WHEN DOES ADAPTIVE REORDERING HELP? MINIMAX LIMITS FOR COSTLY MODEL COMPARISON

Anonymous authors
Paper under double-blind review

ABSTRACT

Human labels turn model comparison into a sequential acquisition problem: judgments are acquired until the verdict between a candidate model and a deployed control is certified. The adjudicator can adapt both when to stop and what to label next. We study the value of the second choice for exact verdicts on a fixed cohort. Under Cartesian feasible support, additive utility and a monotone decision rule, adaptive reordering has zero deterministic minimax value: every deterministic adaptive policy can be forced to reveal every decision-relevant case, while a fixed width-sorted list reveals no irrelevant case. The result covers multivalued outcomes, multi-tier decisions and arbitrary positive label costs; computing the optimal value is weakly NP-hard even though an optimal policy is obtained by one sort. For accuracy comparison, the relevant cases are exactly the model disagreements. The boundary is sharp: hard cross-case coupling yields a $\Theta(n/\log n)$ fixed-to-adaptive gap for bounded outcome alphabets, while priors or randomization can also make reordering useful. On 108,715 fully judged TREC-COVID run–topic comparisons, width order certifies after a median of 60% of the judgments, versus 79% for random order and 43% for a realized-path oracle. On a frozen 2,000-instance subset, the fitted-prior adaptive policy closes 20.4% of the realized width-to-oracle gap; under the same prior, reordering itself saves 0.432 judgments in expected cost relative to the optimal fixed order.

1 INTRODUCTION

A candidate model is ready to replace the deployed control. The decision concerns this realized pair on a fixed evaluation cohort, yet some relevance judgments, audit outcomes or expert labels needed to compare them are still unavailable. Human labels are the bottleneck of modern evaluation (Dorner et al., 2025), so labels are acquired sequentially and adjudication stops once the comparison is settled. A sequential adjudicator can adapt in two ways: *when it stops*, and *what it labels next*. Adaptive stopping obviously helps. It is equally natural to expect that choosing the next label in light of earlier labels—as active testing and active inference do for statistical estimation (Kossen et al., 2021; Zrnic & Candès, 2024)—should help as well. Does it?

It helps to separate three quantities. A *guaranteed* advantage lowers worst-case cost relative to every fixed order. A *hindsight* advantage appears when the realized labels admit a short certificate after the fact. An *exploitable* advantage is the part a policy can predict before buying those labels. The theory settles the first question; the experiments measure the latter two.

Our setting is a fixed cohort with missing labels. Each unresolved case induces a finite set of candidate-minus-control contributions, and a verdict is certified when every feasible completion gives the same decision. For additive retrieval measures, the resulting interval is the signed MED of Tan & Clarke (2015); here the question is which label to acquire next. The key object is *relevance*: a case is relevant if changing its outcome alone can change the verdict. Under Cartesian support and a monotone additive decision, an adversary can keep every unrevealed relevant case relevant, while relevance itself is monotone in uncertainty width. The minimax cost is therefore exactly the cost of the relevant cases, and a width-sorted list attains it with one sort even though computing that value is weakly NP-hard.

The intuition is an absence of *information externalities*. Under Cartesian feasible support—which constrains only which outcomes are logically possible, not how likely they are—observing one case

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.

shifts the aggregate but removes no feasible outcome of another case, so it cannot make another relevant case unnecessary. Hard coupling is exactly what creates externalities: once a revealed outcome can exclude outcomes elsewhere, adaptivity can become near-maximally valuable, and we construct additive sign certification problems on which every fixed order needs $\Omega(n/\log n)$ times more labels than an adaptive tree, which we show is order-wise tight for bounded outcome alphabets.

Relation to classical results. For binary outcomes and unit costs, the product-support equality follows from classical evasiveness of threshold functions together with the desirability order of weighted thresholds. Widest-first rules for SUM (Khanna & Tan, 2001) are another precursor, but there the stopping condition depends only on the unrevealed widths. Signed certification is outcome-dependent: the same residual width may certify after one revealed value and remain unresolved after another. The multivalued case also cannot be encoded as independent binary increments without changing both the support and the reveal operation. Our contribution is the combined extension to heterogeneous multivalued supports, arbitrary positive costs and monotone tiers, together with the sharp boundary under coupling.

