$

$$V_T^{\text{unk}}(r, B) = \Theta_K\left(\sqrt{T} + T\min\{r, (B+1)^{-1/2}\}\right) \quad \textit{for } 0 \leq r \leq 1/16,\ 0 \leq B \leq T. \tag{3.6}$$

The constants are independent of T, B, r, θ. Each known member admits an exactly unbiased gap estimator satisfying $|D_\theta| < 4$. The unknown-calibration lower bound allows full current transcripts and outcome-dependent audits; it already holds for independent, identically distributed utility vectors. The upper bound holds for all predictable utility sequences.

Section 4 constructs the indistinguishable populations; Section 5 gives the two algorithms. Appendix B supplies the complete proof, constants, and endpoints.

Interpreting the rate. When $Tr^2 \lesssim 1$, using the central calibration already attains the square-root scale. For larger Tr^2, the additional term remains of order Tr until B reaches order r^{-2}; beyond that scale, it decreases as $T/\sqrt{B+1}$. For any fixed positive r, a uniform square-root guarantee therefore requires $B = \Omega(T)$.

4 Lower bound: indistinguishable comparisons

We now prove the audit lower bound in Theorem 3.2. Unlike the comparison-only classification, this argument must control the numerical information revealed after the learner selects an audit.

4.1 Matched moments and opposite numerical actions

Let X take values $1/2, -1/2, 1, -1$. For each θ we define a distribution μ_θ on this same support, with every mass at least $1/16$. Set

$$M = \frac{3}{8(1-z^2)}, \qquad d = \frac{3}{32z(1-z^2)}.$$

Give the inner pair total mass M and signed imbalance d, and the outer pair mass $1 - M$ and imbalance $-zd$, as in Table 1. The latent comparison mean $V = g_\theta^{-1}(X)$ satisfies

Table 1: The complete hard population, with $z = 1/2 + \theta$. Every world uses the same numerical support. Only its probabilities and the comparison means of the inner pair change. Expanding these masses gives (B.4).

Numerical half-gap X	Comparison-sign mean V	Probability under μ_θ
$+1/2$	$+z/2$	$(M+d)/2$
$-1/2$	$-z/2$	$(M-d)/2$
$+1$	$+1/2$	$(1-M-zd)/2$
-1	$-1/2$	$(1-M+zd)/2$

$$\mathbb{E}_\theta V = 0, \qquad \mathbb{E}_\theta V^2 = 5/32, \qquad \mathbb{E}_\theta V^3 = -3/256. \tag{4.1}$$

The masses make these three moments independent of θ; Appendix B gives the cancellations. Each ordered three-label likelihood is a polynomial of degree at most three in V, so the full transcript distribution is also parameter-independent.

The numerical mean, however, is

$$\mathbb{E}_\theta X = -\theta \, \frac{3}{32z(1-z^2)}. \tag{4.2}$$

Draw independent $X_t \sim \mu_\theta$ and use utilities $(X_t, -X_t)$, duplicating the second type for larger K. The best mean action reverses between θ and $-\theta$. Cross-type pairs have the matched comparison law; same-type pairs give fair labels. All environments retain the same numerical support, so an audit can distinguish them through their score probabilities, rather than through disjoint possible values.

4.2 Information conditional on a selected transcript

The learner may select an audit after observing the current labels. We therefore bound the score information conditional on every possible pre-audit transcript. For a transcript ω,

$$Q_\theta(j \mid \omega) = \frac{w_j(\theta) L_\omega(\theta, j)}{M_\omega},$$

where j indexes the four scores and M_ω is the common transcript probability. The explicit masses and label probabilities give

$$|\partial_\theta \log Q_\theta(j \mid \omega)| \le 8, \qquad \mathrm{KL}(Q_\delta(\cdot \mid \omega) \| Q_{-\delta}(\cdot \mid \omega)) \le 512\delta^2 \quad (\delta \le 1/64). \tag{4.3}$$

This bound includes same-type audits, which can disclose an absolute score at zero half-gap. At a common entering history, the learner's action and audit-selection kernels agree in both environments. Comparisons contribute zero divergence; each admitted score vector contributes at most (4.3). The chain rule gives

$$\mathrm{KL}(\mathbb{P}_\delta^T \| \mathbb{P}_{-\delta}^T) \le 512 B \delta^2.$$

With $\delta = \min\{r, [64\sqrt{B+1}]^{-1}\}$, total variation is at most $1/4$. The mean utility of the first action has opposite signs in the two environments and magnitude at least $\delta/5$. The two-point decision argument yields

$$V_T^{\mathrm{unk}}(r, B) \ge \frac{T}{512} \min\{r, (B+1)^{-1/2}\}.$$

Independent Rademacher utilities supply the separate $\sqrt{T/2}$ lower bound. Taking the maximum of the two bounds gives their sum up to a constant factor. The argument permits joint learning of the calibration, actions, and audit-selection rule.

4.3 Regular shape changes and an explicit instance

Proposition J.2 gives the audit cost of a regular change of inverse-link shape. It considers a jointly C^2 path of odd increasing maps $g_\theta : [-b, b] \to [-1, 1]$, with $0 < b < 1$, $g_\theta(\pm b) = \pm 1$, and $\partial_v g_\theta > 0$. The common endpoints fix the numerical scale. If $\partial_\theta g_\theta|_{\theta=0} \not\equiv 0$, the local lower bound is $\Omega(\sqrt{T} + T \min\{r, (B+1)^{-1/2}\})$ for every fixed cap m and sufficiently small r. When all g_θ are polynomials of degree at most m, the rate is sharp and known calibration gives square-root regret. Constants may depend on the path, K, and m.

The four-point construction matches only three moments of V; its fourth moment varies with calibration. The general argument instead chooses a fixed distribution of V for which $\mathbb{E} g_\theta(V)$ changes sign at zero. Comparisons depend only on V, so every finite transcript has the same law. Common numerical support and smoothness bound the information in selected scores, allowing Proposition I.1 to convert the conflicting means into an audit lower bound.

A quintic map makes this argument explicit and permits known-calibration reconstruction from five labels. Draw V from the fixed density $\nu(v) = 1 + (20v^3 - 3v)/2$ on $[-1/2, 1/2]$, and set

$$\widetilde{g}_\theta(v) = 2v + (\theta - 5/72)(v^3 - v/4) + \tfrac{1}{4}(v^5 - v/16), \qquad \widetilde{p}_\theta(x) = \tfrac{1}{2}[1 + \widetilde{g}_\theta^{-1}(x)]. \tag{4.4}$$

The density is asymmetric: its first moment is zero, while its third and fifth moments are nonzero. The constant $-5/72$ cancels the fixed fifth-moment contribution, leaving $\mathbb{E}\widetilde{X}_\theta = \theta/280$. Products of at most five labels estimate the powers of V in $\widetilde{g}_\theta$. Write $\widetilde{V}_{T,m}$ for this family's minimax values.

Theorem 4.1 (Quintic inverse and larger fixed caps)**.** *For* (4.4)*, fix $K \geq 2$ and $m \geq 5$, and allow at most m conditionally independent same-pair labels per round. Then*

$$\widetilde{V}^{\mathrm{known}}_{T,m}(\theta, 0) = \Theta_K(\sqrt{T}), \qquad \widetilde{V}^{\mathrm{unk}}_{T,m}(r, B) = \Theta_{K,m}\Big(\sqrt{T} + T\min\{r, (B+1)^{-1/2}\}\Big)$$

for $|\theta| \leq 1/16$, $0 \leq r \leq 1/16$, and $0 \leq B \leq T$. Five labels give an exactly unbiased half-gap statistic bounded by $5/2$. The lower bound permits full m-label transcripts and outcome-selected audits with fixed calibration and independent utility vectors; the upper bound allows predictable utilities. Unknown-calibration constants may depend on m.

Appendix I gives the selected-score information bound, its explicit constants, and a bounded nominal statistic with $O(r)$ bias for the matching upper bound.

5 Algorithms and upper bounds

5.1 Nominal comparisons or numerical scores

The nominal route uses the known family centre. For $r > 0$, $D_0 = 2(W_1 + W_2 + W_3)/3$ satisfies

$$|x - \mathbb{E}_\theta[D_0 \mid x]| < 2r.$$

At $r = 0$ it is exactly unbiased. Apply exponential weights with implicit exploration to $(2 - D_0)/2$, whose range is two. The partial-feedback bound plus the pointwise approximation error gives

$$\mathbb{E}R_T \leq 32\sqrt{KT\log K} + 8rT. \tag{5.1}$$

This route uses no numerical audits. With known calibration, replace D_0 by D_θ and its range bound; the bias term vanishes.

The audit-only route uses numerical scores without a comparison link. With k admissions and n rounds left, admit the round with probability k/n, independently of current feedback. This spends exactly B admissions. On admitted rounds, the score update (C.8) corrects both proposal and admission probabilities and uses implicit exploration. Its bound is

$$\mathbb{E}R_T \leq 8\sqrt{\frac{KT(T+1)\log K}{B+1}} + 2\frac{T-B}{B+1}, \qquad B \geq 1. \tag{5.2}$$

The final term accounts for rounds after the last admission. The analysis retains predictable utilities and the random hindsight comparator, using established label-efficient and exponential-weights tools (Auer et al., 2002; Cesa-Bianchi et al., 2005; Neu, 2015).

Algorithm 1 selects the smaller bound before the run using r, B, T, K, without knowing θ. Together with the trivial $2T$ bound, the two routes establish (3.6), including $B = 0$. Appendix C.5 also lets previous comparisons supply predictable advice for the score update; the experiments evaluate that implementation separately.

A common shape with unknown scale. The positive case of Theorem 3.1 uses the same comparison update. Write the family polynomial as $P(v) = \sum_{j=0}^{m} b_j v^j$, set $H = \sum_j |b_j|$, and form $D_P = \sum_j b_j \prod_{\ell=1}^{j} W_\ell$. Its mean is $\lambda_p x$ and its range is at most H. Updating with D_P/H optimizes the rescaled utilities $(\lambda_p/H)u_t$; their best cumulative action is unchanged. The update needs the common coefficients, but neither the actual link nor its scale. Appendix J.4 gives the complete guarantee.

Appendix B.7 also optimizes the budget when an admission has a specified numerical-utility cost.

5.2 Numerical measurements and comparison advice

The nominal route uses an externally justified radius. The audit-only guarantee applies more broadly: it requires bounded, conditionally unbiased numerical scores but no calibrated or repeatable comparison process. For example, a random draw from a fixed list of rubric scores measures

Algorithm 1 An implementable calibration–audit guarantee

Require: T, K, B and a declared radius r for (3.2).
Choose the smaller certified bound in (5.1) and (5.2); use the nominal route if $B = 0$.
Initialize $w(i) = 1$, $k = B$; set $\eta_c = \sqrt{\log K/(4KT)}$ and $\eta_a = \sqrt{(B+1)\log K/[KT(T+1)]}$.
for $t = 1, \ldots, T$ **do**
  Announce $q(i) = w(i)/\sum_j w(j)$; then draw I, J independently from q.
  if the nominal route was selected **then**
    Obtain three conditionally independent labels of the selected pair; set $D_0 \leftarrow 2(W_1 + W_2 + W_3)/3$.
    $\widehat{L}(i) \leftarrow \mathbf{1}\{I = i\}(2 - D_0)/[2(q(i) + 2\eta_c)]$; $\eta \leftarrow \eta_c$.
  else
    $\rho \leftarrow k/(T - t + 1)$; draw $C \sim \mathrm{Ber}(\rho)$ independently of current feedback.
    If $C = 1$, obtain the selected scores, rescale them to $[0, 1]$, and set $k \leftarrow k - 1$.
    Form $\widehat{L}$ by (C.8), with $\gamma = \eta_a/2$; set it to zero if $C = 0$; $\eta \leftarrow \eta_a$.
  end if
  $w(i) \leftarrow w(i) \exp(-\eta \widehat{L}(i))$ for every action.
end for

that list's mean. Scores may be correlated with comparison feedback under the conditions in Appendix A; the admission gate remains independent of current feedback. Systematic conditional bias contributes an additional error term. Comparison advice is formed from previous observations, so it cannot use the current verdict to change an action already taken.

6 Experiments

6.1 Testing the calibration mechanism

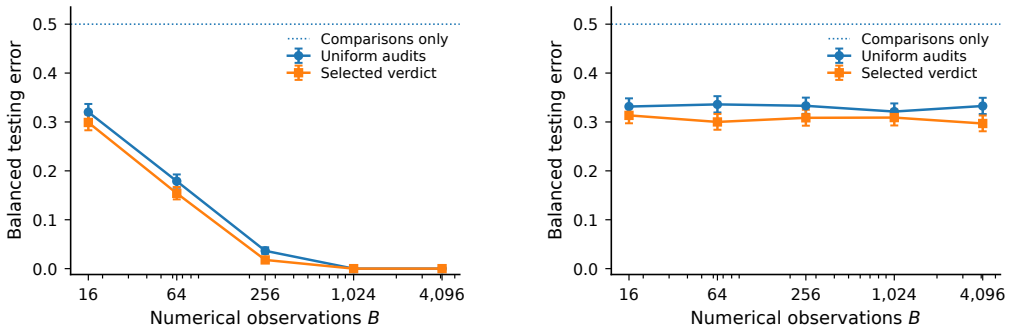
How much does selecting a comparison verdict improve numerical information? We test equally likely worlds $\theta = \delta$ and $-\delta$ from the original three-label construction. A likelihood-ratio test receives B uniformly obtained scores or scores selected after zero positive labels. This category maximizes local audit Fisher information at the centre and is chosen analytically. The selected design receives an unlimited comparison-screening pool, isolating information per purchased score from a fixed-horizon regret problem.

Figure 2(a) fixes $\delta = 1/16$: both tests rapidly identify the environment. Figure 2(b) sets $\delta = 1/(4\sqrt{B+1})$: error remains near 0.33 for uniform and 0.30 for selected audits. Unaudited complete transcripts have balanced error exactly $1/2$. Selection raises local information per audit from 3.081 to 4.113, improving a constant without changing the $B^{-1/2}$ calibration resolution.

6.2 Recorded judgments and a direct-judge baseline

Do extra audits improve decisions on the recorded tasks? We match all 9,125 HelpSteer2 response pairs to numerical scores, individual ratings, human preferences, and RLHFlow judgments from 34 LLMs (Wang et al., 2024; 2025; Li, 2025b;a). The partitions contain 8,677 training and 448 validation pairs. Matching uses full prompt and response text with orientation. An action is a recorded judge's rule for selecting between two saved answers. Its utility is the mean listed helpfulness or correctness score, divided by four. An audit samples one listed score per distinct selected answer. These are replays of existing records, with no new annotations or model queries; Appendix F specifies the access rules.

A frozen diagnostic measures numerical association. Held-out helpfulness-gap RMSE is 0.393 for a zero predictor, 0.337 for GPT-4o's verdict, 0.297 for all 34 verdicts, and 0.155 for human preference strength. Each nonzero predictor uses the same 2,000-pair numerical calibration set, which is unavailable to the online learners. This diagnostic measures prediction error on the records, rather than the theoretical radius r or residual $\varepsilon_{\mathcal{L}}$.



(a) Fixed separation: $\delta = 1/16$.

(b) Local separation: $\delta = 1/(4\sqrt{B+1})$.

Figure 2: **Error in distinguishing the two calibration environments.** Uniform audits and audits selected after zero positive labels use the same three-label family. Panel (a) fixes the separation; panel (b) decreases it with B. Each point uses 4,096 independent tests per world. Error bars combine exact binomial intervals; zero observed error retains a positive upper endpoint.

When the comparison critic is also a candidate rule, that rule selects the critic's preferred answer and never loses its strict comparisons. Inconsistent/error verdicts give neutral updates and precommitted fair fallback choices. Following the existing critic is therefore a direct baseline, specified without numerical test selection. In the nine-rule training-helpfulness replay, it has mean regret 0.92, compared with 90.45 for comparison-only learning and 96.69 for Audit+comparison. Table 2 reports all methods and costs under the matched 64-seed protocol. The separate 128-seed replay in Appendix F.4 is analyzed independently.

Table 2: **Decision quality and observation costs in the nine-rule replay.** Nine-rule training helpfulness, shuffled order, GPT-4o critic, 64 seeds. Audit methods use 20% admissions. Scores and records are mean access counts; regret is mean ± 1.96 Monte Carlo standard errors. Direct critic follows the existing critic; numerical test outcomes are not used to select it.

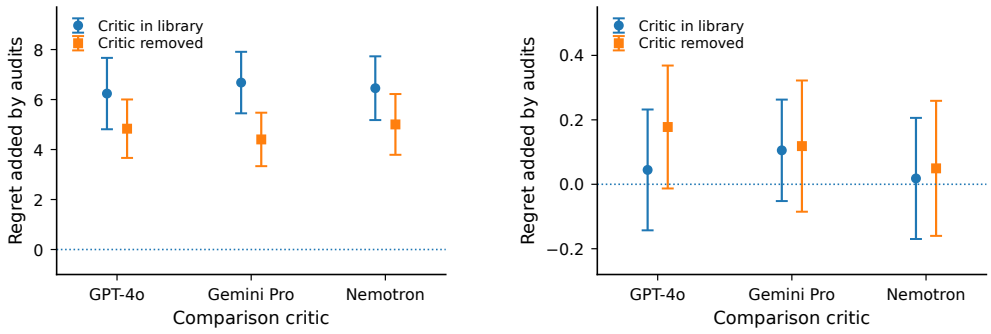
Method	Admissions	Scalar scores	Judge records	Regret
Audit-IX	1,735	2,284	16,286	166.69 ± 3.76
Audit+comparison	1,735	2,187	17,629	96.69 ± 3.27
Comparison-only	0	0	17,395	90.45 ± 2.99
Audited-mean	1,735	1,789	9,432	67.38 ± 11.00
Direct critic	0	0	8,677	0.92 ± 0.53

6.3 Removing the critic's rule

To separate the critic's role as feedback from its role as an action, we repeat the study with GPT-4o, Gemini-1.5-Pro, and Llama-3.1-Nemotron-70B, retaining or removing each critic's own rule. Both partitions use the same nine complete-coverage rules; removal leaves eight, with the comparator recomputed in that library. Methods, learning rates, priors, budgets, and fallbacks are unchanged.

Figure 3(a) reports Audit+comparison minus comparison-only regret at 20% admissions. On shuffled training helpfulness, the increments after removal remain positive: 4.83 ± 1.17, 4.40 ± 1.07, and 5.00 ± 1.22 for the three critics. The corresponding validation intervals in Figure 3(b) include zero. Removing the self-agreement control thus does not reverse the reported training differences; the validation experiment does not resolve their sign. Audited-mean is faster when enough scores are available and more variable at small budgets.

The controls comprise 39,936 sequential method trajectories and 1,536 direct-policy trajectories over two targets, two partitions, shuffled and length-sorted orders, and four budgets. Appendix E.2 reports the full 20% slice; the result files contain every budget and access count. Repeated seeds



(a) Training partition: 8,677 pairs.

(b) Validation partition: 448 pairs.

Figure 3: **Effect of adding audits with and without the critic's rule.** Audit+comparison minus comparison-only regret at 20% admissions; positive values mean higher regret for the audited method. Panels (a) and (b) use different vertical scales. Error bars are paired 1.96 Monte Carlo standard errors over 64 seeds within each fixed library.

quantify algorithmic variation conditional on the records. The comparison tests identify a worst-case information requirement, whereas the replay shows that the evaluated audits need not improve a strong existing decision rule at a finite budget.

7 Related work and conclusion

Known-model observability and Blackwell comparison concern a specified observation system (Blackwell, 1953; Bartók et al., 2014; Lattimore & Szepesvári, 2019). Our constructions give bounded exact reconstruction for each known calibration, while identical comparison laws under different calibrations require opposite numerical decisions. Unknown prediction quality can also impose an adaptation cost (Wei et al., 2020). Active inference and surrogate-reward bandits exploit predictions and references (Zrnic & Candes, 2024; Ji et al., 2025); unknown-link preference optimization obtains positive guarantees under optimization structure (Zhang & Ying, 2025). The present separation concerns numerical external regret with unrestricted predictable utilities and a fixed comparison cap. Appendix G details the objectives and assumptions.

The classification identifies exactly when a fixed comparison cap supports no-audit learning: all links must share polynomial numerical recovery up to uniformly positive scale. The cubic and quintic families violate the shared-shape condition while each known member permits bounded exact estimation; their budget laws quantify the cost of resolving this ambiguity. On the recorded tasks, a direct-judge baseline is strong and audits do not improve the reported training controls; validation differences remain uncertain. Comparison-only learnability, audit complexity, and finite-task value are separate questions with different answers here.

AI use statement

Generative AI tools were used extensively in checking proofs, implementing and debugging code, polishing the manuscript. The submitting authors are responsible for the final text, mathematical claims, code, and interpretation.

Ethics statement

The empirical study reuses public evaluation records and involves no new participant recruitment or model queries. Numerical targets are the means of the specified recorded rubric scores; they are not universal measures of human preference. Human disagreement and position-inconsistent model judgments are retained. The raw files remain unchanged, and the text-based associations used in the

analysis are documented separately. Counts of reused judgments and newly simulated audit requests describe the replay protocol rather than prices or latency observed in deployment.

Reproducibility statement

Appendix B gives the explicit family, population construction, conditional audit-information bound, and matching rates. Appendices A and C specify information sets and numerical updates; Appendix D proves the known-experiment contrast. Appendices E and F describe the calibration tests, original records, direct-critic controls, and complete removal study. Appendix I gives the separation criterion and the quintic extension with complete proofs and explicit constants. Appendix J proves the necessary-and-sufficient comparison-only classification. The supplementary materials include source records, text-based joins, executable experiments, per-run results, and verification scripts. Symbolic and numerical checks accompany the analytic proofs; the proofs have not been formally verified.

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 2023. URL https://doi.org/10.1126/science.adi6000.

Kianoosh Ashouritaklimi, Valentin Kilian, Daolang Huang, Tom Rainforth, and François Caron. Prediction-powered active testing. *arXiv preprint arXiv:2607.08347*, 2026. URL https://arxiv.org/abs/2607.08347.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 2010. URL https://www.jmlr.org/papers/v11/audibert10a.html.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 2002. URL https://doi.org/10.1137/S0097539701398375.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *IEEE Symposium on Foundations of Computer Science*, 2013. URL https://arxiv.org/abs/1305.2545.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 2014. URL https://doi.org/10.1287/moor.2014.0663.

Viktor Bengs, Robert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 2021. URL https://www.jmlr.org/papers/v22/18-546.html.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2011. URL https://proceedings.mlr.press/v15/beygelzimer11a.html.

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

Nicholas Brawand, Nima Leclerc, Anhthy Ngo, Matthew Peterson, Sriram Vishwanath, Laith Alhussein, and Ben Wellner. Active multiple-prediction-powered inference. *arXiv preprint arXiv:2605.08429*, 2026. URL https://arxiv.org/abs/2605.08429.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. URL https://doi.org/10.1017/CBO9780511546921.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 2005. URL https://doi.org/10.1109/TIT.2005.847729.

Chao-Kai Chiang, Takashi Ishida, and Masashi Sugiyama. LLM routing with dueling feedback. *arXiv preprint arXiv:2510.00841*, 2025. URL https://arxiv.org/abs/2510.00841.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W. Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.27414.

Qiwei Di, Jiafan He, and Quanquan Gu. Nearly optimal algorithms for contextual dueling bandits from adversarial feedback. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/di25a.html.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.04475.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011. URL https://arxiv.org/abs/1103.4601.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, 2015. URL https://proceedings.mlr.press/v40/Dudik15.html.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.11665.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3541.

Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *International Conference on Machine Learning*, volume 37, pp. 218–227, 2015. URL https://proceedings.mlr.press/v37/gajane15.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Olga Holtz, Fedor Nazarov, and Yuval Peres. New coins from old, smoothly. *Constructive Approximation*, 33(3):331–363, 2011. doi: 10.1007/s00365-010-9108-5.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024. URL https://arxiv.org/abs/2403.12031.

Wenlong Ji, Yihan Pan, Ruihao Zhu, and Lihua Lei. Multi-armed bandits with machine learning-generated surrogate rewards. *arXiv preprint arXiv:2506.16658*, 2025. URL https://arxiv.org/abs/2506.16658v2. Version 2, April 2026.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, 2013. URL https://proceedings.mlr.press/v28/joulani13.html.

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabás Póczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. In *Advances in Neural Information Processing Systems*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2ba596643cbbbc20318224181fa46b28-Abstract.html.

Michael S. Keane and George L. O'Brien. A Bernoulli factory. *ACM Transactions on Modeling and Computer Simulation*, 1994. URL https://doi.org/10.1145/175007.175019.

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision-making: Algorithms, regret bounds and applications. *arXiv preprint arXiv:2302.03683*, 2023. URL https://arxiv.org/abs/2302.03683.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024. URL https://arxiv.org/abs/2403.13787.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Algorithmic Learning Theory*, volume 98, pp. 529–556, 2019. URL https://proceedings.mlr.press/v98/lattimore19a.html.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. URL https://tor-lattimore.com/downloads/book/book.pdf.

Min Li. Interpreting language model preferences through the lens of decision trees. RLHFlow technical blog, January 22, 2025a. URL https://rlhflow.github.io/posts/2025-01-22-decision-tree-reward-model/.

Min Li. LLM-Preferences-HelpSteer2. RLHFlow dataset, CC-BY-4.0, 2025b. URL https://huggingface.co/datasets/RLHFlow/LLM-Preferences-HelpSteer2.

Puheng Li, Tijana Zrnic, and Emmanuel Candès. Robust sampling for active statistical inference. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2511.08991.

Xuheng Li, Heyang Zhao, and Quanquan Gu. Feel-good thompson sampling for contextual dueling bandits. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/li24co.html.

Mingyang Liu, Yongshan Chen, Zhiyuan Fan, Gabriele Farina, Asuman E. Ozdaglar, and Kaiqing Zhang. Online learning and equilibrium computation with ranking feedback. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.19221.

Virginia L. Ma and Emmanuel J. Candès. Optimized labeling resource allocation for prediction-assisted inference via OPAL. *arXiv preprint arXiv:2606.03211*, 2026. URL https://arxiv.org/abs/2606.03211.

Davide Maran, Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Nicola Gatti, and Marcello Restelli. Bandits with ranking feedback. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/936ce22b767cf1a1496083e4725d3b21-Abstract-Conference.html.

Siddharth Mitra and Aditya Gopalan. On adaptivity in information-constrained online learning. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/5964.

Şerban Nacu and Yuval Peres. Fast simulation of new coins from old. *The Annals of Applied Probability*, 15(1A):93–115, 2005. doi: 10.1214/105051604000000549.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.03271.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*, 2024. URL https://arxiv.org/abs/2406.18665.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.02155.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled RLHF from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024. URL https://arxiv.org/abs/2405.00254v1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, volume 30, 2013. URL https://proceedings.mlr.press/v30/Rakhlin13.html.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/saha22a.html.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *Algorithmic Learning Theory*, 2022. URL https://proceedings.mlr.press/v167/saha22a.html.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, volume 139, pp. 9235–9244, 2021. URL https://proceedings.mlr.press/v139/saha21a.html.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Dueling convex optimization with general preferences. In *International Conference on Machine Learning*, volume 267, pp. 52552–52564, 2025. URL https://proceedings.mlr.press/v267/saha25a.html.

Noa Shoham, Ron Dorfman, Shalev Shaer, Kfir Y. Levy, and Yaniv Romano. Prediction-powered semi-supervised learning with online power tuning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.22586.

Bas van Opheusden, Luigi Acerbi, and Wei Ji Ma. Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLOS Computational Biology*, 16(12):e1008483, 2020. doi: 10.1371/journal.pcbi.1008483.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023a. URL https://arxiv.org/abs/2305.17926.

Xuchuang Wang, Qingyun Wu, Wei Chen, and John C. S. Lui. Multi-fidelity multi-armed bandits revisited. In *Advances in Neural Information Processing Systems*, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/64602b87c31db70a3ef060f6c5d5b01d-Abstract-Conference.html.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024. URL https://arxiv.org/abs/2406.08673.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. HelpSteer2-Preference: Complementing ratings with preferences. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.01257.

Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3583–3634, 2020. URL https://proceedings.mlr.press/v125/wei20a.html.

Yihong Wu and Pengkun Yang. Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends in Communications and Information Theory*, 17(4):402–586, 2020. doi: 10.1561/0100000095. URL https://arxiv.org/abs/2104.07317.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. *Journal of Computer and System Sciences*, 2012. URL https://doi.org/10.1016/j.jcss.2011.12.028.

Qining Zhang and Lei Ying. Sign-SZPO: Provable preference-based reinforcement learning with an unknown link function. *arXiv preprint arXiv:2506.03066*, 2025. URL https://arxiv.org/abs/2506.03066v2. Version 2, July 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05685.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, 2020. URL https://proceedings.mlr.press/v108/zimmert20a.html.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *International Conference on Machine Learning*, volume 235, pp. 62993–63010, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

APPENDIX CONTENTS

A Information sets and reference measurements **17**

B The calibration separation: construction and complete proof **18**

B.1 A smooth family with a uniformly bounded inverse statistic 18

B.2 Utility populations that conceal the calibration parameter 18

B.3 The information in an outcome-selected audit 19

B.4 Regret lower bound and the uncertainty radius 19

B.5 Matching nominal and audit-only upper bounds 20

B.6 Minimality within symmetric binary sessions 20

B.7 Optimizing the cost of a numerical budget 21

C Algorithms and regret guarantees **22**

C.1 From comparison statistics to numerical updates 22

C.2 Implicit exploration with a measurement range 23

C.3 The exact-budget gate 23

C.4 Channel combination and the hindsight comparator 24

C.5 Audit-only scores and practical advice 25

D The known-experiment benchmark **25**

D.1 Proof of the known-experiment benchmark 26

E Calibration and critic-overlap experiments **28**

E.1 Analytic identities and selected-audit tests 28

E.2 Direct-critic controls and removal ablations 29

F Recorded-data construction and experimental protocol **31**

F.1 Source records, matching, and the numerical target 31

F.2 Frozen numerical-prediction diagnostics 33

F.3 Online replay: observations and cost accounting 33

F.4 The separate 128-seed replay 36

G Connections to the closest literature **36**

G.1 Relative feedback and numerical objectives 36

G.2 Observability, experiments, and scarce labels 37

G.3 Correction, multi-fidelity information, and active evaluation 38

G.4 Unknown links and adaptation to predictions 39

G.5 Finite sessions, evaluation, and routing 39

H Scope and remaining questions **39**

I A separation criterion and a quintic extension **40**

I.1 A sufficient condition for a calibration–audit separation 40

I.2 A fixed latent comparison distribution 42

I.3 Proof of the quintic extension 43

J Classification of comparison-only learnability **44**

J.1 What a finite session observes 45

J.2 Failure to recover a numerical gap under a known link 45

J.3 Unknown nonlinear shape creates a second obstruction 46

J.4 A common statistic gives an implementable upper bound 47

J.5 Why uniform signal strength is part of the classification 47

J.6 Audit complexity near a regular change of shape 48

J.7 Reading the criterion for a comparison family 49

A INFORMATION SETS AND REFERENCE MEASUREMENTS

Let $\mathcal{H}_{t-1}$ contain the learner's previous observations and private randomization. Current advice, when used, is predictable from this history. After the learner announces q_t, enlarge the history to $\mathcal{G}_t$ by including the environment's current utility choice. This enlarged information is used for the analysis, not revealed to the learner. Conditional on $\mathcal{G}_t$, the two proposals are independent draws from q_t. Sections 3–4.3 use conditionally independent labels from a single comparison link. The known-experiment benchmark in Appendix D also allows finite sessions with dependent labels. In that benchmark, the prescribed conditional joint law can depend on the queried half-gap and preceding transcript.

An allowed deterministic decision tree includes all of its prefixes. The learner can select its tree using its entering history and private randomness. For example, a finite menu of judges, a finite output alphabet, and a fixed bound on the number of within-round queries produce a finite set of deterministic trees. A transcript includes query choices and returned symbols, so that hiding a query choice does not create artificial indistinguishability. Reversing the pair is included in the allowed menu. The entire nonnumeric feedback used in the lower-bound model is represented by these transcripts. An additional side measurement correlated with the utility would have to be added to the experiment, potentially enlarging $\mathcal{L}$.

Half-gaps and absolute scores. The half-gap $x = (u(i) - u(j))/2$ lies in $[-1, 1]$. An audit reveals exact selected scores in the information theorem and in the calibrated residual algorithm. It is enough for that algorithm to reveal the exact half-gap, but granting both scores makes the lower bound stronger. An identical-action audit can reveal an absolute score even though its half-gap is zero. Each round may have at most one audit admission. The cost of the reference measurements within that admission is separately reported in the experiments.

Expected realized regret. Define $\overline{R}_T = \max_i \sum_t u_t(i) - \sum_t \langle q_t, u_t \rangle$. Then

$$R_T - \overline{R}_T = \sum_t \left(\langle q_t, u_t \rangle - \tfrac{1}{2}[u_t(I_t) + u_t(J_t)] \right)$$

is a sum of martingale differences with respect to the expanded round history. Therefore $\mathbb{E}R_T = \mathbb{E}\overline{R}_T$, even for an adaptive environment and a random best action. The upper proof below controls the expected maximum estimation error simultaneously over all fixed actions.

Noisy reference scores. For the audit-only specialization, exact scores can be replaced by bounded measurements $Y_t(i) \in [0, 1]$ satisfying

$$\mathbb{E}[Y_t(i) \mid \mathcal{G}_t, I_t, J_t, C_t = 1] = u_t(i) \tag{A.1}$$

for each selected i, with $u_t \in [0, 1]^K$ for this score convention. The right side defines the numerical target. One design-based construction fixes an arbitrary finite list of potential rubric scores for each current candidate before the fresh draws, defines its utility as that list's mean, and samples a list entry uniformly on admission. Lists can have different lengths and arbitrary dependence among their entries. The learner's private sampling supplies (A.1).

Two score measurements may be correlated, and an identical output may reuse the same measurement. The score-IX proof uses convexity rather than independence of those measurements. Comparison feedback may also depend on the scores; advice for the current decision uses only previous observations. The admission gate is independent of the joint current feedback. The upper algorithm consequently uses a feedback-independent gate. The lower bound in Theorem 3.2 additionally allows audits selected from current verdicts.

If the conditional means in (A.1) differ from the intended target by at most β_t for all selected actions, the score-regret proof gains at most $2\mathbb{E}\sum_t \beta_t$. The two terms come from the learner's mixture and the comparator. Random admission corrects selective observation, not an arbitrary bias of the measurement device.

B THE CALIBRATION SEPARATION: CONSTRUCTION AND COMPLETE PROOF

We prove Theorem 3.2. Throughout, x denotes a half-gap, and a binary label is encoded by $W \in \{-1, 1\}$. An audit reveals exact selected scores. All constants below are uniform over the specified one-parameter family.

B.1 A SMOOTH FAMILY WITH A UNIFORMLY BOUNDED INVERSE STATISTIC

Let $\theta \in [-1/16, 1/16]$, $z = 1/2 + \theta$, and define

$$b_\theta = \frac{8\theta}{z(1-z^2)}, \qquad a_\theta = 2 - \frac{b_\theta}{4}, \qquad g_\theta(v) = a_\theta v + b_\theta v^3, \quad -\tfrac{1}{2} \le v \le \tfrac{1}{2}. \tag{B.1}$$

On $z \in [7/16, 9/16]$, $z(1-z^2) > 1/3$. Hence $|b_\theta| \le 24|\theta| \le 3/2$ and

$$g_\theta(\pm\tfrac{1}{2}) = \pm 1, \qquad 5/4 \le g'_\theta(v) = 2 + b_\theta(3v^2 - 1/4) \le 11/4.$$

Consequently g_θ has a smooth increasing odd inverse v_θ on $[-1, 1]$. Set $p_\theta(x) = (1 + v_\theta(x))/2$. This is a smooth strictly increasing symmetric comparison law, with

$$\tfrac{1}{4} \le p_\theta(x) \le \tfrac{3}{4}, \qquad \tfrac{2}{11} \le p'_\theta(x) \le \tfrac{2}{5}. \tag{B.2}$$

Its derivatives are uniformly bounded on this compact parameter and gap domain. The probability, slope, and estimator bounds are uniform over this domain.

If W_1, W_2, W_3 are conditionally independent signs with mean $v_\theta(x)$, then

$$D_\theta = \frac{a_\theta}{3}(W_1 + W_2 + W_3) + b_\theta W_1 W_2 W_3 \quad \text{satisfies} \quad \mathbb{E}_\theta[D_\theta \mid x] = x, \quad |D_\theta| \le |a_\theta| + |b_\theta| < 4. \tag{B.3}$$

Thus every fixed member has zero known-law approximation residual. A known-θ learner can use the comparison channel of (C.3), with $A = 4$ and $e = 0$, to obtain at most $64\sqrt{KT \log K}$ expected regret without an audit. Using the actual range gives the same order with a smaller constant. The proof of that update treats the random hindsight comparator, as required by (2.1).

B.2 UTILITY POPULATIONS THAT CONCEAL THE CALIBRATION PARAMETER

For each $z = 1/2 + \theta$, put mass on the same four half-gaps $(1/2, -1/2, 1, -1)$:

$$\begin{aligned} w_{+1/2}(z) &= \frac{3(4z+1)}{64z(1-z^2)}, \quad w_{-1/2}(z) = \frac{3(4z-1)}{64z(1-z^2)}, \\ w_{+1}(z) &= \frac{17 - 32z^2}{64(1-z^2)}, \qquad w_{-1}(z) = \frac{23 - 32z^2}{64(1-z^2)}. \end{aligned} \tag{B.4}$$

These masses sum to one and each is at least $1/16$ on the parameter interval. For example, the smallest numerator in the second fraction is $9/4$, while its denominator is at most 36. The other three lower bounds follow immediately from the displayed formulas. Let μ_θ denote this distribution.

Write $M_{1/2}$ and M_1 for the total masses on the two absolute gaps, and $d_{1/2}$ and d_1 for the positive-minus-negative mass differences. Direct simplification gives

$$M_{1/2} = \frac{3}{8(1-z^2)}, \quad M_1 = 1 - M_{1/2}, \quad d_{1/2} = \frac{3}{32z(1-z^2)}, \quad d_1 = -z d_{1/2}.$$

Since $g_\theta(z/2) = 1/2$ and $g_\theta(1/2) = 1$, the latent comparison mean $V = v_\theta(X)$ takes values $\pm z/2, \pm 1/2$. Its first three moments are

$$\mathbb{E}_\theta V = 0, \qquad \mathbb{E}_\theta V^2 = \frac{5}{32}, \qquad \mathbb{E}_\theta V^3 = -\frac{3}{256}, \tag{B.5}$$

independent of θ. Every ordered triple of independent binary labels conditional on X has a likelihood that is a polynomial of degree at most three in V. Thus the complete three-label law is identical for all θ. Prefixes, randomized early stopping, and reversed presentation orders cannot add information: generate the full canonical triple first and reveal or reverse its entries as requested. Same-type action pairs have half-gap zero and fair labels, also independent of θ.

In contrast, the numerical mean is

$$\mathbb{E}_\theta X = (\tfrac{1}{2} - z)d_{1/2} = -\theta d_{1/2}, \qquad d_{1/2} > \tfrac{1}{5}. \tag{B.6}$$

For the last inequality, $z(1 - z^2) \le 2/(3\sqrt{3}) < 2/5$ implies $d_{1/2} > 15/64$. Use independent $X_t \sim \mu_\theta$ and utilities $(X_t, -X_t)$, duplicating the second action for $K > 2$. All four utility values have uniformly positive probabilities. The link and these probabilities change together, preserving comparisons while reversing the best mean action.

B.3 The information in an outcome-selected audit

We grant the learner the full three-label transcript before every audit decision; this is at least as informative as any allowed stopping or orientation strategy. For an ordered prefix or triple ω with common marginal mass $M_\omega > 0$, its conditional exact-score distribution is

$$Q_\theta(j \mid \omega) = \frac{w_j(\theta)L_\omega(\theta, j)}{M_\omega}. \tag{B.7}$$

Here j indexes the four gap values and L_ω is its conditional comparison likelihood. The denominator is independent of θ, by (B.5). The complete selected score vector is a known function of X_t and also recovers X_t, including for same-type proposals.