Our contributions are:

- **Exact minimax characterization under Cartesian support.** For finite heterogeneous multivalued contribution sets, arbitrary positive reveal costs and monotone multi-tier decisions, the deterministic adaptive and fixed-order minimax costs both equal $c(\mathcal{R})$, the total cost of the decision-relevant cases, and every width-sorted list attains it for all margins and tiers at once, although the optimal value is weakly NP-hard to compute (Theorem 1). For two classifiers compared by accuracy, the relevant cases are exactly the disagreements (Corollary 1).
- **A sharp boundary.** Known cross-case coupling raises the largest possible ratio $\mathrm{OPT}_{\mathrm{fix}} / \mathrm{OPT}_{\mathrm{ad}}$ from exactly one to $\Theta(n/\log n)$ for bounded outcome alphabets; non-monotone or non-additive decisions also break the equality, a prior restores an expected-cost advantage, and randomized adaptive policies can strictly beat every randomization over fixed orders, each already on three coordinates (Propositions 3 and 4). Cross-case information externalities, prior information and randomization, not sequential observation itself, are the sources of reordering value.
- **Guaranteed, hindsight and exploitable value on human judgments.** On 108,715 fully judged TREC-COVID comparisons, width order reaches the exact verdict after a median of 60% of the judgments, against 79% for random order and 43% for a realized-path oracle, and the ordering persists under indifference margins. The hindsight gap is substantial, but on small instances an exact Bayes-optimal adaptive policy under a cross-topic prior recovers only 20.4% of it on average and none at the median.

2 Problem Setup and Identification

Fixed comparison. Consider a fixed control f^0 and a fixed candidate f^1 evaluated on n cases from a fixed cohort. For case i, let $\mathcal{C}_i$ be the set of outcomes (e.g. relevance labels or positive identities) compatible with current information, and $\delta_i(c) = u_i^1(c) - u_i^0(c)$ the candidate-minus-control utility if outcome c is true. The feasible contribution set is $\mathcal{A}_i = \{\delta_i(c) : c \in \mathcal{C}_i\}$, the realized but unrevealed contribution is $d_i \in \mathcal{A}_i$, and the aggregate effect is $\Delta(d) = \frac{1}{n}\sum_{i=1}^n d_i$. Everything below depends only on the finite sets $\mathcal{A}_i$.

Sharp bounds and certificates. Let $m_i = \min \mathcal{A}_i$, $M_i = \max \mathcal{A}_i$ and $w_i = M_i - m_i$. We first assume contribution-level product support $\Gamma = \prod_{i=1}^n \mathcal{A}_i$: every combination of coordinatewise feasible contributions is itself feasible. This is a statement about logical support. Quotas, transitivity or hierarchical consistency can violate it by making one judgment rule out outcomes elsewhere.

Proposition 1 (Sharp aggregate bounds)**.** *Under product support,* $\min_{d\in\Gamma} \Delta(d) = L_0 := \frac{1}{n}\sum_i m_i$ *and* $\max_{d\in\Gamma} \Delta(d) = U_0 := \frac{1}{n}\sum_i M_i$*, and both endpoints are attained.*

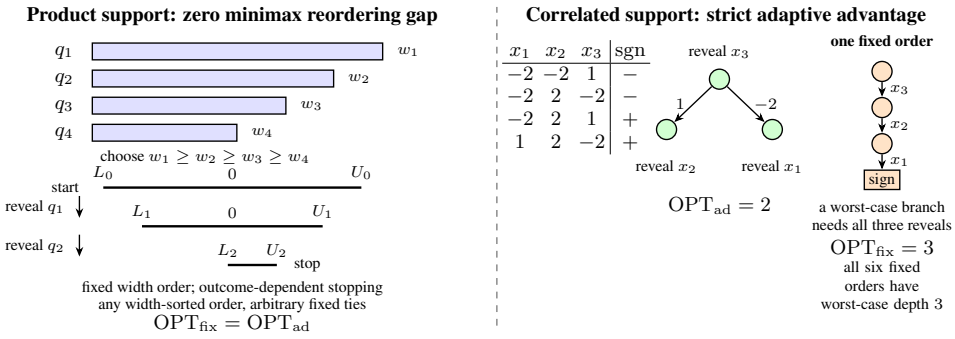


Figure 1: Left: under product support and a monotone link, every nonincreasing-width fixed order attains the fully adaptive minimax value; the intervals show one realized stopping path. Right: in the correlated four-world support of Proposition 2, adaptive reordering reduces the worst-case reveal count from three to two. Proposition 3 amplifies this to a near-linear gap, order-wise tight for bounded outcome alphabets.

The certificate map is

$$\mathrm{Cert}(L,U) = \begin{cases} \textsc{Beneficial}, & L > 0, \\ \textsc{Harmful}, & U < 0, \\ \textsc{Null}, & L = U = 0, \\ \textsc{Unresolved}, & \text{otherwise.} \end{cases} \tag{1}$$

An unresolved interval records non-identification, not sampling uncertainty. A deployment margin τ is handled by applying the same map to $[L-\tau, U-\tau]$, which leaves every width unchanged; we write $\lambda(d) = \mathrm{sgn}(\sum_i d_i - n\tau) \in \{-1, 0, +1\}$ for the certificate class of a world d. If the true joint support Γ' is smaller than Γ, the attainable extrema over Γ' consistent with any history are no wider than those over Γ, so every Cartesian certificate remains valid; sharpness and ordering optimality, however, are relative to the envelope.

We evaluate acquisition cost in the deterministic worst case, which isolates what can be guaranteed from the revealed labels alone. Priors and randomization are revisited in Section 3.1, and the fitted-prior experiments appear in Section 4.1.

Exact reveals and policy cost. A reveal exactly adjudicates one case. After revealing $S \subseteq [n]$ with values d_S,

$$L(S, d_S) = \tfrac{1}{n}\big(\textstyle\sum_{i\in S} d_i + \sum_{i\notin S} m_i\big), \qquad U(S, d_S) = \tfrac{1}{n}\big(\textstyle\sum_{i\in S} d_i + \sum_{i\notin S} M_i\big), \tag{2}$$

so revealing i updates $L \leftarrow L + (d_i - m_i)/n$, $U \leftarrow U - (M_i - d_i)/n$ and removes exactly w_i/n of width, independent of the outcome. On a general support Γ, a history certifies once all worlds of Γ consistent with it share the class λ; on a product support this coincides with the interval test, because the endpoints are attained. A deterministic adaptive policy π maps histories to a next case or a stop. With $S_\pi(d)$ the set of cases revealed before stopping on world d and $C_\pi(d) = |S_\pi(d)|$,

$$\mathrm{OPT}_{\mathrm{ad}} = \min_{\pi} \max_{d\in\Gamma} C_\pi(d), \qquad \mathrm{OPT}_{\mathrm{fix}} = \min_{\sigma} \max_{d\in\Gamma} C_\sigma(d), \tag{3}$$

where σ ranges over fixed permutations that still stop adaptively at the first certificate. Policies may stop only at certified histories, and we first consider deterministic policies; randomized policies are treated in Section 3.1 and Appendix E. The adaptive class permits both outcome-dependent stopping and outcome-dependent reordering; the fixed class permits only the former. Selective adjudication (pooled assessment, expert review, audit) makes the order a decision; if outcomes mature exogenously, the same certificates apply but the order is not controllable.

3 Minimax Acquisition and Its Boundary

Coordinate i is *relevant* on Γ if there exist $d, d' \in \Gamma$ that differ only in coordinate i and satisfy $\lambda(d) \neq \lambda(d')$; let $\mathcal{R} \subseteq [n]$ be the set of relevant coordinates. Let $\mathfrak{S}_w$ be the set of permutations sorted by

nonincreasing width, with arbitrary tie-breaking. Throughout this section, $\lambda(d) = g(\sum_i d_i)$ for a monotone (nondecreasing or nonincreasing) link $g : \mathbb{R} \to \mathcal{Y}$ into a totally ordered set of decision classes. The three-way sign certificate with margin τ is the case $g(x) = \mathrm{sgn}(x - n\tau)$; multi-tier decisions such as harmful / negligible / beneficial with cut points $\pm n\tau$, $g(x) = \mathrm{sgn}(x - n\tau) + \mathrm{sgn}(x + n\tau)$, are also covered.