To control score information after selection, differentiate (B.7) at each possible transcript. For the two inner masses, cancellation gives

$$\frac{d}{dz}\log w_{+1/2} = -\frac{1}{z(4z+1)} + \frac{2z}{1-z^2}, \qquad \frac{d}{dz}\log w_{-1/2} = \frac{1}{z(4z-1)} + \frac{2z}{1-z^2}.$$

Their absolute values are less than 2 and 5, respectively. For the outer masses the derivatives simplify to $-30z/[(17 - 32z^2)(1 - z^2)]$ and $-18z/[(23 - 32z^2)(1 - z^2)]$, with absolute values less than 4 and 2. Each inner-gap label probability is $1/2 \pm z/4$, so each label contributes at most $16/23$ to the absolute logarithmic derivative. Outer-gap probabilities are constant. Therefore

$$|\partial_\theta \log Q_\theta(j \mid \omega)| \le 8 \tag{B.8}$$

for every possible pre-audit transcript. The same bound holds for a same-type proposal, where the conditional law is simply $w(\theta)$.

Let $0 < \delta \le 1/64$. Equation (B.8) implies

$$\left|\log \frac{Q_\delta(j \mid \omega)}{Q_{-\delta}(j \mid \omega)}\right| \le 16\delta, \qquad \mathrm{KL}(Q_\delta(\cdot \mid \omega) \| Q_{-\delta}(\cdot \mid \omega)) \le (e^{16\delta} - 1)^2 \le 512\delta^2.$$

The middle inequality follows by bounding KL by chi-squared divergence; the last uses $16\delta \le 1/4$ and $256e^{1/2} < 512$. Uniformity over the transcript lets us apply the same bound after the learner selects an audit.

For any adaptive learner, the current cheap transcript has the same conditional law in the two worlds at each common entering history. Actions and the audit decision use the same learner kernel there. Only an admitted numerical observation contributes to the chain-rule divergence. With at most B admissions on every path,

$$\mathrm{KL}(\mathbb{P}_\delta^T \| \mathbb{P}_{-\delta}^T) \le 512B\delta^2. \tag{B.9}$$

The common-history calculation covers simultaneous calibration learning, action choice, and audit selection under the fixed parameter.

B.4 Regret lower bound and the uncertainty radius

Suppose the declared family is $|\theta| \le r$, where $0 < r \le 1/16$, and choose

$$\delta = \min\left\{r, \frac{1}{64\sqrt{B+1}}\right\}.$$

Then (B.9) is at most $1/8$, so Pinsker gives total variation at most $1/4$ for the full history and all its prefixes. Let $s_t \in [-1, 1]$ be the average signed action type. The mean utility of the first action has

opposite signs in the two environments and magnitude at least $\delta/5$. Since current X_t is independent of the past and the fresh proposals, and the hindsight comparator is at least the best mean action,

$$\frac{1}{2}(\mathbb{E}_\delta R_T + \mathbb{E}_{-\delta} R_T) \geq \frac{\delta}{5} \sum_{t=1}^{T} \left(1 + \frac{\mathbb{E}_\delta s_t - \mathbb{E}_{-\delta} s_t}{2}\right) \geq \frac{3\delta T}{20}.$$

Using $\min\{r, a/64\} \geq \min\{r, a\}/64$ yields

$$V_T^{\mathrm{unk}}(r, B) \geq \frac{T}{512} \min\{r, (B+1)^{-1/2}\}. \tag{B.10}$$

A separate independent Rademacher utility sequence under any one family member gives $\sqrt{T/2}$, even when every past score is revealed. Taking the maximum of the two lower bounds is equivalent up to constants to their sum. The $r = 0$ endpoint is just the known central link and retains this square-root baseline.

B.5 Matching nominal and audit-only upper bounds

The centre-link statistic $D_0 = 2(W_1 + W_2 + W_3)/3$ is implementable without knowing θ. If the actual link is p_θ and $r > 0$, then

$$|x - \mathbb{E}_\theta[D_0 \mid x]| = |b_\theta(v^3 - v/4)| \leq 24r \cdot \frac{1}{12\sqrt{3}} < 2r.$$

At $r = 0$ the error is exactly zero. Apply the one-channel implicit-exploration proof in Appendix C, with range $A = 2$. Its expected loss corresponds to numerical utility up to an error of at most $2r$ in every half-gap. The mixture-versus-comparator difference costs at most $8r$ per round in the utility convention of (2.1). Hence an implementable no-audit policy satisfies

$$\mathbb{E}R_T \leq 32\sqrt{KT \log K} + 8rT. \tag{B.11}$$

The learner selects this bound using the declared family radius.

Alternatively, the audit-only score update (C.8) with uniform-subset admissions gives, after rescaling $[-1, 1]$ to $[0, 1]$,

$$\mathbb{E}R_T \leq 8\sqrt{\frac{KT(T+1)\log K}{B+1}} + 2\frac{T-B}{B+1}, \qquad B \geq 1. \tag{B.12}$$

It uses exactly B admissions, no comparison-law input, and allows arbitrary predictable utility sequences. For $B = 0$ use (B.11) or the trivial $2T$ bound. Choose the smaller available guaranteed upper bound before the run. Equations (B.10)–(B.12) prove the stated joint rate, uniformly over T, B, r for fixed K. Both routes use the inputs specified in Algorithm 1.

Uniformity and scope. The rate constants are uniform as $r \downarrow 0$ and may depend on fixed K. The audit-only guarantee applies to comparison processes beyond this family. The lower bound uses the specified parametric construction; the recorded-data experiments evaluate their own numerical targets without fitting this family.

B.6 Minimality within symmetric binary sessions

The restriction to three comparisons is the first nontrivial nonlinear case in the following exact-reconstruction class.

Proposition B.1 (One or two labels need no unknown positive scale)**.** *Consider a family of continuous strictly increasing symmetric binary links on $[-1, 1]$. Suppose every known member permits an unbiased gap statistic using at most $m \leq 2$ conditionally independent labels, of absolute range at most the same A. Then a learner not told the link attains $O(A\sqrt{KT \log K})$ numerical regret without audits. No analogue of the fixed-radius separation in Theorem 3.2 is possible within this class.*

Proof. Put $v = 2p(x) - 1$. A stopped binary experiment of depth at most two can be completed with unused independent labels. Its statistic has expectation equal to a polynomial $g(v)$ of degree at most two. Strict monotonicity implies that the feasible v interval has nonempty interior; symmetry of the inverse implies $g(-v) = -g(v)$ throughout that interval. The constant and quadratic coefficients therefore vanish, so $x = av$ for some $a > 0$.

Average the statistic over exchange of its labels and over sign reversal with output negation. These operations preserve unbiasedness and its bound A. For one label the resulting statistic is aW_1. For two labels, its values at $(+,+)$, mixed signs, and $(-,-)$ are $a, 0, -a$, respectively, by uniqueness of the degree-two Bernstein representation. Thus $a \le A$, including for a privately randomized original statistic by taking its conditional expectation on the completed transcript.

Use a single sign W with the one-channel update applied to $(1 - W)/2$ and a learning rate set for range one. Conditional on the announced distribution, its expected loss for action i is $1/2 - [u_t(i) - \langle q_t, u_t\rangle]/(4a)$. The implicit-exploration regret argument therefore scales by the fixed positive factor $4a$; it does not require a as an input. The bound is at most $16A\sqrt{KT \log K}$ with the conservative constants used here. Fixed positive scaling preserves the hindsight-best numerical action, even for a predictable utility sequence. □

The proposition concerns identical binary labels from one symmetric link and exact known-link reconstruction. A panel of different judges or a different numerical objective is a different experiment. The cubic inverse in (B.1) is the first odd polynomial beyond this positive-scale case.

B.7 Optimizing the cost of a numerical budget

A cost per audited round. For admission cost $\lambda \ge 0$ in numerical-utility units, optimizing the budget gives

$$\inf_B \{V_T^{\mathrm{unk}}(r, B) + \lambda B\} = \Theta_K\left(\sqrt{T} + \min\{Tr, \lambda^{1/3}T^{2/3}\}\right). \tag{B.13}$$

Choose zero audits or $B \asymp (T/\lambda)^{2/3}$, with rounding and the cap T. For fixed positive radius, the interior choice improves the calibration term when λ is below order Tr^3. The argument below proves the endpoints. This is a utility-denominated consequence of the rate, not a measured annotation-price or latency threshold.

We justify (B.13) as a consequence of the uniform constants in Theorem 3.2. Consider

$$F(B) = T \min\{r, (B+1)^{-1/2}\} + \lambda B.$$

The square-root baseline is separate. The choice $B = 0$ gives $F(0) = Tr$ because $r \le 1/16$. If $\lambda \le T^{-1/2}$, choosing $B = T$ costs at most $2\sqrt{T}$ in addition to constant multiples of the baseline. If $T^{-1/2} < \lambda < T$, choose $B = \lceil (T/\lambda)^{2/3} \rceil \wedge T$. Then $T/\sqrt{B+1} + \lambda B \le C\lambda^{1/3}T^{2/3}$ for a numerical constant C. If $\lambda \ge T$, zero audits suffice for the upper bound in (B.13).

For the matching lower bound, any budget with $B + 1 \le r^{-2}$ pays at least Tr in the calibration term. Otherwise $B \ge 1$, and $B + 1 \le 2B$ gives

$$F(B) \ge \frac{T}{\sqrt{2B}} + \lambda B \ge c\lambda^{1/3}T^{2/3}.$$

The last inequality follows by minimizing over positive real B; it also holds with zero right-hand side when $\lambda = 0$. Minimize over integer budgets and combine this with the ordinary square-root lower bound. When $r = 0$ the calibration term vanishes, and $B = 0$ is already optimal up to its square-root constants. The parameter λ measures cost in the numerical-utility units used to define $F(B)$.

C ALGORITHMS AND REGRET GUARANTEES

C.1 FROM COMPARISON STATISTICS TO NUMERICAL UPDATES

We first separate the part estimated from comparisons from the residual requiring an audit. For a known comparison experiment, choose a bounded statistic D with

$$\mathbb{E}[D \mid x] = h(x), \qquad |D| \le A, \qquad h \in \mathcal{L}, \quad \|x - h(x)\|_\infty \le e. \tag{C.1}$$

An exact audit computes $r = x - h(x)$, whose range is at most e. The useful decomposition is at the level of *conditional means*; the individual comparison statistic need not approximate x on each observation. Define two nonnegative channels

$$X^{\mathrm{c}} = (A - D)/2 \in [0, A], \qquad X^{\mathrm{a}} = (e - r)/2 \in [0, e]. \tag{C.2}$$

Their conditional means sum to $(A+e)/2 - x/2$. Only the second channel needs a numerical audit.

An exact audit budget. With k_t admissions left and $n_t = T - t + 1$ rounds left, set $\rho_t = k_t/n_t$. Draw the audit indicator $C_t \sim \mathrm{Ber}(\rho_t)$ privately, independently of current observations. This uses exactly B admissions on every path. It also avoids an additional exploration budget. After the last admission, continue updating the comparison channel alone.

The update. Start with equal action weights. Sample the two proposals from normalized weights q_t. Use the estimate

$$\widehat{L}_t(i) = \mathbf{1}\{I_t = i\} \left[\frac{X_t^{\mathrm{c}}}{q_t(i) + \eta A} + \frac{C_t X_t^{\mathrm{a}}}{\rho_t q_t(i) + \eta e} \right], \qquad w_{t+1}(i) = w_t(i) e^{-\eta \widehat{L}_t(i)}. \tag{C.3}$$

Omit zero-range channels and the second term after the budget is exhausted. This is a two-channel implicit-exploration update (Neu, 2015); its full specification is Algorithm 2 in Appendix C.

Theorem C.1 (Residual audit bound)**.** *Let $B \ge 1$ and suppose* (C.1) *holds for the actual comparison experiment. Put*

$$W = TA^2 + \frac{2e^2 T(T+1)}{B+1}, \qquad \eta = \sqrt{\frac{\log K}{KW}}.$$

The update (C.3) *uses exactly B admissions and, for every predictable utility sequence, satisfies*

$$\mathbb{E}R_T \le 16\sqrt{KW \log K} + 4e\frac{T-B}{B+1}. \tag{C.4}$$

When $W = 0$, the gap is identically zero and no update is needed.

The first term separates as $O(\sqrt{K \log K}[A\sqrt{T} + eT/\sqrt{B+1}])$. Taking a best observable approximation sets $e = \varepsilon_{\mathcal{L}}$; A records its estimation cost. This is the known-experiment residual guarantee. Known-calibration learning uses D_θ with zero residual. The nominal route uses the one-channel proof with its bias bounded separately in Appendix B. The audit-only route sets the comparison statistic to zero.

Where the bound comes from. A scaled implicit-exploration inequality controls each channel's mixture bias, second moment, and exponential comparator error. Conditional Cauchy–Schwarz combines the channels even though they share the same proposal. The uniform-subset audit gate satisfies

$$\mathbb{E} \sum_{t:k_t>0} \rho_t^{-1} \le \frac{2T(T+1)}{B+1}, \qquad \mathbb{E}(T - \tau) = \frac{T-B}{B+1},$$

where τ is the last audit. These two quantities give W and the remaining-tail term, respectively. A log-sum-exp argument keeps the random hindsight comparator inside the expectation.

Audit-only learning with comparison advice. The choice $h = 0$, $A = 0$, $e = 1$ yields an audit-only learner with $O(T\sqrt{K \log K/(B+1)})$ regret. It can use a bounded unbiased numerical gap, or the two individual scores, instead of exact scores. Predictable comparison-based advice can be incorporated by the standard expert-advice mixture (Auer et al., 2002; Beygelzimer et al., 2011); the guarantee remains against every fixed action. The replay uses this specialization with predictable comparison advice. Its numerical measurements satisfy the conditional-mean requirement in Appendix A; the comparison process needs neither a fitted link nor repeated queries.

Algorithm 2 Split-channel implicit exploration

Require: Horizon T, admission budget $B \geq 1$, statistic D, mean function h, ranges A, e.
$k \leftarrow B$, $w(i) \leftarrow 1$ for $i \in [K]$.
$W \leftarrow TA^2 + 2e^2T(T+1)/(B+1)$, $\eta \leftarrow \sqrt{\log K/(KW)}$.
for $t = 1, \ldots, T$ **do**
  $q(i) \leftarrow w(i)/\sum_j w(j)$; draw I, J independently from q.
  Execute the allowed comparison session and obtain D; put $X^{\mathrm{c}} \leftarrow (A - D)/2$.
  $\rho \leftarrow k/(T - t + 1)$; draw $C \sim \mathrm{Ber}(\rho)$ independently of the current session.
  if $C = 1$ **then**
    Audit the selected scores; $x \leftarrow [u_t(I) - u_t(J)]/2$.
    $r \leftarrow x - h(x)$; $X^{\mathrm{a}} \leftarrow (e - r)/2$; $k \leftarrow k - 1$.
  end if
  Form $\widehat{L}$ by (C.3), omitting zero-range or exhausted channels.
  $w(i) \leftarrow w(i) \exp(-\eta \widehat{L}(i))$ for all i.
end for

C.2 Implicit exploration with a measurement range

Lemma C.2. *Condition on an entering history. Let $I \sim q$, and let $C \sim \mathrm{Ber}(\rho)$ be independent of the current potential observation, with $\rho > 0$. On admission let $X \in [0, L]$, and write $x_i = \mathbb{E}[X \mid I = i, C = 1]$. For $\lambda > 0$, define*

$$\widehat{x}_i = \frac{C\mathbf{1}\{I = i\}X}{\rho q_i + \lambda L}.$$

For every predictable $a \in \Delta_K$,

$$\mathbb{E}e^{2\lambda(\langle a, \widehat{x}\rangle - \langle a, x\rangle)} \leq 1. \tag{C.5}$$

Also, $\mathbb{E}\langle q, \widehat{x}^2\rangle \leq KL^2/\rho$ and $\langle q, x\rangle - \mathbb{E}\langle q, \widehat{x}\rangle \leq K\lambda L^2/\rho$.

Proof. Put $p_i = \rho q_i$. On the event $(C, I) = (1, i)$ let $z = 2\lambda a_i X/(p_i + \lambda L) < 2$. The elementary inequality $e^z \leq 1 + z/(1 - z/2)$ for $0 \leq z < 2$ gives

$$e^{2\lambda\langle a, \widehat{x}\rangle} \leq 1 + \frac{2\lambda a_i X}{p_i + \lambda L - \lambda a_i X} \leq 1 + \frac{2\lambda a_i X}{p_i}.$$

Summing over the disjoint (C, I) events yields expectation at most $1 + 2\lambda\langle a, x\rangle \leq e^{2\lambda\langle a, x\rangle}$. Positive weights keep $q_i > 0$; zero-probability events could equivalently be omitted. Direct calculation gives

$$\mathbb{E}\langle q, \widehat{x}^2\rangle = \sum_i \frac{\rho q_i^2 \mathbb{E}[X^2 \mid I = i, C = 1]}{(\rho q_i + \lambda L)^2} \leq KL^2/\rho,$$

and bias $\sum_i q_i \lambda L x_i/(\rho q_i + \lambda L) \leq K\lambda L^2/\rho$. The exponential inequality is the range-rescaled implicit-exploration argument of Neu (2015). □

C.3 The exact-budget gate

Lemma C.3. *Initialize $k_1 = B \in \{1, \ldots, T\}$. At round t, put $\rho_t = k_t/(T - t + 1)$, draw $C_t \sim \mathrm{Ber}(\rho_t)$ privately and independently of current feedback, and set $k_{t+1} = k_t - C_t$. Exactly B admissions occur. If τ is their last time, then*

$$\mathbb{E}(T - \tau) = \frac{T - B}{B + 1}, \qquad \mathbb{E}\sum_{t:k_t > 0} \rho_t^{-1} \leq \frac{2T(T+1)}{B+1}. \tag{C.6}$$

Proof. Sequentially choosing k of n remaining positions with probability k/n produces a uniform B-subset of $[T]$. One can verify this by assigning probability $\binom{T}{B}^{-1}$ to every possible binary sequence of B admissions. The $B + 1$ spacings have equal expectations, giving the first identity.

When n positions remain, $X = k_t \sim \mathrm{Hypergeom}(T, B, n)$. Since $\mathbf{1}\{X > 0\}/X \le 2/(X+1)$ and, by Vandermonde's identity,

$$\mathbb{E}\frac{1}{X+1} = \frac{T+1}{(B+1)(n+1)}\left[1 - \frac{\binom{T-B}{n+1}}{\binom{T+1}{n+1}}\right],$$

we obtain $\mathbb{E}[\mathbf{1}\{X > 0\}n/X] \le 2(T+1)/(B+1)$. Sum over $n = 1, \dots, T$. Out-of-range binomial coefficients are zero. The marginal schedule is uniform even when the environment reacts to its past: admission decisions depend only on remaining counts and independent coins. The proof never conditions on the future schedule when analysing utility or estimation errors. □

C.4 Channel combination and the hindsight comparator

We prove Theorem C.1 and its predictable-advice version. Let $a_t^h \in \Delta_K$ be predictable advice rules, including the K fixed actions. The master has weights $\pi_t(h)$ and chooses $q_t = \sum_h \pi_t(h) a_t^h$. For a pure-action learner, these are simply the K coordinate vectors. Let H upper bound both $\log K$ and the negative log prior mass of each fixed action. Equal action priors give $H = \log K$; allocating mass $\beta < 1$ to an extra advice rule gives $H = \log[K/(1-\beta)]$.

Write $m_t^{\mathrm{c}}(i) = \mathbb{E}[X_t^{\mathrm{c}} \mid \mathcal{G}_t, I_t = i]$ and, on active rounds $k_t > 0$, $m_t^{\mathrm{a}}(i) = \mathbb{E}[X_t^{\mathrm{a}} \mid \mathcal{G}_t, I_t = i]$. In the latter expression the conditional distribution of J_t is q_t, since the gate is independent. Set $m_t^{\mathrm{a}} = 0$ after the budget is exhausted. Let $m_t = m_t^{\mathrm{c}} + m_t^{\mathrm{a}}$.