Lemma 1 (Relevance is monotone in width)**.** *Let $\Gamma = \prod_i \mathcal{A}_i$ and g be monotone. If coordinate j is relevant and $w_i \geq w_j$, then coordinate i is relevant. In particular, zero-width coordinates are irrelevant, equal-width coordinates are both relevant or both irrelevant, and $\mathcal{R}$ is the prefix of every $\sigma \in \mathfrak{S}_w$.*

Proof. Assume g nondecreasing (otherwise reverse the order on $\mathcal{Y}$). Relevance of j yields values of the other coordinates, with sum b over $[n] \setminus \{i, j\}$ and some $a_i \in \mathcal{A}_i$, such that $g(b + a_i + m_j) < g(b + a_i + M_j)$. Let $p = b + m_j + m_i$, $q = b + m_j + M_i$, $r = b + M_j + m_i$ and $t = b + M_j + M_i$; all four sums are attained in Γ by product support. Because $w_j \leq w_i$, $r \leq q$; also $p \leq b + a_i + m_j$ and $b + a_i + M_j \leq t$. If $g(p) \neq g(q)$, the two worlds with $d_j = m_j$ that differ only in coordinate i show that i is relevant. Otherwise $g(r) \leq g(q) = g(p) \leq g(b + a_i + m_j) < g(b + a_i + M_j) \leq g(t)$, and the two worlds with $d_j = M_j$ that differ only in coordinate i show that i is relevant. □

Lemma 2 (An adversary preserves relevance)**.** *Let $\Gamma = \prod_k \mathcal{A}_k$ and g be monotone, and fix a coordinate i. There is a value $a \in \mathcal{A}_i$ such that, on the residual product $\prod_{k \neq i} \mathcal{A}_k$ with $d_i = a$, the relevant set is exactly $\mathcal{R} \setminus \{i\}$.*

Proof. Fixing a coordinate cannot create relevance: a residual witness pair is also a witness pair in Γ. If $\mathcal{R} \setminus \{i\} = \emptyset$, any a works. If $i \notin \mathcal{R}$, any a works: moving both worlds of a witness pair for $j \in \mathcal{R}$ to $d_i = a$ changes neither class, since each move changes only the irrelevant coordinate i. Otherwise let j be a narrowest coordinate of $\mathcal{R} \setminus \{i\}$, take a witness pair for j, and let a be its common value in coordinate i. Then j is relevant for the residual label $x \mapsto g(a + x)$, which is again a monotone function of an additive sum on a product support. By Lemma 1 applied to the residual problem, every $k \in \mathcal{R} \setminus \{i\}$, having $w_k \geq w_j$, is relevant there. □

Theorem 1 (Exact minimax cost and universal fixed ordering)**.** *Assume (i) $\Gamma = \prod_i \mathcal{A}_i$; (ii) $\lambda(d) = g(\sum_i d_i)$ with g monotone; and (iii) exact reveals. Then (a) for every deterministic adaptive policy there is a world $d \in \Gamma$ on which the policy reveals every coordinate of $\mathcal{R}$ before stopping; and (b) every fixed order $\sigma \in \mathfrak{S}_w$ with adaptive stopping certifies every world without revealing any coordinate outside $\mathcal{R}$. Consequently, for any positive reveal costs c_i, the optimal worst-case cost of adaptive policies, that of fixed orders, and $\max_d \sum_{i \in S_\sigma(d)} c_i$ for every $\sigma \in \mathfrak{S}_w$ all equal $c(\mathcal{R}) = \sum_{i \in \mathcal{R}} c_i$. With unit costs, $\max_{d \in \Gamma} C_\sigma(d) = \mathrm{OPT}_{\mathrm{fix}} = \mathrm{OPT}_{\mathrm{ad}} = |\mathcal{R}|$.*