Apply Lemma C.2 to each channel with $\lambda = \eta$, $L = A$ or e, and observation probability 1 or ρ_t. Put

$$V_t = A^2 + e^2 \frac{\mathbf{1}\{k_t > 0\}}{\rho_t},$$

where the second term is defined as zero when inactive. Mixture bias is at most $\eta K V_t$. Since $(a+b)^2 \le 2a^2 + 2b^2$, the conditional second moment of their sum is at most $2KV_t$. Conditional Cauchy–Schwarz combines the two exponential bounds at exponent η:

$$\mathbb{E}_t \exp\{\eta\langle a_t^h, \widehat{L}_t - m_t\rangle\} \le 1.$$

This does not assume independence of the two channels or of two scores. The active indicator is predictable, so each fixed action gives a nonnegative supermartingale even after admission stops. Iterating and applying log-sum-exp gives

$$\mathbb{E}\max_{i\in[K]} \sum_t (\widehat{L}_t(i) - m_t(i)) \le \frac{\log K}{\eta}. \tag{C.7}$$

The exponential-weights potential inequality is pathwise and valid for arbitrary nonnegative estimates:

$$\sum_t \langle q_t, \widehat{L}_t\rangle - \min_i \sum_t \widehat{L}_t(i) \le \frac{H}{\eta} + \frac{\eta}{2}\sum_t \sum_h \pi_t(h)\langle a_t^h, \widehat{L}_t\rangle^2.$$

It follows from $e^{-z} \le 1 - z + z^2/2$ for $z \ge 0$ and comparison with the prior of the hindsight-best fixed action. Jensen bounds the squared expert loss by $\langle q_t, \widehat{L}_t^2\rangle$. Together with (C.7), the bias bound, and Lemma C.3, expected target-loss regret is at most

$$\frac{2H}{\eta} + 2\eta K \mathbb{E}\sum_t V_t \le \frac{2H}{\eta} + 2\eta K W.$$

With $\eta = \sqrt{H/(KW)}$, this is $4\sqrt{KWH}$.

To translate to utility, let $\mu_t = \langle q_t, u_t\rangle$. On active rounds,

$$m_t(i) = \frac{A+e}{2} - \frac{u_t(i) - \mu_t}{4}.$$

Hence numerical regret is four times target-loss regret there. On inactive rounds, the remaining comparison target differs from $A/2 - [u_t(i) - \mu_t]/4$ by $\frac{1}{2}\mathbb{E}[r_t \mid \mathcal{G}_t, I_t = i]$, whose absolute value

is at most $e/2$. The mixture-comparator difference costs at most e in loss units, or $4e$ in numerical units, per inactive round. The first identity in (C.6) proves

$$\mathbb{E}R_T \le 16\sqrt{KWH} + 4e(T-B)/(B+1).$$

This proves Theorem C.1 and its advice form, using the expected realized-regret identity from Appendix A. Regret against a predictable advice rule would be evaluated on the actual run's history; no counterfactual rerun of an adviser is claimed.

C.5 Audit-only scores and practical advice

The real-data implementation uses the two observed scores directly, instead of discarding their common offset. With scores $Y_t \in [0,1]$ satisfying (A.1), define $L_t = 1 - Y_t$ and

$$\widehat{\ell}_t(i) = \frac{C_t}{2}\left[\frac{\mathbf{1}\{I_t = i\}L_t(I_t)}{\rho_t q_t(i) + \gamma} + \frac{\mathbf{1}\{J_t = i\}L_t(J_t)}{\rho_t q_t(i) + \gamma}\right]. \tag{C.8}$$

Each summand before the factor $1/2$ satisfies Lemma C.2 with $L = 1$. Convexity of the exponential bounds the average by the average of its two exponential bounds, with no assumption of independent numerical errors. Jensen similarly controls the second moment. Using $\gamma = \eta/2$ and

$$\eta = \sqrt{\frac{(B+1)H}{KT(T+1)}}$$

in the master update gives expected score regret at most

$$4\sqrt{\frac{KT(T+1)H}{B+1}} + \frac{T-B}{B+1}. \tag{C.9}$$

Indeed, the comparator and potential terms give $2H/\eta$, while variance and mixture bias together give $\eta K\mathbb{E}\sum_{k_t>0}\rho_t^{-1}$. After the last audit, at most one score-regret unit is lost per round. For scores in $[-1,1]$, map to $[0,1]$ and multiply the score-regret bound by two.

Audit+comparison uses K fixed actions and one online comparison rule. Its prior on the latter is $\beta = 0.9$, so $H = \log(10K)$. At the start of round t, let a_t be the comparison rule's current distribution. The master samples q_t as the mixture of the fixed-action weights and the weight on a_t. After obtaining a comparison $z_t \in \{-1, 0, 1\}$ of the two selected outputs, the auxiliary rule updates its two sampled action log-weights by

$$v_{t+1}(i) = v_t(i) - \frac{\eta_c}{2}\left[\frac{\mathbf{1}\{I_t = i\}(1 - z_t)/2}{q_t(i) + \gamma_c} + \frac{\mathbf{1}\{J_t = i\}(1 + z_t)/2}{q_t(i) + \gamma_c}\right],$$

where a_t is the softmax of v_t, $\eta_c = \sqrt{\log K/(KT)}$, and $\gamma_c = \eta_c/2$. Denominators use the actual master proposal probabilities q_t, not the auxiliary rule's hypothetical probabilities. On admission the master's fixed-action losses are (C.8) and the advice loss is $\langle a_t, \widehat{\ell}_t\rangle$, using the *pre-feedback* a_t. Thus the online advice is predictable and (C.9) applies, whatever the comparison law. The auxiliary rule is not a separate theorem or a newly claimed preference-learning primitive; it is a concrete source of advice for the standard master.

When $B = 0$ the audit-only bound is replaced by the trivial score regret T (or numerical regret $2T$ under $[-1,1]$ scores). No unbiased numeric target can be inferred solely from an arbitrary comparison stream. A noisily measured x cannot generally be substituted into $x - h(x)$ for nonlinear h; only the zero-component score specialization avoids that extra calibration and exact-residual requirement.

D The known-experiment benchmark

Theorem 3.2 concerns unknown calibration with zero approximation residual in every known model. Here the comparison experiment is known, but its likelihood span leaves a positive numerical residual. Uniform-approximation duality identifies that residual, and the selected-audit experiment determines its observation cost.

Let $\mathcal{C}$ be a finite collection of comparison decision trees. All prefix likelihoods $\ell_z(x)$ are continuous, constants belong to their span, and the collection permits reversal of the pair. Labels within a session may be dependent. Define $\mathcal{L}$ and $\varepsilon_{\mathcal{L}}$ by (3.3); write $V_T(\mathcal{C}, B)$ for its numerical-regret minimax value when the law is supplied. A bounded statistic has expectation in $\mathcal{L}$. Conversely, for $h = \sum_j a_j \ell_j$, select event j in proportion to $|a_j|$ and return $\operatorname{sign}(a_j) \sum_r |a_r|$ if that event occurs, and zero otherwise. This realizes h with a bounded statistic, though its range may be large.

The next statement varies a population-mixture index α within one fixed comparison experiment. In contrast, Theorem 3.2 varies the calibration parameter θ. The experiment equivalence below preserves the full observed history, including audit arrival times.

Theorem D.1 (Every selected audit is one coin)**.** *Suppose $\varepsilon_{\mathcal{L}} > 0$. There are mutually singular distributions μ_+, μ_- on half-gaps, with means $\varepsilon_{\mathcal{L}}, -\varepsilon_{\mathcal{L}}$, such that every allowed comparison transcript has the same distribution under both. On the oblivious family*

$$x_t \overset{\text{iid}}{\sim} \mu_\alpha = \tfrac{1+\alpha}{2}\mu_+ + \tfrac{1-\alpha}{2}\mu_-, \qquad u_t = (x_t, -x_t), \quad -1 \le \alpha \le 1,$$

the complete observed history of any adaptive learner using at most B audits is a randomized function of B independent $\operatorname{Ber}((1+\alpha)/2)$ observations and randomness independent of α. If exactly B rounds are audited, the B component labels can also be recovered from the history. These two experiments are statistically equivalent.

For $K > 2$, duplicate the second action. The theorem grants exact scores even when both proposals have the same type. Such an audit still reveals x_t on the constructed family. A weaker audit cannot reveal more information.

Corollary D.2 (An explicit numerical price)**.** *For every $0 \le B \le T$,*

$$V_T(\mathcal{C}, B) \ge \max\left\{\sqrt{T/2},\ T\varepsilon_{\mathcal{L}} c_B\right\}, \qquad c_B = \sup_{0\le\alpha\le 1} \alpha\big[1 - \operatorname{TV}(P_\alpha^{\otimes B}, P_{-\alpha}^{\otimes B})\big], \tag{D.1}$$

where $P_\alpha = \operatorname{Ber}((1+\alpha)/2)$. In particular, $c_0 = 1$ and $c_B \ge 1/(8\sqrt{B+1})$. For $B \ge 1$, $c_B = O(B^{-1/2})$.

The testing constant has a direct binomial-tail formula:

$$1 - \operatorname{TV}(P_\alpha^{\otimes B}, P_{-\alpha}^{\otimes B}) = 2\mathbb{P}\{N < B/2\} + \mathbb{P}\{N = B/2\}, \quad N \sim \operatorname{Bin}\big(B, (1+\alpha)/2\big). \tag{D.2}$$

The lower bound holds even for an oblivious environment with known comparison calibration.

D.1 Proof of the known-experiment benchmark

D.1.1 Observable statistics and uniform-approximation duality

For a finite deterministic tree, each transcript statistic is a finite linear combination of its leaf likelihoods. A prefix likelihood is the sum of descendant leaf likelihoods, including an early-stop leaf when appropriate. Integrating over private randomization preserves the span. Conversely, suppose $h = \sum_{j=1}^{M} a_j \ell_j$, with ℓ_j a realizable leaf or prefix probability. Put $A = \sum_j |a_j|$. For $A > 0$, independently select j with probability $|a_j|/A$, execute its allowed tree (stopping at its prefix if desired), and output

$$D = A \operatorname{sign}(a_j) \mathbf{1}\{\text{event } j \text{ occurs}\}.$$

Then $|D| \le A$ and $\mathbb{E}[D \mid x] = h(x)$. When all coefficients vanish use $D = 0$. This construction uses one allowed session, not the union of its component trees. Its internal random choice is part of the observed transcript. Since reflection preserves $\mathcal{L}$, a best approximation can be chosen odd. Known identical-action gaps can be set to zero directly.

Lemma D.3 (Dual pair)**.** *Let $\mathcal{L} \subset C([-1,1])$ be finite dimensional, contain constants, and be preserved by reflection. If $e = \inf_{h\in\mathcal{L}} \|x - h(x)\|_\infty > 0$, there are mutually singular probability measures μ_+, μ_- such that $\int h\, d\mu_+ = \int h\, d\mu_-$ for all $h \in \mathcal{L}$, and $\int x\, d\mu_\pm = \pm e$.*

Proof. On $\mathcal{L} + \operatorname{span}\{x\}$, the functional $h + ax \mapsto ae$ has norm one. Indeed, $\|h + ax\|_\infty \ge |a|e$, and a best approximation attains this bound when $a = 1$. Hahn–Banach extends the functional to

$C([-1,1])$ with the same norm. Riesz representation gives a signed measure λ with $\|\lambda\|_{\mathrm{TV}} = 1$, $\int h\, d\lambda = 0$ for $h \in \mathcal{L}$, and $\int x\, d\lambda = e$.

Let $S(x) = -x$ and replace λ by $\lambda_o = (\lambda - S_{\#}\lambda)/2$. It still annihilates $\mathcal{L}$, still pairs with x to e, and has total variation at most one. Its variation is exactly one: for a best approximation h_*,

$$e = \int (x - h_*)\, d\lambda_o \le e\|\lambda_o\|_{\mathrm{TV}}.$$

Since λ_o annihilates constants, its Jordan parts each have mass $1/2$. Oddness implies $S_{\#}\lambda_o^+ = \lambda_o^-$. Let $\mu_+ = 2\lambda_o^+$ and $\mu_- = 2\lambda_o^-$. Their h-integrals agree; their means are negatives of one another and have difference $2e$. The Jordan parts have disjoint measurable supports, proving mutual singularity. On a finite symmetric domain, the same argument uses finite-dimensional linear-program duality. □

D.1.2 An exact simulator for selectively audited histories

Apply Lemma D.3 to the span of all permitted comparison probabilities. Let $S_t \in \{+,-\}$ be the component generating x_t, with $\mathbb{P}_\alpha(S_t = +) = (1+\alpha)/2$, independently across rounds. Define μ_α as in Theorem D.1. For any fixed entering history and any selected deterministic comparison tree, every prefix z has likelihood

$$M_z := \int \ell_z(x)\, d\mu_+(x) = \int \ell_z(x)\, d\mu_-(x).$$

Consequently, its law is M_z under every α. If $M_z > 0$, its conditional component distributions are

$$\mu_{s,z}(dx) = \frac{\ell_z(x)\mu_s(dx)}{M_z}, \qquad s \in \{+,-\}, \tag{D.3}$$

and Bayes' formula gives $\mathbb{P}_\alpha(S_t = + \mid z, \mathcal{H}_{t-1}) = (1+\alpha)/2$. Zero-likelihood prefixes are never visited.

A learner may randomize whether to audit after z. Conditional on its private seed and entering history, this decision is a function of z alone. It does not alter the component weights. The following simulation is therefore exact, by induction over rounds and session events.

Before the first audit in a round, generate the transcript using the common prefix probabilities M_z. Follow the learner's choices on this transcript. If it never audits, complete the session from the common law; no α-dependent observation is required. If it audits after z, consume the next independent $\mathrm{Ber}((1+\alpha)/2)$ bit S and sample x from (D.3). Reveal the scores of the selected action types, and generate any later labels conditional on this x and the prefix. Every kernel except the bit distribution is independent of α.

For two actions, set the scores to $(x, -x)$. With $K > 2$, let action 1 have score x and all others have score $-x$. A cross-type proposal sees half-gap x or $-x$; reflection closure covers both. A same-type proposal has half-gap zero, so its comparison law is independent of x and satisfies the same matching statement with a constant likelihood. The learner's choice of action indices adds no unaccounted information.

At most B rounds consume a bit. Extra unused bits can be appended without changing the history, so the complete observed experiment is a randomized transformation of B bits. Conversely, every exact audit score reveals the underlying x: its action type and score determine x even when the two proposals have the same type. Since μ_+ and μ_- are mutually singular, x determines S almost surely. When exactly B rounds are audited, these B recovered labels are independent with the stated common parameter. This is the reverse transformation and proves Theorem D.1.

The simulator concerns the *observed history*. Unobserved scores in unaudited rounds need not be simulated from a α-independent kernel, and are not claimed to be. The regret calculation below uses their known means separately. Neither statistical equivalence nor the simulation removes the time at which a score becomes available.

D.1.3 Regret and the finite-budget testing constant

Let $\mathbb{P}_\alpha^{t-1}$ be the law of the learner's observed past before its current pair is drawn. Data processing under the simulator gives

$$\mathrm{TV}(\mathbb{P}_\alpha^{t-1}, \mathbb{P}_{-\alpha}^{t-1}) \le \mathrm{TV}(P_\alpha^{\otimes B}, P_{-\alpha}^{\otimes B}).$$

Write $s_t \in [-1, 1]$ for the average signed type of the two proposed actions. Current x_t is independent of the entering history and the fresh action draws, and has mean αe. The expected payoff is $\alpha e \mathbb{E}_\alpha s_t$. The hindsight comparator is at least the score of the action with the larger *mean* utility. This use of $\mathbb{E}\max \ge \max \mathbb{E}$ is a lower bound, not an interchange in an upper bound. Hence for $\alpha \ge 0$,

$$\begin{aligned}\tfrac{1}{2}(\mathbb{E}_\alpha R_T + \mathbb{E}_{-\alpha} R_T) &\ge \alpha e T - \frac{\alpha e}{2} \sum_t (\mathbb{E}_\alpha s_t - \mathbb{E}_{-\alpha} s_t) \\ &\ge \alpha e T [1 - \mathrm{TV}(P_\alpha^{\otimes B}, P_{-\alpha}^{\otimes B})].\end{aligned}$$

Optimize over α. One of the two worlds attains at least the average. The randomized oblivious sequence can be fixed to an environment seed attaining at least the averaged regret for each learner.

For the separate baseline term, take independent Rademacher X_t and scores $(X_t, -X_t)$, duplicating the second type if necessary. The learner's expected payoff is zero even if every previous numerical score is observed. The expected hindsight comparator is $\mathbb{E}|\sum_t X_t| \ge \sqrt{T/2}$, by the Rademacher Khintchine inequality. Since these are different instances, the lower bound is the maximum of the terms.

For the B-coin experiment, its likelihood ratio is monotone in the number of positive bits. The optimal test between α and $-\alpha$ thresholds their count at $B/2$, splitting ties equally. The sum of testing errors is exactly (D.2). When $B = 0$, the two empty experiments coincide and $c_0 = 1$.

For an elementary finite-B bound, take $\alpha = 1/(4\sqrt{B+1})$. For $0 \le \alpha \le 1/2$,

$$\mathrm{KL}(P_\alpha \| P_{-\alpha}) = \alpha \log \frac{1+\alpha}{1-\alpha} \le 8\alpha^2.$$

Pinsker and product additivity give $\mathrm{TV}(P_\alpha^{\otimes B}, P_{-\alpha}^{\otimes B}) \le 2\alpha\sqrt{B} \le 1/2$, implying $c_B \ge 1/(8\sqrt{B+1})$. Conversely, Hoeffding's inequality in (D.2) gives $1 - \mathrm{TV} \le 2e^{-B\alpha^2/2}$, so $c_B \le 2e^{-1/2}/\sqrt{B}$ for $B \ge 1$. These bounds suffice for the order statement; our plots evaluate the exact binomial expression rather than either inequality.

E Calibration and critic-overlap experiments

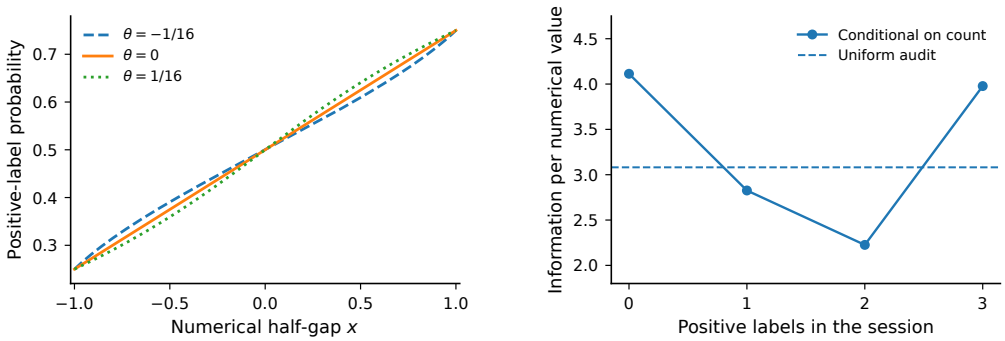
E.1 Analytic identities and selected-audit tests

The tests use the cubic family (B.1) and four-point population (B.4). We evaluate all four three-label counts over 401 equally spaced parameters in $[-1/16, 1/16]$. A separate check enumerates every ordered prefix at 61 positive and negative parameter separations. The largest discrepancy between triple probabilities and their centre-link values is 1.12×10^{-16}; the largest discrepancy in the conditional unbiased-statistic identity is 3.34×10^{-16}. The check independently computes the range of D_θ and the inverse derivative. These computations supplement the interval-wide proof with checks of the explicit formulas.

Figure 4(a) displays the three links at $\theta \in \{-1/16, 0, 1/16\}$. To evaluate how a verdict changes audit quality, let N be the number of positive labels and compute

$$I_n = \sum_j Q_0(j \mid N = n) \left[\partial_\theta \log Q_\theta(j \mid N = n)|_{\theta=0}\right]^2.$$

The numerical derivative uses a centred difference of size 10^{-6} and is checked at half that step. Figure 4(b) gives the resulting four values. Averaging them under the common count probabilities gives information 3.08148 per uniform audit; $N = 0$ gives the largest value, 4.11312. Its common probability is $379/2048$, so retaining this category requires a mean $2048/379 \simeq 5.40$ cheap sessions, or 16.21 labels, per numerical observation. The screening pool isolates information per admitted score; its session cost is additional to the numerical budget.



(a) Uniformly regular comparison links.

(b) Conditional numerical information.

Figure 4: The calibration witness and the value of verdict-based selection. The label-count marginal is parameter-independent, while its conditional score distribution changes with the parameter. The right panel evaluates local Fisher information at the central calibration.

The testing experiment has budgets $16, 64, 256, 1024, 4096$. Its fixed separation is $\delta = 1/16$; its local separation is $\delta = 1/(4\sqrt{B+1})$. For each design and each of the two equally weighted worlds, 4,096 independent multinomial replications supply the observations to the exact likelihood-ratio test. The uniform design samples the joint (N, X) law. The selected design samples $X \mid N = 0$ after being granted enough cheap sessions. It always pays for exactly B numerical values. The world is held fixed within a replication, and the parameter does not change between numerical observations.

This gives 163,840 independent testing replications across all designs and worlds. Balanced error is the average of their two empirical error frequencies. Each world receives a two-sided 97.5% Clopper–Pearson interval; averaging the endpoints gives simultaneous coverage at least 95% for the balanced risk by a union bound. At zero observed errors, the upper endpoint remains positive. Figure 2(a) and (b) display these intervals. The testing design was chosen from the analytical construction, not from the replay results.

E.2 Direct-critic controls and removal ablations

The three fixed critics are GPT-4o, Gemini-1.5-Pro, and Llama-3.1-Nemotron-70B. We use the same nine selection rules with complete training coverage for *both* partitions. The fixed library separates the partition comparison from the 34-rule validation experiment in Appendix F.3. For each critic, the removal condition excludes only that critic's selection rule and keeps all other records. The comparator is recomputed within the resulting eight-rule library. The verdict used to judge the selected outputs is still the excluded critic's observed record; only its role as an action is removed.

The direct-critic policy always chooses that critic's rule. It uses no numerical labels or learned weights. It is run for all three critics rather than selecting one by numerical test outcomes. Returned inconsistent or erroneous judgments use the same precommitted per-rule, per-round fallback coin as the online replay. Table 3 gives both numerical targets and both partitions; length-ordered runs are also included in the result files.

The methods use the updates in Appendix C.5. Audit-IX is (C.8); Audit+comparison adds the predictable comparison rule with prior weight 0.9. Audited-mean keeps empirical admitted scores and a 5% uniform exploration component; comparison-only is the original auxiliary update used directly, with no audits. The learning rates, fallbacks, initial means, prior weights, and critic choices are fixed across the full-library and removal conditions.

The budgets are 1%, 5%, 20%, and 100%, rounded to the nearest integer and at least one. Every audited method spends that many admissions on every path. The study crosses two partitions, two targets, two orders (shuffled and response-length sorted), three critics, and both overlap conditions. There are 64 seeds per configuration. Comparison-only is run once per combination, not once per nominal budget. The result is 39,936 sequential trajectories and 1,536 direct-policy trajectories.

Table 3: Direct-critic controls, shuffled input, 64 seeds. Every policy uses one existing judge record per prompt and no audit. The three policies are specified by the critics already used by the learning algorithms.

Target	Critic	Training regret	Validation regret
Helpfulness	GPT-4o	0.92 ± 0.53	2.24 ± 0.44
Helpfulness	Gemini Pro	12.81 ± 2.47	3.52 ± 0.56
Helpfulness	Nemotron	17.31 ± 2.59	0.88 ± 0.34
Correctness	GPT-4o	1.64 ± 0.82	1.84 ± 0.41
Correctness	Gemini Pro	10.09 ± 2.17	3.29 ± 0.55
Correctness	Nemotron	16.24 ± 2.52	1.09 ± 0.37

All trajectories replay existing human and model records; the replications introduce algorithmic randomization rather than new inference or annotation.

Within a fixed action library, the action/fallback/gate tape uses seed $30000 + s$, numerical score sampling uses $20000 + s$, and the shuffle uses $5000 + s$. Algorithms in the same comparison share their random tapes. Each score index is sampled only from the corresponding observed answer's listed scores when admitted. Changing the action library changes the number of fallback draws, so a cross-library difference is not interpreted as a within-path paired causal effect. Paired intervals are reported only for method differences within a fixed library.

Tables 4 and 5 show the entire 20% slice, including correctness and length-ordered inputs. The tables include every configuration in that slice, regardless of the sign of the difference. They complement the helpfulness plots in Figure 3(a) and (b). Complete per-run score and judgment-access counts accompany every budget, including the cases where paid information increases regret.

Table 4: All 20% admission comparisons on the train partition: Audit+comparison minus comparison-only regret. Positive means the audited method is worse. Columns use their own hindsight comparator and fixed-library paired randomization; they are not a paired intervention across action-library sizes.

Target	Order	Critic	In library	Removed
Helpfulness	Shuffled	GPT-4o	6.24 ± 1.43	4.83 ± 1.17
Helpfulness	Shuffled	Gemini Pro	6.68 ± 1.23	4.40 ± 1.07
Helpfulness	Shuffled	Nemotron	6.45 ± 1.28	5.00 ± 1.22
Helpfulness	Length	GPT-4o	6.94 ± 1.11	4.37 ± 0.96
Helpfulness	Length	Gemini Pro	4.60 ± 1.33	3.91 ± 1.05
Helpfulness	Length	Nemotron	6.78 ± 1.15	4.43 ± 1.13
Correctness	Shuffled	GPT-4o	5.42 ± 1.30	5.12 ± 1.23
Correctness	Shuffled	Gemini Pro	6.59 ± 1.15	4.53 ± 1.10
Correctness	Shuffled	Nemotron	5.97 ± 1.20	5.37 ± 1.22
Correctness	Length	GPT-4o	6.01 ± 1.09	3.98 ± 0.97
Correctness	Length	Gemini Pro	4.33 ± 1.22	3.63 ± 0.89
Correctness	Length	Nemotron	6.11 ± 1.11	4.27 ± 1.09

Interpreting a direct critic. For a consistent verdict, the critic's rule selects its preferred response; against another rule it therefore ties or wins under that critic. For an inconsistent/error verdict the comparison update is neutral, regardless of the fallback choice. The rule therefore weakly dominates in the critic's *comparison* outcomes. Its numerical value is assessed separately using the recorded scores. In the training helpfulness records, its expected-fallback numerical means are 0.80245 for GPT-4o, 0.80128 for Gemini Pro, and 0.80069 for Nemotron. In the matched validation library the order changes: Nemotron has 0.79762, versus 0.79492 for GPT-4o. The best-in-hindsight comparator additionally reflects the actual fallback draws in each path. These descriptive record summaries are kept separate from the learning methods' inputs.

Interpreting the removal control. Removing the critic's rule controls its direct participation in the action library; correlations among the remaining judges are retained. At 20% admissions, the

Table 5: All 20% admission comparisons on the validation partition: Audit+comparison minus comparison-only regret. Positive means the audited method is worse. Columns use their own hindsight comparator and fixed-library paired randomization; they are not a paired intervention across action-library sizes.

Target	Order	Critic	In library	Removed
Helpfulness	Shuffled	GPT-4o	0.04 ± 0.19	0.18 ± 0.19
Helpfulness	Shuffled	Gemini Pro	0.11 ± 0.16	0.12 ± 0.20
Helpfulness	Shuffled	Nemotron	0.02 ± 0.19	0.05 ± 0.21
Helpfulness	Length	GPT-4o	0.20 ± 0.18	0.16 ± 0.19
Helpfulness	Length	Gemini Pro	-0.02 ± 0.16	0.05 ± 0.22
Helpfulness	Length	Nemotron	0.03 ± 0.16	0.17 ± 0.22
Correctness	Shuffled	GPT-4o	-0.05 ± 0.17	-0.00 ± 0.19
Correctness	Shuffled	Gemini Pro	0.06 ± 0.16	0.11 ± 0.19
Correctness	Shuffled	Nemotron	0.02 ± 0.19	0.08 ± 0.19
Correctness	Length	GPT-4o	0.24 ± 0.18	0.18 ± 0.17
Correctness	Length	Gemini Pro	-0.03 ± 0.14	0.08 ± 0.21
Correctness	Length	Nemotron	0.07 ± 0.16	0.11 ± 0.21

audited update still has higher regret on the shuffled-training helpfulness runs, while the matched validation intervals include zero. The empirical-mean baseline shows that a simpler numerical update can be faster when enough scores are available. These observations describe finite-budget behavior on the records; the minimax theorem instead concerns the hardest utility sequence within its comparison family.

F Recorded-data construction and experimental protocol

F.1 Source records, matching, and the numerical target

HelpSteer2 supplies numerical ratings of answers on five attributes: helpfulness, correctness, coherence, complexity, and verbosity (Wang et al., 2024). Its companion preference dataset supplies independently collected pairwise annotations, including seven-level signed preference strength (Wang et al., 2025). RLHFlow supplies recorded LLM judgments on response pairs, collected under both presentation orders (Li, 2025b;a). We keep the numerical scores and pairwise preferences as separate measurements.

We match a response using the exact tuple (prompt text, response text), with a SHA-256 hash serving only as its index. A pair requires both response keys and their orientation. All 9,125 LLM-judged pairs match the NVIDIA response records, individual-score lists, and preference records. One byte-identical duplicate response record is deduplicated; conflicting duplicates would fail the join. There is no prompt overlap between the 8,677-pair training partition and the 448-pair validation partition. The training partition has 8,676 unique prompts, while every validation prompt is unique.

The reference score for answer a on prompt t is

$$\bar{y}_t(a) = \frac{1}{4n_{t,a}} \sum_{r=1}^{n_{t,a}} y_{t,a,r},$$

where $y_{t,a,r} \in \{0, 1, 2, 3, 4\}$ are all the individual scores listed in the public disagreement file. This differs from the separately released integer score, which involves the dataset's aggregation procedure. We use the listed-score mean consistently for training diagnostics, payoff evaluation, and audit sampling. A single-score audit draws an index uniformly from this fixed list, separately for each distinct answer. The target is the mean of this finite, declared panel of scores.

Field alignment. Table 6 records the alignment signatures. In both partitions, every LLM row's response texts match the source preference row indicated by the numeric suffix of its ID. The LLM rows are in lexicographic ID order. Meanwhile every value of the supplied `Human` column matches the preference direction at its *file-row position* in the original preference partition. Consequently, that column disagrees with the text-matched preference on 5,861 pairs. Our processed records retain

`human_source_column` and create a separate `human` value from the exact text join. No LLM judgment is changed. This alignment result concerns the six supplied file versions, identified by their recorded hashes.

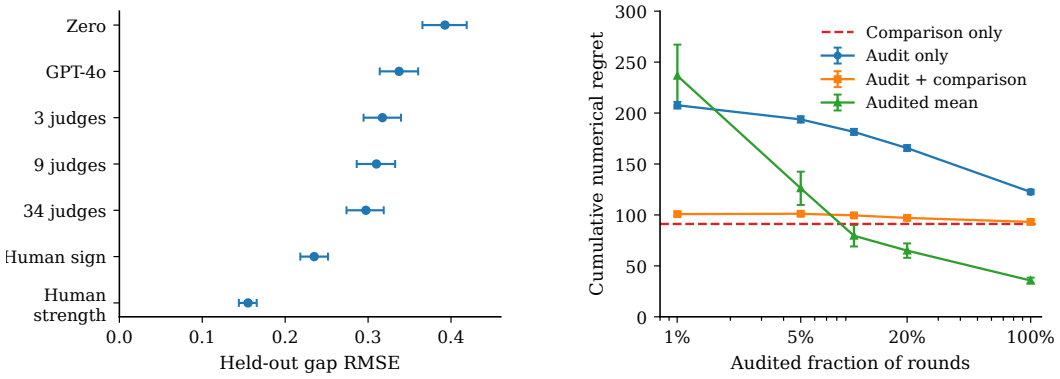
Table 6: Exact matching and coverage of the supplied records. The source files and source `Human` values are preserved. "Attempted" includes returned format/API errors; absent records are different from attempts with errors.

Check	Training	Validation
Response pairs	8,677	448
Text matches preference at numeric ID	8,677	448
Source `Human` matches preference at file-row position	8,677	448
Source `Human` differs from text-joined preference	5,570	291
Pairs with all 34 judges attempted	2,000	448
Judges with complete partition coverage	9	34

Coverage and response categories. The nine judges with complete training coverage are Claude-3.5-Haiku, GPT-4o, Gemini-1.5-Flash, Gemini-1.5-Flash-8b, Gemini-1.5-Pro, Gemma-2-27B, Llama-3.1-Nemotron-70B, Llama-3.3-70B, and Mistral-7B-v0.3. Every judge with complete coverage is included, regardless of its test performance. The all-34 comparison uses only the common 2,000 training pairs and the complete validation partition.

`Prefer A` and `Prefer B` encode consistent choices under the source's position-reversal protocol. `No Preference` indicates inconsistency across the two presentations, not a third response option explicitly chosen by the LLM. Format errors, API errors, and absent records remain distinct categories. For selection policies, a returned inconsistent/error verdict falls back to a uniform choice of the two answers. For a critic's comparison, it produces the neutral update value zero. Absent judgments are never converted to an observed zero: the replay chooses only action sets with complete attempted-record coverage.

Figure 5(a) and (b) show the frozen prediction diagnostic and the independent 128-seed sequential replay, respectively. Their different uncertainty conventions are stated in the caption.



(a) Held-out numerical-gap prediction.

(b) Sequential regret versus paid admissions.

Figure 5: **Information and finite-sample efficiency are separate.** Left: RMSE on 448 validation prompts; intervals are prompt bootstraps with the fitted predictors held fixed. Right: shuffled nine-rule replay, single-score helpfulness audits; the comparison rule uses recorded GPT-4o judgments. The right panel uses 128 seeds; the main text reports the separate matched 64-seed direct-critic control.

F.2 FROZEN NUMERICAL-PREDICTION DIAGNOSTICS

The diagnostic predicts $\bar{y}_t(A) - \bar{y}_t(B)$. Each nonzero model uses the same 2,000 common training pairs, or 4,000 numerical response means (13,462 listed helpfulness scores in total). This calibration expenditure is explicit and is *not* used in the online replay. Categorical inputs are one-hot encoded, keeping returned API/format errors distinct from inconsistent preferences. Ridge regression uses penalty 10 and no intercept. Swapping the two answer positions provides symmetry augmentation by reversing consistent directions and negating the target; errors and inconsistent labels remain unchanged. Predictions are clipped to $[-1, 1]$. No hyperparameter is selected on the 448 validation pairs.

Inputs are the human direction, human strength, a GPT-4o verdict, a three-judge panel (GPT-4o, Gemini-1.5-Pro, Llama-3.1-Nemotron-70B), the nine fully covered judges, or all 34 judges. Table 7 gives the complete held-out errors underlying Figure 5a. Its 5,000 percentile bootstraps resample validation prompts with the fitted predictor fixed (seed 7191). These intervals quantify variation across validation records conditional on the fitted model. They do not include the variability of refitting from another calibration sample.

Table 7: Numerical helpfulness-gap prediction on the 448 validation pairs. All nonzero predictors have identical numerical calibration access. The human inputs are separately collected reference annotations, not cheap LLM substitutes.

Input	RMSE	95% prompt interval	Calibration response means
Zero	0.3927	[0.3655, 0.4189]	0
Human direction	0.2350	[0.2183, 0.2516]	4,000
Human strength	0.1552	[0.1441, 0.1657]	4,000
GPT-4o	0.3374	[0.3140, 0.3603]	4,000
Three judges	0.3173	[0.2945, 0.3396]	4,000
Nine judges	0.3100	[0.2863, 0.3326]	4,000
All 34 judges	0.2973	[0.2740, 0.3188]	4,000

We also select one judge by human-direction agreement and one by numerical gain on the common training pairs. Agreement selects Llama-3.1-405B; numerical gain selects Qwen-2.5-72B. Their validation gains relative to random answer selection are 0.08927 with interval $[0.07430, 0.10359]$ and 0.09017 with interval $[0.07560, 0.10469]$, respectively. The small difference between these two gains does not establish a significant improvement. Table 8 reports all judges.

Figure 6(a) and (b) further show that both positive and negative numerical gaps occur within each recorded verdict category. This measures dispersion of the numerical target within each recorded category. Its interpretation is specific to these categories and records. The theoretical $\varepsilon_{\mathcal{L}}$, by contrast, is a worst-case uniform approximation error for a specified experiment.

F.3 ONLINE REPLAY: OBSERVATIONS AND COST ACCOUNTING

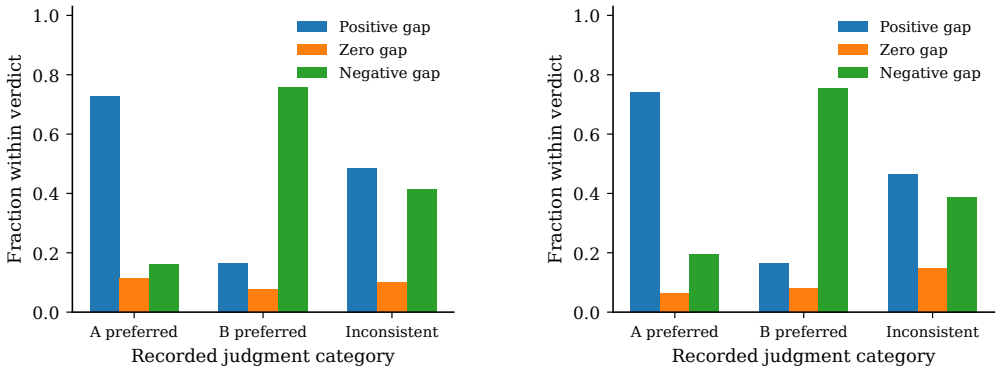
The replay performs no LLM inference, no fresh human annotation, and no synthetic comparison-label generation. An action is a recorded judge's policy for selecting one of the two cached responses. A round fixes each policy's output before the learner draws its two actions. For an inconsistent or erroneous verdict, one fair fallback coin is drawn per policy per round and reused if that policy is selected twice. These precommitted outputs define the current utility vector. The utility of a policy is the reference mean of its selected response. This makes the best fixed policy in hindsight well defined, including fallback randomization.

The learner observes only the requested policy records, the selected output identities, an optional critic record for comparing distinct outputs, and admitted scores. It does not receive the current unselected policies' judgments or the full score table. The full table is used only by the scorer to compute regret. When the outputs coincide, their comparison is a known tie; otherwise the critic's original verdict is reversed if necessary to match the selected orientation.

One audit admission requests at most two distinct answer scores. For a single-score audit, each is a uniformly drawn listed human score; identical outputs reuse one draw. For the panel-mean sensitivity run, each requested mean consumes all listed scores for that answer, and they are counted

Table 8: All 34 judges on the same 448 validation pairs. Agreement uses the text-joined human direction and excludes API/format errors. Fallback includes position-inconsistent verdicts and errors; it is not a human tie rate. Numerical gain is relative to uniformly selecting an answer, using the raw helpfulness-score-list mean divided by four. Rows are alphabetical rather than selected by performance.

Recorded judge	Agreement (%)	Fallback (%)	Error (%)	Numerical gain
Claude-3.5-Haiku	26.8	93.1	0.9	0.0102
Claude-3.5-Sonnet	61.3	25.2	2.0	0.0913
DeepSeek-LLM-67B	33.3	82.4	0.2	0.0303
DeepSeek-V3	61.3	32.6	0.7	0.0867
GPT-4-Turbo	60.0	29.9	0.0	0.0906
GPT-4o	54.8	31.9	0.7	0.0845
GPT-4o-Mini	56.7	38.6	0.0	0.0780
Gemini-1.0-Pro	51.4	29.7	1.3	0.0525
Gemini-1.5-Flash	59.6	27.0	0.0	0.0829
Gemini-1.5-Flash-8b	48.9	40.0	0.0	0.0577
Gemini-1.5-Pro	56.5	36.4	0.0	0.0822
Gemma-2-27B	44.1	58.9	0.2	0.0535
Gemma-2-9B	50.7	45.5	0.4	0.0622
Llama-3-70B	54.9	43.1	0.0	0.0743
Llama-3-8B	42.0	48.7	0.0	0.0385
Llama-3.1-405B	58.3	37.9	0.0	0.0893
Llama-3.1-70B	58.7	30.6	0.0	0.0892
Llama-3.1-8B	37.1	55.4	0.0	0.0350
Llama-3.1-Nemotron-70B	59.8	27.0	0.0	0.0872
Llama-3.2-11B-Vision	43.1	50.0	0.0	0.0417
Llama-3.2-3B	21.7	96.7	0.0	0.0010
Llama-3.2-90B-Vision	59.6	29.5	0.0	0.0909
Llama-3.3-70B	58.9	42.9	0.0	0.0834
Mistral-7B-v0.1	24.0	96.9	8.9	0.0047
Mistral-7B-v0.2	25.9	92.0	0.0	0.0100
Mistral-7B-v0.3	25.0	93.8	0.0	0.0080
Mixtral-8x22B	42.6	64.5	1.6	0.0468
Mixtral-8x7B	33.4	76.1	0.4	0.0322
Qwen-2-72B	58.1	28.3	0.4	0.0822
Qwen-2-VL-72B	56.2	28.8	0.0	0.0798
Qwen-2.5-72B	58.7	21.2	0.0	0.0902
Qwen-2.5-7B	45.1	57.6	0.0	0.0551
Qwen-2.5-Coder-32B	55.3	41.5	7.6	0.0750
SOLAR-10.7B	38.2	60.7	7.1	0.0270



(a) GPT-4o.