Proof. (a) Answer each query by Lemma 2. By induction, after any history the residual relevant set is $\mathcal{R}$ minus the revealed coordinates. While it is nonempty, a residual witness pair shows that the residual support is not class-homogeneous, so the policy cannot stop; hence the resulting path reveals all of $\mathcal{R}$. Completing the unrevealed coordinates arbitrarily gives a world on which the policy follows this path. (b) Two worlds that agree on $\mathcal{R}$ are connected by changing the other coordinates one at a time inside Γ, and no such step changes λ. So once all of $\mathcal{R}$ is revealed the residual support is class-homogeneous and σ stops. By Lemma 1, $\mathcal{R}$ is a prefix of σ, so σ never reaches a coordinate outside $\mathcal{R}$. Writing $\mathrm{OPT}^c_{\mathrm{ad}}, \mathrm{OPT}^c_{\mathrm{fix}}$ for the worst-case cost optima, and since fixed orders are adaptive policies, (a) gives $c(\mathcal{R}) \leq \mathrm{OPT}^c_{\mathrm{ad}} \leq \mathrm{OPT}^c_{\mathrm{fix}} \leq \max_d \sum_{i \in S_\sigma(d)} c_i$ for every $\sigma \in \mathfrak{S}_w$, and (b) bounds the right-hand side by $c(\mathcal{R})$. □

Interpretation. Theorem 1 fixes the order, not the stopping rule: different label realizations still stop at different depths. Many fixed orders are minimax-optimal, since any order that places $\mathcal{R}$ first attains $c(\mathcal{R})$. Width order is useful because it gets this guarantee without computing $\mathcal{R}$, works for every margin and decision tier, and never reaches an irrelevant case. On the TREC-COVID instances checked exactly, every nonzero-weight document is relevant, so realized-path efficiency—rather than worst-case optimality—is what distinguishes the practical orders in Section 4.1. Appendices A–C give an independent constructive proof for the unit-cost case.

The width list itself is independent of g; only the stopping rule changes. A single adjudication order is therefore minimax-optimal across margins and tiers, echoing universal non-adaptive lists in stochastic score classification (Ghuge et al., 2025).

Corollary 1 (Comparing two classifiers)**.** *Let f^0, f^1 be frozen classifiers evaluated by 0–1 accuracy on a finite pool whose labels are unknown, carry no hard cross-example constraints, and for which both predicted labels are feasible on every example, with any margin τ and any positive label costs. Let D be the set of examples on which the predictions differ. If the verdict $\operatorname{sgn}(\mathrm{Acc}(f^1) - \mathrm{Acc}(f^0) - \tau)$ is not determined before labeling, then every example in D is relevant and no other example is. Hence every adaptive labeling policy has a label realization on which it labels all of D, and labeling D in any order is minimax optimal.*

The proof is in Appendix J. Appendix J gives an LLM multiple-choice replay illustrating Corollary 1: restricting acquisition to model disagreements preserves every exact verdict.

Relation to evasiveness. Part (a) is evasiveness of monotone functions of an additive sum on their relevant coordinates, and Lemma 1 is a multivalued desirability order; in the binary case both are classical (Section 1). Knowing that $\mathcal{R}$ is a width prefix matters because computing $\mathcal{R}$ is hard. With $\mathcal{A}_k = \{0, a_k\}$ for positive integers $a_1, \ldots, a_K$, an extra coordinate $\mathcal{A}_j = \{0, 1\}$ and $g(x) = \operatorname{sgn}(x - T - \frac{1}{2})$, coordinate j is relevant iff some subset of the a_k sums to T; since j is narrowest, Lemma 1 gives $\mathrm{OPT}_{\mathrm{ad}} = K + 1$ iff the SUBSET SUM instance is a yes-instance. The optimal value is therefore (weakly) NP-hard to compute, yet the optimal policy needs only one sort and $O(1)$ arithmetic per reveal: on a product support the residual is class-homogeneous iff g takes the same value at the two residual endpoints, which are maintained incrementally.

3.1 WHERE ADAPTIVITY COMES FROM: COUPLING, NON-MONOTONICITY AND PRIORS

On a general support $\Gamma \subseteq \prod_i \mathcal{A}_i$, a history certifies once all consistent worlds share the class λ.