(b) Llama-3.1-Nemotron-70B.

Figure 6: Signs of numerical helpfulness gaps within each recorded judgment category on the validation pairs. A positive gap favours answer A. Panels (a) and (b) use the same target and cases, without a fitted comparison link.

accordingly. Accessed judge records count the one or two distinct selector records plus an additional critic record if the outputs differ. Possible cache hits when the critic coincides with a selected rule remain included in the count. The unit is a record access, and each source record summarizes two presentation orders. Dollar costs, token usage, human time, and latency were not measured.

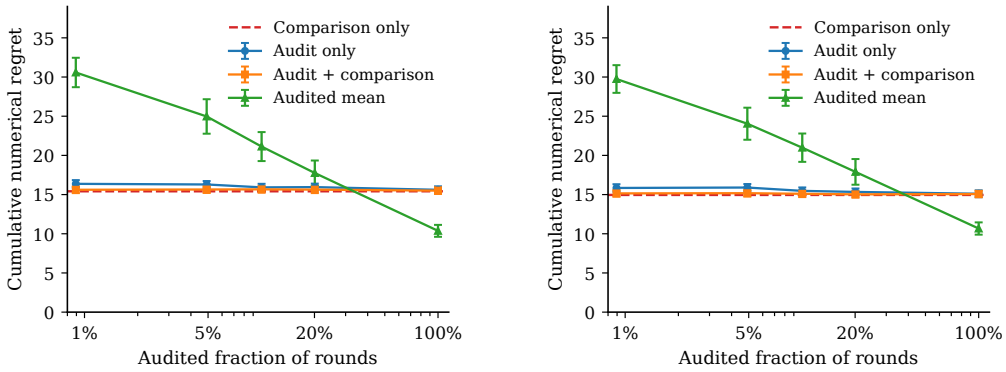
Algorithms. Audit-IX uses (C.8) with equal fixed-action priors and no critic. Audit+comparison uses the predictable-advice construction in Appendix C.5, with $\beta = 0.9$ and one of three separately reported critics: GPT-4o, Gemini-1.5-Pro, or Llama-3.1-Nemotron-70B. The prior was fixed independently of validation outcomes, and all hyperparameters were frozen before the critic-removal study. The study was not preregistered.

Preference-IX applies the auxiliary relative-comparison update directly to its own policy, with no audits. It shares the same three critic variants. It is included to measure the value of the recorded relative feedback, separately from the numerical-regret guarantees in the theory. Audited-mean records each selected action's admitted score, chooses the action with largest empirical mean, and mixes in 5% uniform exploration. An unobserved action has initial estimate 0.5; exact ties are broken by index. This simple baseline has no adversarial guarantee. Uniform samples actions uniformly and uses no audit or critic.

Configurations and seeds. The two complete-coverage datasets have $(T, K) = (8677, 9)$ and $(448, 34)$. Each is replayed for helpfulness and correctness, in numeric-ID order, a random permutation, and stable order by the two responses' total word count. Length sorting is a controlled order-sensitivity test based on response word counts alone. Deployment timestamps are unavailable in the source data.

Budgets are 1%, 5%, 10%, 20%, and 100% of the horizon, rounded to the nearest integer with minimum one. Every audit method spends exactly that number of admissions on every path. There are 128 seeds per setting. The fallback/action/gate tape uses seed $30000+s$, the score-index tape uses $20000+s$, and the shuffled order uses $5000+s$. A panel-mean sensitivity is run only for helpfulness, shuffled input, and Audit-IX. Altogether the real-record replay contains 358 configurations, 45,824 method trajectories, and 209,072,000 decisions. Repeated seeds revisit the same finite dataset under new algorithmic randomization; they are not additional human observations.

Figure 7(a) and (b) report the two target attributes on the 34-rule validation partition. Table 10 reports the order and objective comparisons at 20% admissions. Table 11 reports score and judgment-record costs across admission budgets. The accompanying results report all configurations, including the two additional critics and panel-mean sensitivities.



(a) Helpfulness.

(b) Correctness.

Figure 7: The 448-round, 34-rule validation replay. Panels (a) and (b) use the same prompts and access protocol with different declared numerical targets. All algorithms start afresh. Error bars are Monte Carlo standard errors times 1.96, conditional on these records.

F.4 The separate 128-seed replay

Table 9 reports the 128-seed replay separately from the 64-seed control study. The two sets of runs are analyzed independently. Within these 128-seed runs the paired Audit+comparison minus Preference-IX regret difference is 5.9245 ± 0.9747, and the difference from Audited-mean is 32.0952 ± 7.3664 (Monte Carlo standard errors times 1.96). Both favour the simpler comparator on this fixed record set. Preference-IX is called comparison-only in the main text.

Table 9: **Paid resources and decision quality.** Nine-rule helpfulness replay, $T = 8{,}677$, shuffled inputs. Audit methods use a 20% admission budget. Counts of scalar scores and accessed judge records are means over 128 seeds; regret is mean ± 1.96 Monte Carlo standard errors. Advice and comparison-only use recorded GPT-4o judgments.

Method	Admissions	Scalar scores	Judge records	Regret
Audit-IX	1,735	2,284	16,287	165.6 ± 2.9
Audit+comparison	1,735	2,182	17,595	97.1 ± 2.1
Preference-IX	0	0	17,347	91.1 ± 2.0
Audited-mean	1,735	1,788	9,430	65.0 ± 7.1
Uniform	0	0	16,388	222.4 ± 2.2

G Connections to the closest literature

G.1 Relative feedback and numerical objectives

Dueling bandits study sequential decisions from pairwise preferences; their objectives can be defined through preference probabilities, utility models, or pairwise winners (Yue et al., 2012; Bengs et al., 2021). REX3 uses a utility-based comparison model (Gajane et al., 2015), whereas adversarial dueling bandits with arbitrary preference matrices measure regret against the Borda winner (Saha et al., 2021). Best-of-both-world analyses, contextual policy classes, realizability, and adversarial contextual feedback extend the decision problem in complementary directions (Saha & Gaillard, 2022; Dudík et al., 2015; Saha & Krishnamurthy, 2022; Li et al., 2024; Di et al., 2025). Dueling convex optimization also permits general preference mechanisms under its optimization structure (Saha et al., 2025). Their guarantees depend on the specified preference objective, comparison model, and optimization assumptions.

Table 10: All objectives and input orders at 20% admissions. GPT-4o is the comparison source; means and Monte Carlo intervals use 128 seeds. The two rule sets have different horizons and are not directly comparable.

Rules / target	Order	Audit-IX	+ Comparison	Comparison only	Audited mean
9 rules / Help.	id	166.5 ± 2.8	90.3 ± 2.1	82.8 ± 2.0	67.0 ± 8.7
9 rules / Help.	shuffled	165.6 ± 2.9	97.1 ± 2.1	91.1 ± 2.0	65.0 ± 7.1
9 rules / Help.	length	169.6 ± 3.0	104.1 ± 2.2	97.3 ± 2.1	71.1 ± 9.5
9 rules / Corr.	id	159.7 ± 2.6	86.5 ± 2.0	79.5 ± 1.9	64.8 ± 8.4
9 rules / Corr.	shuffled	159.8 ± 2.8	92.6 ± 1.9	87.2 ± 1.9	60.6 ± 6.6
9 rules / Corr.	length	162.7 ± 2.9	99.7 ± 2.1	93.7 ± 2.0	62.7 ± 7.7
34 rules / Help.	id	16.3 ± 0.5	15.4 ± 0.5	15.3 ± 0.5	16.0 ± 1.6
34 rules / Help.	shuffled	16.0 ± 0.4	15.6 ± 0.4	15.4 ± 0.4	17.8 ± 1.6
34 rules / Help.	length	15.9 ± 0.4	15.1 ± 0.4	15.0 ± 0.4	16.3 ± 1.6
34 rules / Corr.	id	15.7 ± 0.4	15.0 ± 0.4	14.9 ± 0.4	15.4 ± 1.6
34 rules / Corr.	shuffled	15.3 ± 0.4	15.0 ± 0.4	15.0 ± 0.4	17.9 ± 1.6
34 rules / Corr.	length	15.5 ± 0.4	14.8 ± 0.4	14.7 ± 0.4	16.2 ± 1.6

Table 11: Primary nine-rule helpfulness replay across all budgets, with shuffled inputs and single-score audits. Admissions are exact; other counts are empirical means. Each source judge record combines two answer-order presentations, so a record is not a single model call.

Method	Admissions	Scalar scores	Judge records	Regret
Audit-IX	87	116	16,354	207.7 ± 3.3
Audit+comparison	87	109	17,639	100.8 ± 2.0
Audit-IX	434	577	16,336	193.8 ± 3.1
Audit+comparison	434	546	17,642	101.1 ± 2.0
Audit-IX	868	1,150	16,317	181.5 ± 2.9
Audit+comparison	868	1,091	17,620	99.5 ± 2.0
Audit-IX	1,735	2,284	16,287	165.6 ± 2.9
Audit+comparison	1,735	2,182	17,595	97.1 ± 2.1
Audit-IX	8,677	11,211	16,161	122.6 ± 2.3
Audit+comparison	8,677	10,878	17,546	93.2 ± 2.0

Maran et al. (2024) study ranking feedback with the numerical regret objective, while Liu et al. (2026) distinguish instantaneous and time-average rankings, prove impossibilities, and obtain positive guarantees through variation control or historical information. Our audit question keeps the numerical objective but changes the resource available to overcome instantaneous ambiguity. Our results separate known and unknown calibration under bounded current-pair comparisons, even when every known model admits bounded exact reconstruction. Theorem 3.1 also characterizes every continuous symmetric link family that permits comparison-only no-regret under a fixed label cap. A hard budget limits the numerical audits selected after those comparisons.

Preference-based reinforcement learning and language-model alignment often impose a reward representation and a comparison link (Ouyang et al., 2022; Zhu et al., 2023; Rafailov et al., 2023). Heterogeneous feedback makes the target aggregation consequential (Park et al., 2024). These works motivate the distinction between a recorded preference and a declared numerical reference; our theorem concerns online regret for that declared reference.

G.2 Observability, experiments, and scarce labels

The span of feedback probabilities is central to partial monitoring. Its classification separates games by which loss differences their signals can identify (Bartók et al., 2014; Lattimore & Szepesvári, 2019); linear formulations provide a common treatment of many sequential experiments (Kirschner et al., 2023). A finite comparison experiment is a special feedback structure within this broader tradition. The supplied-law benchmark in Appendix D quantifies the cost of a numerical channel for the part outside the cheap signal span. The main construction instead has zero such residual for every member: the additional uncertainty lies in which observation model is present. An augmented partial-monitoring description can include the calibration as a latent state. The construction estab-

lishes the separation while keeping that state fixed throughout the run, with uniformly observable and well-conditioned known-state slices; the calibration in our lower-bound populations is fixed while numerical utilities are drawn independently.

Uniform-approximation duality produces matching distributions in many statistical lower bounds (Wu & Yang, 2020). We use that classical duality to produce mutually singular components and then prove that their mixture weights survive every allowed pre-audit transcript. For the supplied-law benchmark, the exact simulator identifies the resulting audit experiment in the sense of Blackwell (1953). Matching the unconditional feedback law alone would not justify conditioning on the learner's selected audit events; the prefix-matching property supplies that step. This equivalence preserves audit timing without optimizing it. The main unknown-calibration construction instead uses overlapping four-point numerical populations. Selection can improve its per-audit Fisher information, while a transcript-uniform divergence bound controls its order. The coin equivalence belongs to the known-experiment benchmark; the unknown-calibration proof uses the conditional-divergence argument formalized in Proposition I.1.

Label-efficient prediction already has the $T/\sqrt{B}$ dependence when only B labels are available (Cesa-Bianchi et al., 2005). Information-constrained online learning studies adaptation to easier loss sequences under related budgets (Mitra & Gopalan, 2020). Our audit-only rate uses this existing principle. The split algorithm uses the exponential-weights and partial-feedback tools developed in Auer et al. (2002); Audibert & Bubeck (2010); Neu (2015), with expert advice as in Beygelzimer et al. (2011). General background appears in Cesa-Bianchi & Lugosi (2006); Lattimore & Szepesvári (2020). Resource constraints in bandits with knapsacks usually price action-dependent consumption while maximizing total reward (Badanidiyuru et al., 2013); here the scarce resource reveals numerical information after actions already accrue utility.

Theorem 3.1 combines finite-session polynomial observability with a decision requirement across fixed unknown links. A common polynomial may recover utility only up to a link-dependent positive scale, which is sufficient because the scale is fixed across rounds. Its necessity proof separates a failure within one experiment from conflicting numerical interpretations of the same latent comparisons. The moment and testing ingredients are classical; the result is the necessary-and-sufficient condition for this fixed-link, current-pair protocol. Proposition J.2 then connects regular shape variation to the conditional score information needed for an audit rate.

G.3 Correction, multi-fidelity information, and active evaluation

Inverse-probability weighting corrects unequal observation probabilities (Horvitz & Thompson, 1952). Doubly robust policy evaluation combines a numerical predictor with a weighted correction (Dudík et al., 2011), and optimistic online learning exploits predictable structure when its errors are small (Rakhlin & Sridharan, 2013). These are conceptual precedents for separating a useful prediction from the information required to repair it. Our positive algorithms use these established correction and exponential-weights tools; the main result is the calibration–audit separation.

Multi-fidelity bandits formalize a hierarchy of evaluations with different costs and accuracies (Kandasamy et al., 2016; Wang et al., 2023b). Prediction-powered inference uses a few labels to correct model-based predictions (Angelopoulos et al., 2023); active inference and stratified hybrid language-model evaluation allocate labels to improve precision (Zrnic & Candes, 2024; Fisch et al., 2024). Selection bias in active learning, online power tuning, and prediction-powered active testing further study correction under adaptive acquisition or learning (Farquhar et al., 2021; Shoham et al., 2025; Ashouritaklimi et al., 2026). These works study estimation, confidence intervals, or learning with purchased labels. Sequential reward improvement is also explicitly addressed by Ji et al. (2025): their MLA-UCB method uses surrogate rewards and learns covariance information, with a jointly Gaussian analysis and a batched non-Gaussian extension. Their stochastic assumptions enable efficiency improvements in sequential decisions. Our lower bound instead shows that exact reconstruction under each known comparison model can coexist with an additional numerical-information requirement when calibration is unknown.

Recent allocation methods address robustness to unreliable uncertainty scores (Li et al., 2025), learned labeling policies (Ma & Candès, 2026), and cost-aware combinations of several predictors (Cowen-Breen et al., 2026; Brawand et al., 2026). These are direct precedents for learning numerical corrections and pricing their inputs. Our matching upper bound uses nominal and label-efficient

updates. Its role is to attain the calibration–audit rate for numerical external regret; the lower bound identifies the extra information needed relative to known calibration.

G.4 Unknown links and adaptation to predictions

Wei et al. (2020) prove that loss-prediction errors and whether those errors are known change the attainable contextual-bandit rates; they also identify costs of choosing among multiple predictors. This is an important precedent for information-dependent adaptation gaps. Their observations are numerical loss predictions and bandit outcomes. Here the unknown object is the numerical interpretation of a finite comparison experiment, and every supplied interpretation has a bounded exact statistic. The hard pair preserves the full comparison law while changing that interpretation.

Zhang & Ying (2025) address preference-based reinforcement learning without a known link. Their sign-based zeroth-order update aligns with the policy gradient and converges to a stationary policy under their optimization conditions. The optimization conditions support a positive result with an unknown link. Our objective is cumulative numerical external regret under unrestricted utility variation, with a cap on the current-pair experiment. Proposition B.1 gives a positive result inside our own setting when the unknown link contributes only a positive numerical scale.

G.5 Finite sessions, evaluation, and routing

A bounded session is different from a random session with bounded *expected* length. Bernoulli factories and inverse-binomial sampling construct nonlinear unbiased estimators by allowing random stopping (Keane & O'Brien, 1994; Nacu & Peres, 2005; Holtz et al., 2011; van Opheusden et al., 2020). Such a protocol changes the observable experiment and is different from the fixed caps in Theorems 3.2 and 4.1 and the finite-tree benchmark. Similarly, a cap on queries does not specify a wall-clock deadline. Delayed-feedback learning treats the statistical effect of unavailable results as a separate resource (Joulani et al., 2013; Zimmert & Seldin, 2020); our bounds count synchronous observations rather than elapsed time.

Classifier confidence calibration concerns agreement with correctness probabilities (Guo et al., 2017), rather than the numerical interpretation of pairwise preferences studied here. LLM evaluation also exhibits position and length effects (Zheng et al., 2023; Wang et al., 2023a; Dubois et al., 2024). HelpSteer2 and its preference extension supply the distinct numerical and ordinal annotations used here (Wang et al., 2024; 2025); the recorded LLM verdicts and their two-position protocol come from RLHFlow (Li, 2025b;a). RewardBench assesses reward models (Lambert et al., 2024), while RouterBench, RouteLLM, and dueling-feedback routing address selection among language models (Hu et al., 2024; Ong et al., 2024; Chiang et al., 2025). Our action is instead a recorded judgment rule selecting between cached answers. This permits observation-budget replays on fixed answers; generation costs and the answers' source-model identities are not evaluated. The frozen numerical diagnostic uses the ridge implementation in scikit-learn (Pedregosa et al., 2011).

H Scope and remaining questions

Theorem 3.1 classifies zero-audit learnability for every continuous symmetric single-link family under fixed-cap conditionally independent comparisons. Its alternatives are square-root and linear uniform minimax regret. Theorem 3.2 gives a uniform rate for the explicit cubic inverse family with at most three labels per round. Theorem 4.1 covers a different, quintic inverse family and every fixed cap $m \geq 5$, with constants allowed to depend on m. Proposition I.1 identifies sufficient conditions shared by both constructions. Applying it to another comparison family requires checking equality of the allowed transcripts, conflicting numerical optima, and conditional audit information; the zero-audit classification alone does not determine the conditional information in purchased scores. The families are analytic constructions, rather than fitted models of the recorded judgments.

The rates are uniform in T, B, r for fixed K and, in the extension, fixed m. Sharp growing-K dependence, caps that increase with the horizon, and unbounded random sessions require additional analysis. For zero audits, the common-polynomial criterion identifies the positive classes. Equation (J.8) gives the full audit-budget rate for any class containing a nonpolynomial inverse, and Proposition J.2 treats regular normalized shape variation. Audited rates for arbitrary remaining classes depend on

the numerical-score experiment and parameterization. A learner may improve on the worst-case rate on easier populations, even though uniform adaptation to the known-calibration benchmark is ruled out for the constructed families at small audit budgets.

Numerical audits measure a declared target. Exact selected scores support the lower bounds and nonlinear residual implementation; bounded, conditionally unbiased references support the audit-only update. Sampling a listed scorer measures the corresponding finite score-list mean. Extending this guarantee to another scoring process requires the same conditional-mean property or an explicit bias allowance. Audit admissions, scalar measurements, and response delay remain distinct quantities.

The empirical study reuses a finite corpus. Paired Monte Carlo intervals describe algorithmic variation on these records; the frozen diagnostic uses prompt bootstraps conditional on its fitted predictor. At 20% admissions, Audit+comparison has higher regret than comparison-only in the reported shuffled-training helpfulness removal controls; the matching validation differences remain uncertain. New user evaluations with additional judge libraries and populations would assess generalization; measurements of scoring prices and latency would quantify operating costs.

I A SEPARATION CRITERION AND A QUINTIC EXTENSION

The four-point construction in Appendix B changes a comparison link and a utility population together. This appendix separates the conditions needed by that argument from its particular formulas. We then verify them for a quintic inverse link and every fixed comparison cap $m \geq 5$. The new population has a continuous density on the common numerical support $[-1, 1]$; it is distinct from Table 1.

I.1 A SUFFICIENT CONDITION FOR A CALIBRATION–AUDIT SEPARATION

Fix the payoff and timing in Section 2, with $K \geq 2$. Partition the actions into two nonempty types, with utilities X_t and $-X_t$. A comparison transcript contains all permitted labels, query choices, orientations, early-stop decisions, and public auxiliary observations. The conditions below concern conditional kernels at a *common* observed history, including any learner randomization exposed there.