Lemma 3 (Fixed orders and determining sets)**.** *For any nonempty $\Gamma \subseteq \prod_i \mathcal{A}_i$ and any fixed order σ with adaptive stopping, $\max_{d \in \Gamma} C_\sigma(d)$ is the smallest k such that the first k coordinates of σ determine λ on Γ. Consequently $\mathrm{OPT}_{\mathrm{fix}}(\Gamma)$ is the minimum size of a set S such that $d_S = d'_S$ implies $\lambda(d) = \lambda(d')$ for all $d, d' \in \Gamma$.*

Proof. Let $K = \max_d C_\sigma(d)$ and let k^* be the smallest k such that the first k coordinates of σ determine λ. After k^* reveals every residual support is class-homogeneous, so every world stops by depth k^* and $K \le k^*$. Conversely, every world has stopped by depth K, and its residual support after K reveals is contained in its class-homogeneous residual support at stopping; hence worlds that agree on the first K coordinates share λ, and $k^* \le K$. The second claim follows by placing any set S first. □

Adaptive stopping therefore does not change the worst-case depth of a fixed order: $\mathrm{OPT}_{\mathrm{fix}}$ is the complexity of an optimal oblivious decision tree, which equals the minimum reduct complexity of the decision table (Ostonov & Moshkov, 2025), and on general supports the fixed-versus-adaptive comparison is the regular-versus-oblivious comparison of decision-tree theory. The smallest correlated witness has three coordinates.

Proposition 2 (Strict gap under correlated support)**.** *Let Γ be certified by the sign of the coordinate sum, with*

$$\Gamma = \{(-2, -2, 1),\ (-2, 2, -2),\ (-2, 2, 1),\ (1, 2, -2)\}.$$

Then $\mathrm{OPT}_{\mathrm{ad}} = 2$ and $\mathrm{OPT}_{\mathrm{fix}} = 3$.

Reveal x_3; then reveal x_2 if $x_3 = 1$ and x_1 if $x_3 = -2$. Every fixed pair leaves two opposite-sign worlds indistinguishable (Appendix D), so the adaptive depth is two while every fixed order has depth three. The gain comes from an information externality: observing x_3 rules out outcomes elsewhere.

Proposition 3 (Each assumption is essential; the coupled gap is maximal for bounded outcome sets)**.** *(a)* Additive sign without product support. *For every $d \ge 1$ there is a support Γ_d on $n_d = 2^d - 1$ coordinates, with $|\Gamma_d| = 2^d$ and $|\mathcal{A}_i| \le 3$, certified by the sign of the coordinate sum, such that $\mathrm{OPT}_{\mathrm{ad}}(\Gamma_d) \le d$ and $\mathrm{OPT}_{\mathrm{fix}}(\Gamma_d) \ge 2^{d-1}$; hence $\mathrm{OPT}_{\mathrm{fix}} / \mathrm{OPT}_{\mathrm{ad}} \ge (n_d + 1)/(2\log_2(n_d +$*

1)). *(b)* Maximality. *For every* Γ *with* $|\mathcal{A}_i| \le k$ *(*$k \ge 2$*),* $\mathrm{OPT}_{\mathrm{fix}} \le (k^{\mathrm{OPT}_{\mathrm{ad}}} - 1)/(k-1)$. *Hence* $\mathrm{OPT}_{\mathrm{fix}} / \mathrm{OPT}_{\mathrm{ad}} = O(n \log k / \log n)$*, so for bounded* k *(in particular* $k = 3$ *as in (a)) the separation in (a) is tight up to a constant factor. (c)* Product support and additive sum, non-monotone link. *On* $\mathcal{A}_1 = \{0,1\}$*,* $\mathcal{A}_2 = \{0,2\}$*,* $\mathcal{A}_3 = \{0,3\}$ *with* $g(x) = \mathbf{1}[x \in \{3,5,6\}]$*, all three coordinates are relevant,* $\mathrm{OPT}_{\mathrm{ad}} = 2$ *and* $\mathrm{OPT}_{\mathrm{fix}} = 3$*. (d)* Product support without additivity. *On* $\{0,1\}^{k+2^k}$*, the address function has adaptive depth* $k+1$*, while every fixed order needs all* $k + 2^k$ *coordinates.*

The upper bound in (b) ranges over all finite supports with alphabet size at most k, while the lower bound already holds for additive sign certification with $k = 3$. Part (a) realizes the regular-versus-oblivious decision-tree separation inside an additive sign problem; part (b) shows that its order is tight for bounded alphabets. Part (c) isolates monotonicity: even on a product support, a non-monotone link gives a strict gap. Proofs are in Appendix G.