Proposition I.1 (A sufficient condition)**.** *Let a comparison family have parameter range $[-r_0, r_0]$, where $0 < r_0 \leq 1$. Suppose there are constants $a, C > 0$ and $0 < \delta_0 \leq r_0$ such that, for each $0 < \delta \leq \delta_0$, two environments with parameters $+\delta$ and $-\delta$ satisfy the following conditions, after a fixed relabeling of the parameter sign if necessary.*

1. *Each parameter is fixed for the run. Within each environment, X_t are independent and identically distributed, independently of past observations and the learner's fresh current draws.*
2. *For every allowed action pair and comparison strategy, the complete pre-audit transcript has the same conditional law in both environments. This includes random query selection and every allowed stopping and orientation choice.*
3. *The numerical means satisfy $\mathbb{E}_{+\delta} X_t \geq a\delta$ and $\mathbb{E}_{-\delta} X_t \leq -a\delta$.*
4. *Conditional on any positive-probability pre-audit transcript, chosen actions, and common history, the laws of* all *disclosed numerical scores have KL divergence at most $C\delta^2$, in the direction $+\delta$ to $-\delta$. This includes identical-action and same-type audits.*

For $0 < r \leq r_0$, any learner with unknown parameter in $[-r, r]$ and at most B admissions on every path has worst-case expected regret at least

$$\frac{3aT}{4} \min\left\{ r, \delta_0, \frac{1}{\sqrt{8C(B+1)}} \right\}. \tag{I.1}$$

At $r = 0$ the asserted additional lower bound is zero. In particular, for fixed a, C, δ_0, the bound is $\Omega\big(T \min\{r, (B+1)^{-1/2}\}\big)$ uniformly over T, B, r in these ranges.

Proof. First fix an admissible δ. Factor the observed-history law round by round. At a common entering history the learner uses the same kernels for its announced distribution, private random draws, and current proposals. Condition 2 makes the comparison kernel identical. The audit-selection kernel, which may depend on the entire current transcript, is also the same at this common history. Thus only an admitted score contributes divergence. Condition 4 and the KL chain rule give

$$\mathrm{KL}(P^T_{+\delta} \| P^T_{-\delta}) \le C\delta^2 \mathbb{E}_{+\delta}\left[\sum_{t=1}^{T} C_t\right] \le CB\delta^2. \tag{I.2}$$

Condition 4 applies to the full score vector, so this calculation does not discard absolute-score information when the selected half-gap is zero. Learner-private randomization can be included in the history or integrated through its common kernel. Adaptive query and audit decisions are therefore covered.

Let $s_t \in [-1, 1]$ be the average signed type of the two proposals. Independence of the current population draw gives expected round payoff $(\mathbb{E}_{\pm\delta} X_t)\mathbb{E}_{\pm\delta} s_t$. The hindsight comparator is at least the cumulative payoff of the best mean type. Consequently the two expected regrets are bounded below by

$$a\delta \sum_{t=1}^{T}(1 - \mathbb{E}_{+\delta} s_t) \quad \text{and} \quad a\delta \sum_{t=1}^{T}(1 + \mathbb{E}_{-\delta} s_t),$$

respectively. Prefix total variation is at most the total variation of the full history; adjoining fresh action draws through a common kernel cannot increase it. Since s_t is bounded by one,

$$\frac{\mathbb{E}_{+\delta} R_T + \mathbb{E}_{-\delta} R_T}{2} \ge a\delta T \left(1 - \mathrm{TV}(P^T_{+\delta}, P^T_{-\delta})\right).$$

Choose $\delta = \min\{r, \delta_0, [8C(B+1)]^{-1/2}\}$. Pinsker's inequality and (I.2) give total variation at most $1/4$, proving (I.1). This choice also covers $B = 0$. For the last assertion, put $c_* = \min\{1, \delta_0, (8C)^{-1/2}\}$. The chosen δ is at least $c_* \min\{r, (B+1)^{-1/2}\}$ because $r \le 1$. □

Conditions for matching upper bounds. The proposition is a lower-bound criterion. To obtain the known-calibration benchmark, suppose each known member admits a legal transcript statistic D_θ with $\mathbb{E}_\theta[D_\theta \mid x] = x$ for every $x \in [-1, 1]$ and $|D_\theta| \le A$, uniformly over the parameter range. The one-channel update in Appendix C then gives $16A\sqrt{KT \log K}$ regret without audits. To match the unknown-calibration rate, suppose in addition that the learner can compute a nominal statistic D_*, without knowing θ, such that

$$|D_*| \le A_*, \qquad \sup_{|\theta| \le r,\, |x| \le 1} |\mathbb{E}_\theta[D_* \mid x] - x| \le Lr$$

with fixed A_*, L. The same one-channel proof, with the pointwise bias retained, gives $16A_*\sqrt{KT \log K} + 4LrT$. The numerical update (B.12) gives the other route. Selecting the smaller bound before the run yields

$$O\Big(\sqrt{T} + T \min\{r, (B+1)^{-1/2}\}\Big)$$

for fixed K, A_*, L, with $B = 0$ handled by the nominal route. These upper bounds hold for predictable utility sequences, beyond the independent populations used in the proposition.

For completeness, the separate square-root lower bound uses utilities $(\xi_t, -\xi_t)$ with independent Rademacher ξ_t, duplicating one type as needed. Even a learner shown all past numerical utilities has expected payoff zero. The hindsight comparator has expectation $\mathbb{E}|\sum_t \xi_t| \ge \sqrt{T/2}$. Thus the minimax value is at least the maximum of this bound and (I.1), which is within a factor of two of their sum. The two terms need not be realized by the same population. This also establishes the endpoint $r = 0$.

Checking the original construction. Appendix B verifies the criterion with parameter sign reversed, $a = 1/5$, $C = 512$, and $\delta_0 = 1/64$. The comparison equality is (B.5); the full-score condition is (B.8) and its KL consequence, including same-type pairs. Equation (B.3) supplies $A = 4$, and the nominal statistic supplies $A_* = 2$, $L = 2$. This recovers the mechanism of Theorem 3.2. The criterion is sufficient: absence of a common unbiased statistic alone would not establish its decision-conflict and conditional-information conditions.

I.2 A FIXED LATENT COMPARISON DISTRIBUTION

A useful way to satisfy comparison equality is to fix the latent distribution before choosing its numerical interpretation. Let V have a positive C^1 density ν on $[-b, b]$, with $b < 1$. Let g_θ be increasing odd maps from this interval onto the same interval $[-1, 1]$. Define

$$X_\theta = g_\theta(V), \qquad p_\theta(x) = \frac{1 + g_\theta^{-1}(x)}{2}.$$

Assume these maps are continuously differentiable in θ and twice continuously differentiable in v, with uniform bounds

$$\nu \geq n_0 > 0, \quad |\nu'| \leq N, \quad c \leq g'_\theta \leq c', \quad |g''_\theta| \leq M, \quad |\partial_\theta g_\theta| \leq H, \quad |\partial_\theta g'_\theta| \leq H_1,$$

where $c > 0$. The prime denotes a derivative in v. These are checkable bounds on densities and maps, not assumptions about a learner's access to V.

Given $V = v$, a cross-type comparison has sign mean v or $-v$, according to orientation. For a realized transcript ω of length $n \leq m$, the label likelihood is a product of factors $(1 + w_j \sigma_j v)/2$, where $w_j, \sigma_j \in \{-1, 1\}$. Realized query and stopping probabilities contribute an additional factor independent of θ. Hence its marginal

$$M_\omega = \int_{-b}^{b} \nu(v) L_\omega(v)\, dv$$

is independent of θ. This equality applies to every allowed prefix and randomized stopping strategy by conditioning on the query randomization. Same-type comparisons have zero half-gap and fair labels, so their entire transcript is independent of X_θ and θ.

Let $v_\theta(x) = g_\theta^{-1}(x)$. The common numerical support and change of variables give

$$f_\theta(x) = \frac{\nu(v_\theta(x))}{g'_\theta(v_\theta(x))}, \qquad Q_\theta(x \mid \omega) = \frac{f_\theta(x) L_\omega(v_\theta(x))}{M_\omega}. \tag{I.3}$$

The inverse function theorem gives $|\partial_\theta v_\theta(x)| \leq H/c$. Moreover, $|\partial_v \log L_\omega(v)| \leq m/(1-b)$. Differentiating the conditional density at a fixed numerical x yields

$$\begin{aligned} \partial_\theta \log Q_\theta(x \mid \omega) &= \left(\frac{\nu'}{\nu} + \frac{L'_\omega}{L_\omega} - \frac{g''_\theta}{g'_\theta} \right)(v_\theta(x))\, \partial_\theta v_\theta(x) - \frac{\partial_\theta g'_\theta}{g'_\theta}(v_\theta(x)), \\ \left|\partial_\theta \log Q_\theta(x \mid \omega)\right| &\leq \Lambda_m := \frac{H}{c}\left(\frac{N}{n_0} + \frac{m}{1-b} + \frac{M}{c} \right) + \frac{H_1}{c}. \end{aligned} \tag{I.4}$$

The normalization has no derivative because M_ω is parameter-independent. For a same-type pair, omit the likelihood derivative; an exact selected score still reveals X_θ up to its known sign. For a cross-type pair the two scores are $(X_\theta, -X_\theta)$ in known order. Thus both cases reduce to the displayed density, even when the audit discloses both absolute scores.

A bounded conditional log derivative gives

$$\mathrm{KL}(Q_\delta(\cdot \mid \omega) \| Q_{-\delta}(\cdot \mid \omega)) \leq \left(e^{2\Lambda_m \delta} - 1\right)^2 \leq 8\Lambda_m^2 \delta^2, \qquad 0 < \delta \leq (8\Lambda_m)^{-1}. \tag{I.5}$$

Indeed, the log likelihood ratio is bounded by $2\Lambda_m\delta$, KL is at most chi-squared divergence, and $e^y - 1 \leq y e^y$ for $0 \leq y \leq 1/4$. After coarsening a transcript, the conditional weights of its refinements are the same in both environments because every refined marginal is parameter-independent. Retaining the refinement gives the same uniform KL bound, and data processing gives the bound for the coarser observation. This proves the conditional audit condition, not just equality of the unconditional comparison laws.

To obtain exact known-calibration reconstruction within d labels, additionally suppose $g_\theta(v) = \sum_{k=0}^{d} a_k(\theta) v^k$, with $\sum_k |a_k(\theta)| \leq A$. Define S_k as the average product over all k-element subsets of d conditional independent signs, and $S_0 = 1$. Then $\mathbb{E}[S_k \mid x] = v_\theta(x)^k$ and $|S_k| \leq 1$, so $\sum_k a_k(\theta) S_k$ is unbiased for x and bounded by A. The statistics S_k can be correlated. Their product-expectation identities, rather than their independence, establish unbiasedness. A separation follows when $\int g_\theta(v)\nu(v)\, dv$ changes sign at a linear rate and a bounded nominal statistic has $O(r)$ bias. The next construction verifies these remaining requirements explicitly.

I.3 PROOF OF THE QUINTIC EXTENSION

Let $\varepsilon_0 = 1/2$, $\beta_0 = 1/4$, and $I = [-1/2, 1/2]$. Define

$$\nu(v) = 1 + \varepsilon_0(20v^3 - 3v), \qquad \alpha_\theta = \theta - 5\beta_0/18,$$
$$\widetilde{g}_\theta(v) = 2v + \alpha_\theta(v^3 - v/4) + \beta_0(v^5 - v/16), \quad \widetilde{p}_\theta(x) = \frac{1 + \widetilde{g}_\theta^{-1}(x)}{2}, \qquad |\theta| \le 1/16. \tag{I.6}$$

The coefficient of v^5 is nonzero. This family therefore differs from the cubic inverse in (B.1). For the lower bound take independent $V_t \sim \nu$ and utilities $(\widetilde{g}_\theta(V_t), -\widetilde{g}_\theta(V_t))$, with θ fixed.

Density and numerical mean. The odd perturbation integrates to zero on I, so $\int_I \nu = 1$. The critical points of $20v^3 - 3v$ are $\pm 1/\sqrt{20}$, with values $\mp 1/\sqrt{5}$; its endpoint values are ± 1. Consequently $1/2 \le \nu \le 3/2$ and $|\nu'| \le 6$. Direct integration gives

$$\mathbb{E}_\nu V = 0, \qquad \mathbb{E}_\nu V^3 = \frac{\varepsilon_0}{140}, \qquad \mathbb{E}_\nu V^5 = \frac{\varepsilon_0}{504}, \qquad \mathbb{E}\widetilde{X}_\theta = \frac{\varepsilon_0\theta}{140} = \frac{\theta}{280}. \tag{I.7}$$

The constant $-5\beta_0/18$ in α_θ cancels the fixed fifth-moment contribution because $140/504 = 5/18$. The density is asymmetric even though $\mathbb{E}V = 0$. The best mean type therefore reverses between $\theta = \delta$ and $-\delta$.

Uniform regularity. Both perturbation polynomials vanish at $v = \pm 1/2$, giving $\widetilde{g}_\theta(\pm 1/2) = \pm 1$. Throughout the parameter interval,

$$-\frac{19}{144} \le \alpha_\theta \le -\frac{1}{144}, \qquad \widetilde{g}'_\theta(v) = 2 + \alpha_\theta(3v^2 - 1/4) + \beta_0(5v^4 - 1/16).$$

Using $|3v^2 - 1/4| \le 1/2$ and $|5v^4 - 1/16| \le 1/4$ gives

$$\frac{7}{4} < \frac{539}{288} \le \widetilde{g}'_\theta(v) \le \frac{613}{288} < \frac{9}{4}, \qquad \frac{1}{4} \le \widetilde{p}_\theta(x) \le \frac{3}{4}, \qquad \frac{2}{9} < \widetilde{p}'_\theta(x) < \frac{2}{7}. \tag{I.8}$$

The odd polynomial is thus a bijection between the stated intervals, with a smooth increasing odd inverse. The link is smooth jointly in x and θ and satisfies $\widetilde{p}_\theta(-x) = 1 - \widetilde{p}_\theta(x)$. Endpoint derivatives are understood through its smooth local extension. Also,

$$|\widetilde{g}''_\theta(v)| \le 3(19/144) + 5/8 = 49/48 < 2.$$

These interval bounds establish regularity for every parameter, independently of the numerical checks.

A five-label exact statistic and a nominal one. For five conditionally independent signs set

$$S_1 = \frac{1}{5}\sum_{i=1}^{5} W_i, \qquad S_3 = \binom{5}{3}^{-1} \sum_{i<j<k} W_i W_j W_k, \qquad S_5 = \prod_{i=1}^{5} W_i.$$

With the known parameter, use

$$\widetilde{D}_\theta = (2 - \alpha_\theta/4 - \beta_0/16)S_1 + \alpha_\theta S_3 + \beta_0 S_5. \tag{I.9}$$

Each conditional product has expectation equal to the corresponding power of $\widetilde{g}_\theta^{-1}(x)$. Thus $\mathbb{E}_\theta[\widetilde{D}_\theta \mid x] = x$ for every $x \in [-1, 1]$. Since $|S_j| \le 1$,

$$|\widetilde{D}_\theta| \le 2 + \frac{5}{4}\frac{19}{144} + \frac{17}{16}\frac{1}{4} = \frac{175}{72} < \frac{5}{2}.$$

This gives zero known-model approximation residual and at most $40\sqrt{KT\log K}$ regret. When $m > 5$, the known-calibration learner uses only five labels.

For unknown calibration, $\widetilde{D}_0$ is computable from the family centre. Put $h(v) = v^3 - v/4$. Its maximum absolute value on I is $1/(12\sqrt{3})$, attained at $v = \pm 1/\sqrt{12}$. Because $\widetilde{g}_\theta - \widetilde{g}_0 = \theta h$,

$$\sup_{|\theta| \le r,\, |x| \le 1} |x - \mathbb{E}_\theta[\widetilde{D}_0 \mid x]| \le \frac{r}{12\sqrt{3}}. \tag{I.10}$$

The nominal route therefore gives $40\sqrt{KT\log K} + rT/(3\sqrt{3})$. Its inputs are the known family centre, T, and K; choosing between routes additionally uses the declared r, B.

Complete transcripts and selected numerical scores. The latent distribution ν is fixed, so the argument in Appendix I.2 matches every ordered transcript for *any fixed* finite cap m, including reversals, stopping, and same-type pairs. This step uses all m labels available to the unknown learner, not only the five used by the known learner. The marginal comparison law is independent of θ while the numerical density in (I.3) has common support $[-1, 1]$ and satisfies $2/9 \le f_\theta \le 6/7$.

For an explicit conditional-information constant, $|h| \le 1/16$, $|h'| \le 1/2$, and (I.8) give $|\partial_\theta v_\theta| \le 1/28$. Substituting $n_0 = 1/2$, $N = 6$, $c = 7/4$, $M = 2$, and $b = 1/2$ into (I.4) gives

$$|\partial_\theta \log Q_\theta(x \mid \omega)| \le \frac{37}{49} + \frac{m}{14} \le L_m := 1 + \frac{m}{14}, \qquad \mathrm{KL}(Q_\delta \| Q_{-\delta}) \le C_m \delta^2, \quad C_m = 8L_m^2, \tag{I.11}$$

for $\delta \le \min\{1/16, (8L_m)^{-1}\}$. The conditional KL bound holds for each transcript. The full disclosed score vector is a known function of X_θ; it reveals X_θ even on same-type pairs. This covers the full audit access in Section 2.

Regret and endpoints. Apply Proposition I.1 with $a = 1/280$ and choose

$$\delta = \min\left\{r, \frac{1}{8L_m\sqrt{B+1}}\right\}.$$

The total history KL is at most $1/8$. Hence

$$\widetilde{V}^{\mathrm{unk}}_{T,m}(r, B) \ge \frac{3T}{1120} \min\left\{r, \frac{1}{8L_m\sqrt{B+1}}\right\} \ge \frac{3T}{8960L_m} \min\{r, (B+1)^{-1/2}\}. \tag{I.12}$$

Combining the independent Rademacher lower bound with (I.10) and the audit-only upper bound proves Theorem 4.1. This includes $r = 0$, $B = 0$, and $B = T$; for $B = 0$ the nominal route is available, and for $r = 0$ its bias vanishes. The lower-bound population is independent over rounds, whereas both upper routes allow all predictable utility sequences.

The constants in the unknown-calibration rate may depend on fixed m through L_m; the known-calibration estimator always uses five labels. The result does not assert a uniform rate when m grows with T or for unbounded random session lengths. The necessity of a new population is visible in the original four-point construction: its fourth latent moment is

$$\mathbb{E}_\theta V^4 = \frac{Mz^4 + 1 - M}{16} = \frac{5 - 3z^2}{128}, \qquad z = 1/2 + \theta, \tag{I.13}$$

which varies with θ. Matching the first three moments there does not match a four-label experiment. The continuous population here fixes the entire latent comparison distribution instead.

J Classification of comparison-only learnability

This appendix proves Theorem 3.1. The class may contain any number of continuous symmetric links $p : [-1, 1] \to [0, 1]$. The unknown link is fixed throughout the run. The learner receives at most $m \ge 1$ conditionally independent labels from the current pair, with the proposal rule and numerical regret of Section 2, and receives no numerical scores. In particular, the theorem covers every family of smooth links in this protocol without assuming monotonicity, polynomial inverses, a common latent interval, or a lower bound on signal strength.

Put $v_p(x) = 2p(x) - 1$. Symmetry means $v_p(-x) = -v_p(x)$. Define the finite-dimensional space

$$\mathcal{L}_{p,m} = \mathrm{span}\{1, v_p, \dots, v_p^m\} \subset C([-1, 1]). \tag{J.1}$$

Whenever v_p is injective, it has an inverse g_p on $I_p = [-a_p, a_p]$, where $a_p = |v_p(1)| > 0$. We show below that injectivity and a polynomial inverse are necessary for comparison-only learning, even when p is supplied. The class and cap are fixed as T increases; the minimax supremum over the class is taken separately at each horizon. All constants in the classification may depend on this fixed class, K, and m.

J.1 What a finite session observes

For ordered signs $w = (w_1, \dots, w_m)$ in a fixed presentation, the likelihood conditional on $V = v$ is

$$L_w(v) = 2^{-m} \prod_{j=1}^{m} (1 + w_j v). \tag{J.2}$$

Its degree in v is at most m. More generally, conditioning on the learner's private randomization makes its presentation and stopping rule a deterministic tree. A leaf of length $n \le m$ has likelihood $2^{-n} \prod_{j=1}^{n} (1 + w_j \sigma_j v)$, where the realized presentation signs $\sigma_j \in \{-1, 1\}$ can depend on the earlier labels. Its likelihood is again a polynomial of degree at most m. Thus matching moments through degree m matches every allowed transcript, including the query choices and stopping event. Integrating back over private randomization preserves this equality.

Conversely, every polynomial $P(v) = \sum_{j=0}^{m} b_j v^j$ has a bounded transcript statistic. With m labels, use

$$D_P = \sum_{j=0}^{m} b_j \prod_{\ell=1}^{j} W_\ell, \qquad \mathbb{E}[D_P \mid V = v] = P(v), \qquad |D_P| \le H := \sum_{j=0}^{m} |b_j|. \tag{J.3}$$

The empty product is one. Conditional independence establishes each product expectation; the products themselves can be dependent. These are the finite-experiment span properties used in partial monitoring and approximation-based lower bounds (Bartók et al., 2014; Kirschner et al., 2023; Wu & Yang, 2020).

J.2 Failure to recover a numerical gap under a known link

Fix one link p for which the coordinate function $x \mapsto x$ does not belong to $\mathcal{L}_{p,m}$. Let σ be uniform probability measure on $[-1, 1]$, and let Π be the $L^2(\sigma)$ projection onto $\mathcal{L}_{p,m}$. Writing $\iota(x) = x$, define

$$f = \iota - \Pi\iota, \qquad M = \|f\|_\infty.$$

Reflection preserves $\mathcal{L}_{p,m}$, so uniqueness of the projection makes f odd. Continuity and failure of membership give $M > 0$ and $\int f^2 \, d\sigma > 0$. In particular,

$$\int v_p(x)^j f(x) \, d\sigma(x) = 0 \quad (0 \le j \le m), \qquad \int x f(x) \, d\sigma(x) = \int f(x)^2 \, d\sigma(x) > 0. \tag{J.4}$$

The two probability measures

$$d\mu_\pm(x) = \left(1 \pm \frac{f(x)}{2M}\right) d\sigma(x) \tag{J.5}$$

have densities between $1/2$ and $3/2$ relative to σ. Their comparison moments agree, but

$$\mathbb{E}_{\mu_\pm} X = \pm d_p, \qquad d_p = \frac{1}{2M} \int f(x)^2 \, d\sigma(x) > 0. \tag{J.6}$$

Choose independent draws X_t from either population and utilities $(X_t, -X_t)$; duplicate the second type when $K > 2$. Cross-type comparisons have matching moments in either presentation, and same-type comparisons are fair signs because their gap is zero. Appendix J.1 therefore gives the same full comparison history in both environments, for every learner. The link is the same in both and can even be supplied to the learner.

To convert this equality into regret, let s_t be the average signed type of the two proposals: action 1 has sign $+1$, all other actions have sign -1. The current population draw is independent of the entering history and fresh proposals. The hindsight comparator is at least the cumulative payoff of the best mean type, so

$$\mathbb{E}_+ R_T \ge d_p \sum_{t=1}^{T} (1 - \mathbb{E}_+ s_t), \qquad \mathbb{E}_- R_T \ge d_p \sum_{t=1}^{T} (1 + \mathbb{E}_- s_t). \tag{J.7}$$

The common history law gives $\mathbb{E}_+ s_t = \mathbb{E}_- s_t$. Averaging the two inequalities yields $\max\{\mathbb{E}_+ R_T, \mathbb{E}_- R_T\} \ge d_p T$. Thus no-regret requires $x \in \mathcal{L}_{p,m}$ for every member. In this case

$x = P_p(v_p(x))$ for a polynomial P_p of degree at most m, which can be made odd by symmetrization. This identity forces v_p to be injective. A continuous injective map on an interval is strictly monotone, so $g_p = v_p^{-1} = P_p$ on I_p. We have proved both the necessity of a polynomial inverse and the numerical decision conflict when that requirement fails.

The audit rate when even one known link fails this test. The same residual construction quantifies numerical audits. Let $h = f/(2M)$ and replace (J.5) by $d\mu_\alpha = (1 + \alpha h)d\sigma$, $|\alpha| \le 1$. Comparison moments are independent of α, and $\mathbb{E}_{\mu_\alpha} X = \alpha d_p$. Conditional on any transcript with positive probability, the numerical density is proportional to $(1 + \alpha h(x))L_\omega(v_p(x))$, with normalization independent of α. Its logarithmic derivative in α has absolute value at most one, because $|h| \le 1/2$. Therefore (I.5), with $\Lambda = 1$, gives a conditional audit KL bound $8\delta^2$ between $\alpha = \delta$ and $-\delta$ for $\delta \le 1/8$. The same bound holds for all disclosed score vectors, including same-type pairs. Proposition I.1, together with the audit-only upper bound, gives

$$V_T(\mathcal{F}, m, B) = \Theta_{\mathcal{F},K,m}\left(\frac{T}{\sqrt{B+1}}\right), \qquad 0 \le B \le T, \tag{J.8}$$

whenever $\mathcal{F}$ contains at least one link with $x \notin \mathcal{L}_{p,m}$; in particular, this includes every class containing an invertible link with nonpolynomial inverse. Here $V_T(\mathcal{F}, m, B)$ is the same uniform minimax value with the hard audit budget restored. The lower bound already holds when that particular link is known. Equation (J.8) concerns finite-session numerical information rather than an additional cost of unknown calibration.

J.3 Unknown nonlinear shape creates a second obstruction

The following elementary fact identifies exactly which numerical reinterpretations preserve all mean decisions.

Lemma J.1 (Mean decisions determine shape up to positive scale)**.** *Let g_0, g_1 be continuous, odd, strictly monotone functions on $[-b, b]$, with $b > 0$. Their expectations have the same sign under every probability measure on this interval if and only if $g_1 = cg_0$ for some $c > 0$. If positive proportionality fails, a measure supported on at most two points has expectations of strictly opposite signs.*

Proof. Positive proportionality preserves signs. If the two functions have opposite orientations, a point mass at any positive point already gives opposite signs. Otherwise, multiply both functions by -1 if needed to make them increasing. The ratio $g_1(v)/g_0(v)$ is nonconstant for $v > 0$ when positive proportionality fails. Choose $a, b' \in (0, b]$ where the ratios differ. Under $\nu = t\delta_a + (1-t)\delta_{-b'}$,

$$\mathbb{E}_\nu g_j(V) = tg_j(a) - (1-t)g_j(b'), \qquad t_j = \frac{g_j(b')}{g_j(a) + g_j(b')}$$

is its unique zero as a function of t. The unequal ratios imply $t_0 \ne t_1$. Choosing t strictly between these thresholds makes the two means have opposite signs. □

Suppose now that every member satisfies numerical recovery and hence has a polynomial inverse g_p of degree at most m on I_p. If two of these polynomials are not positively proportional, they remain outside a common positive ray on the common interval $I_{p_0} \cap I_{p_1}$: a polynomial identity on a nonempty open interval holds everywhere. Lemma J.1 gives a common latent population ν with

$$\mathbb{E}_\nu g_{p_0}(V) > 0, \qquad \mathbb{E}_\nu g_{p_1}(V) < 0$$

after relabeling the environments. Generate independent $V_t \sim \nu$ and use utilities $(g_{p_j}(V_t), -g_{p_j}(V_t))$ in environment j. Their numerical means conflict, whereas the whole latent law, and hence every current comparison transcript, is the same. This matches all finite caps simultaneously, not just m moments. Same-type pairs again produce fair signs. Induction over common histories handles the adaptive proposals and session rules.

Let $d = \min\{\mathbb{E}_\nu g_{p_0}(V), -\mathbb{E}_\nu g_{p_1}(V)\} > 0$. The two inequalities in (J.7), with d replacing d_p, again give minimax regret at least dT. Numerical scores are absent in this argument: the two-point latent construction need not have common numerical supports, and it supplies no selected-audit KL

bound. The continuous construction in Section 4.3 supplies that additional property when audits are allowed.

Consequently, comparison-only no-regret requires all inverse polynomials to lie on one positive ray. Fixing any representative yields an odd polynomial $P \not\equiv 0$ of degree at most m and positive numbers λ_p such that

$$g_p(v) = \frac{P(v)}{\lambda_p} \quad (v \in I_p), \qquad P(v_p(x)) = \lambda_p x \quad (|x| \le 1). \tag{J.9}$$

The polynomial is determined up to a single positive multiplier, which changes neither the ray condition nor the existence of a uniform positive lower bound on λ_p.

J.4 A COMMON STATISTIC GIVES AN IMPLEMENTABLE UPPER BOUND

Assume (J.9) and $\lambda_* := \inf_{p\in\mathcal{F}} \lambda_p > 0$. Since P is odd, write $P(v) = \sum_{j\le m,\ j \text{ odd}} b_j v^j$ and $H = \sum_j |b_j| > 0$. Use D_P from (J.3) with the coefficients of this fixed polynomial. This statistic and the update use the known family, not the identity of its unknown member. For every p,

$$|D_P/H| \le 1, \qquad \mathbb{E}_p[D_P/H \mid x] = \frac{\lambda_p}{H} x.$$

At $x = 1$, $\lambda_p = P(v_p(1)) \le H$ because $0 < a_p \le 1$. Thus $u'_t = (\lambda_p/H)u_t$ lies in $[-1, 1]^K$, and D_P/H is exactly unbiased for its half-gap. These scaled utilities remain predictable and have the same best fixed action as the original utilities. The scale is fixed for the whole run.

Use the one-channel update (C.3) with $A = 1$, $e = 0$, $D = D_P/H$, and $\eta = \sqrt{\log K/(KT)}$. Equivalently,

$$\widehat{L}_t(i) = \frac{\mathbf{1}\{I_t = i\}(1 - D_P/H)}{2(q_t(i) + \eta)}, \qquad w_{t+1}(i) = w_t(i)e^{-\eta \widehat{L}_t(i)}. \tag{J.10}$$

No admission gate is used. The zero-residual proof of Theorem C.1 applies directly, including its random hindsight comparator, and gives

$$\mathbb{E}_p R_T = \frac{H}{\lambda_p}\mathbb{E}_p R'_T \le 16\frac{H}{\lambda_p}\sqrt{KT\log K} \le 16\frac{H}{\lambda_*}\sqrt{KT\log K}. \tag{J.11}$$

To see the zero-audit specialization directly, put $\ell_t(i) = \mathbb{E}[(1 - D_P/H)/2 \mid I_t = i]$. Conditional on the entering history and current utilities,

$$\ell_t(i) = \frac{1}{2} - \frac{1}{4}u'_t(i) + \frac{1}{4}\langle q_t, u'_t\rangle.$$

The exponential-weights potential bound gives $\log K/\eta + (\eta/2)\sum_t \langle q_t, \widehat{L}_t^2\rangle$ for estimated losses. Lemma C.2 bounds the expected second moment by K, the mixture bias by $K\eta$, and the expected maximum cumulative comparator error by $\log K/(2\eta)$. Consequently the true loss regret is at most $3\log K/(2\eta) + (3/2)K\eta T$. Its numerical counterpart is four times this quantity in expectation; the realized two-proposal average adds a mean-zero martingale difference. Thus $\mathbb{E}R'_T \le 12\sqrt{KT\log K}$, which implies the stated constant 16 in (J.11) without an audit gate.

The algorithm does not require λ_p or even the value of λ_*; that lower bound is used only for its uniform performance guarantee. Selecting a representative inverse polynomial from the known family specifies P once, before observing the run.

Independent Rademacher utilities $(\xi_t, -\xi_t)$ give a complementary $\sqrt{T/2}$ lower bound for any nonempty family. The learner's expected payoff is zero even if shown every past utility, while the expected hindsight comparator is $\mathbb{E}|\sum_t \xi_t| \ge \sqrt{T/2}$. This proves the square-root case of Theorem 3.1.

J.5 WHY UNIFORM SIGNAL STRENGTH IS PART OF THE CLASSIFICATION

It remains to consider a common polynomial ray with $\inf_p \lambda_p = 0$. All members of a positive ray have the same orientation $e \in \{-1, 1\}$, so the endpoint identity is more generally $\lambda_p = P(ea_p)$. It

implies $\inf_p a_p = 0$. Indeed, if $a_p \geq \varepsilon > 0$ for all members, $P(e\varepsilon) > 0$ and monotonicity gives $\lambda_p \geq P(e\varepsilon)$. Conversely, $a_p \to 0$ implies $\lambda_p \to 0$ by continuity of P.

For each horizon choose a member with

$$0 < a_p \leq \frac{1}{8\sqrt{mT}}.$$

Keep this link fixed and take the two deterministic utility environments $(1, -1)$ and $(-1, 1)$, again duplicating the second type. A cross-type label has sign means a_p and $-a_p$ in the two worlds, up to their common orientation sign. For $a_p \leq 1/2$ its KL divergence is

$$a_p \log \frac{1 + a_p}{1 - a_p} \leq 4a_p^2.$$

The same bound holds after reversing presentation, and same-type labels contribute zero. The adaptive chain rule over at most mT labels gives full-history KL at most $4mTa_p^2 \leq 1/16$, hence total variation at most $1/4$. For the signed proposal average $s_t \in [-1, 1]$, prefix total variation and data processing give

$$\frac{\mathbb{E}_+ R_T + \mathbb{E}_- R_T}{2} = T - \frac{1}{2}\sum_t (\mathbb{E}_+ s_t - \mathbb{E}_- s_t) \geq \frac{3}{4}T.$$

The hard member may depend on T because the minimax supremum is uniform over the family; it remains fixed within each run. By contrast, (J.11) gives square-root regret for each fixed member of this same ray. This distinguishes pointwise learnability from a uniform class guarantee without imposing a sensitivity assumption silently.

The three obstructions exhaust all failures of (3.1): a member fails finite-session numerical recovery; all inverses are polynomials but two lie on different positive rays; or a common shape has scales approaching zero. Each gives a linear lower bound, and $R_T \leq 2T$ gives the corresponding upper bound. Together with Appendix J.4, this completes the proof of Theorem 3.1.

J.6 Audit complexity near a regular change of shape

The preceding classification identifies decision-relevant shape uncertainty. The next result connects it to the audit mechanism for a general smooth path, rather than choosing new coefficients for another example.

Proposition J.2 (A regular normalized shape variation)**.** *Fix $K \geq 2$. Let $g_\theta : [-b, b] \to [-1, 1]$, $0 < b < 1$, be odd increasing diffeomorphisms for θ in a neighborhood of zero. Assume $(\theta, v) \mapsto g_\theta(v)$ is C^2, $g_\theta(\pm b) = \pm 1$, and $\partial_v g_\theta > 0$. Suppose $\dot{g}_0 := \partial_\theta g_\theta|_{\theta=0} \not\equiv 0$. For the links $p_\theta(x) = [1 + g_\theta^{-1}(x)]/2$, write $V^{\mathrm{unk}}_{T,g,m}(r, B)$ for the value with $|\theta| \leq r$. For every fixed finite label cap m, there are $r_0, c > 0$ such that*

$$V^{\mathrm{unk}}_{T,g,m}(r, B) \geq c\left(\sqrt{T} + T\min\{r, (B+1)^{-1/2}\}\right), \qquad 0 \leq r \leq r_0, \quad 0 \leq B \leq T. \tag{J.12}$$

If, in addition, every g_θ is a polynomial of degree at most m, then this bound is sharp, and the known-calibration value with zero audits is $\Theta_K(\sqrt{T})$ uniformly on $|\theta| \leq r_0$. Constants may depend on the fixed path, cap, and K.

Proof. Write $g = g_0$, $h = \dot{g}_0$, and let σ be uniform probability measure on $[-b, b]$. Endpoint normalization gives $h(\pm b) = 0$, so the nonzero function h cannot be a scalar multiple of g. The projection residual

$$z = h - \frac{\int hg\, d\sigma}{\int g^2\, d\sigma}\, g$$

is nonzero and odd, and satisfies $\int zg\, d\sigma = 0$. Set $\epsilon = (2\|z\|_\infty)^{-1}$ and choose the fixed latent density

$$\nu(v) = \frac{1 + \epsilon z(v)}{2b}. \tag{J.13}$$

It is positive, integrates to one, and is C^1. Its induced mean $M(\theta) = \int g_\theta(v)\nu(v)\, dv$ satisfies

$$M(0) = 0, \qquad M'(0) = \epsilon \int z(v)^2\, d\sigma(v) > 0.$$

The first identity follows from oddness of g and orthogonality of z to g; for the second use oddness of h and $h = z + ag$. By continuity of M', there are $r_0 > 0$ and $a > 0$ with $M(\delta) \geq a\delta$ and $M(-\delta) \leq -a\delta$ for all $0 < \delta \leq r_0$.

For each fixed θ, draw independent $V_t \sim \nu$ and use utilities $(g_\theta(V_t), -g_\theta(V_t))$, duplicating one type for larger K. Holding ν fixed preserves every comparison transcript. The maps have the same numerical support. On a sufficiently small compact parameter interval, joint C^2 regularity and $\partial_v g_\theta > 0$ give all the finite constants in (I.4); ν has a positive minimum and bounded derivative. Equations (I.3)–(I.5) therefore give a transcript-uniform conditional KL bound $C_m\delta^2$, including same-type audits of absolute scores. Apply Proposition I.1 and the separate Rademacher lower bound. This proves (J.12), after reducing $r_0 \leq 1$ if necessary.

For the matching claim, polynomial coefficients are uniformly bounded on a compact parameter interval. One way to see this is to interpolate g_θ at $m + 1$ fixed points of $[-b, b]$: the inverse interpolation matrix is fixed and the values are bounded. Thus every known link has a bounded exact statistic (J.3). Use the coefficients of g_0 for the nominal statistic. If $x = g_\theta(v)$, its bias is

$$|\mathbb{E}_\theta[D_0 \mid x] - x| = |g_0(v) - g_\theta(v)| \leq r \sup_{|\eta| \leq r_0,\, |v| \leq b} |\partial_\eta g_\eta(v)|.$$

The upper-bound argument following Proposition I.1 now applies: choose between the bounded nominal update and the audit-only update before the run. It gives the matching rate uniformly in T, B, r, including $r = 0$ and $B = 0$. □

The polynomial requirement in the final assertion is exactly the known-link finite-cap requirement identified by Theorem 3.1. The regularity assumptions instead control the information in selected numerical scores. Every nonconstant normalized C^2 path has a point with nonzero parameter derivative, so the proposition applies locally after recentering there. At a flat parameterization the radius term can change: replacing θ by θ^3 in the original cubic family changes its uncertainty set to $|\theta_{\text{original}}| \leq r^3$, and its exact rate becomes $\Theta_K(\sqrt{T} + T \min\{r^3, (B+1)^{-1/2}\})$. The zero-audit classification is independent of this parameterization. It holds a family fixed as the horizon grows; shrinking radius classes can have intermediate rates, as the uniform formula in Theorem 3.2 shows.

J.7 Reading the criterion for a comparison family

A known link is the singleton-class case: its inverse must be an odd polynomial of degree at most m. For the logistic link $p(x) = (1 + e^{-x/\tau})^{-1}$, the inverse in sign-mean coordinates is $g(v) = 2\tau \operatorname{artanh}(v)$, which is not a finite-degree polynomial on its interval. Thus any fixed cap leaves linear worst-case numerical regret even when its temperature is given. The obstruction is the finite session, not unknown calibration.

A common nonlinear shape can nevertheless be learnable when only its positive scale is unknown. For example, if the family satisfies $P(v_p(x)) = \lambda_p x$ with $P(v) = v + v^3$ and $\lambda_p \in [1/4, 1/2]$, three labels implement $D_P = (W_1 + W_2 + W_3)/3 + W_1W_2W_3$; its expectation is $\lambda_p x$ and its absolute value is at most two. This illustrates the positive branch without estimating the scale. For $m \leq 2$, an odd polynomial can only be linear, recovering Proposition B.1 and exposing its uniform-sensitivity requirement when the class is allowed to approach an uninformative link.

For an increasing normalized family with the same endpoint $a_p = a > 0$ for every member, a common ray would satisfy $P(a) = \lambda_p$. Hence all scales would be equal and all links identical. Any two distinct normalized members with polynomial inverses therefore give calibration ambiguity. This applies to both constructed families in the main text. Theorem 3.1 decides comparison-only learnability; Theorems 3.2 and 4.1 further quantify the budget needed to resolve the ambiguity in their regular parameter families. A classification of audited rates for every smooth family would additionally have to account for the conditional score experiment, not only the comparison transcript.