A prior can create a gain even without coupling. With independent $X_1 \sim \mathrm{Bern}(0.8)$, $X_2 \sim \mathrm{Bern}(0.9)$ and $X_3 \sim \mathrm{Bern}(0.4)$, unit costs and majority threshold two, every policy needs three reveals in the worst case. In expectation, however, revealing X_1 first and then choosing X_2 or X_3 according to its value costs 2.16, versus 2.26 for the best fixed order. Randomization creates another boundary: for majority on three bits, a uniformly random pair costs at most $8/3$ even though every deterministic policy has worst-case depth three. Proposition 4 shows that randomized adaptive policies can improve further over randomized fixed orders.

Proposition 4 (Randomized reordering can pay on a product support)**.** *Let* $\mathcal{A}_1 = \{-2,2\}$*,* $\mathcal{A}_2 = \{-1,0\}$*,* $\mathcal{A}_3 = \{0,2\}$*,* $\Gamma = \prod_i \mathcal{A}_i$ *and* $\lambda(d) = \mathrm{sgn}(\sum_i d_i - \frac{3}{2})$*. Every deterministic policy needs* 3 *reveals in the worst case. The minimax expected cost of distributions over fixed orders with adaptive stopping is* $8/3$*, and that of distributions over adaptive policies is* $5/2$*.*

The class is $+$ iff $d_1 = 2$ and ($d_2 = 0$ or $d_3 = 2$). Mixing "d_1, then d_2, then d_3" with "d_3; then d_1 if $d_3 = 2$, else d_2 and then d_1" in equal proportions costs at most $5/2$ on every world, and the second policy is not a fixed order. Appendix E gives matching Yao lower bounds with exact rational weights. So the zero-gap result is specific to deterministic policies. The value of adaptive evaluation thus comes from cross-case structure, a prior or randomization, not from sequential observation alone.

4 Experiments

The theoretical claims are proved in Section 3. Exact rational-arithmetic solvers provide regression checks on exhaustive and random multivalued instances, random costs and multi-tier links (Appendix F). The experiments ask two practical questions: how early a genuine judgment stream certifies the verdict, and how much of the hindsight gap can be captured before the labels are known.

4.1 Natural incomplete judgments: TREC-COVID

We run a retrospective adjudication replay over genuine human relevance judgments from TREC-COVID Round 1 (Voorhees et al., 2020), pooled to depth 7. The judgments and participant runs required for this replay are publicly downloadable. For every pair of runs and every topic, the cases are the documents in the union of the two top-7 lists. A document counts as relevant if judged partially or fully relevant (qrel ≥ 1); counting only fully relevant documents changes no conclusion (Appendix I). With truncated rank-biased precision with persistence 0.8 (Moffat & Zobel, 2008), document i contributes $c_i y_i$ to the candidate-minus-control difference, where c_i is its RBP weight difference; hence $\mathcal{A}_i = \{0, c_i\}$, one judgment reveals one contribution exactly, and we evaluate on the Cartesian feasible-support envelope $\prod_i \{0, c_i\}$. For exact replay we keep the 108,715 of 304,590 run-pair–topic candidates (35.7%) whose union is fully judged; the interval coincides with the signed MED on every instance.

Observed certification depth. Width order reaches the exact verdict with a median of 40% of the union left unjudged (Table 1). Depth is governed by the effect size: every instance requiring all judgments has $|\Delta| \le 0.066$, and more than half of them have $\Delta = 0$. On the 402 instances checked by exact dynamic programming, the minimax depth equals the number of nonzero contributions, so

Quantity	Value
Median fraction judged: width order / random / realized oracle	0.60 / 0.79 / 0.43
Mean depth: width / rank priority / Moffat A / B / C / random	8.20 / 8.46 / 8.55 / 8.54 / 8.71 / 10.15
Certified within 25% / 50% / 75% of judgments (width order)	2.1% / 41.5% / 77.2%
Width order requires every judgment	9.46%
Initially unresolved (no judgments)	99.74%

Table 1: TREC-COVID Round 1, conditional on the 108,715 fully judged run-pair–topic instances, RBP ($p = 0.8$). Fractions are of the union size n. The oracle is a realized-path lower bound, not a policy: the fewest judgments that certify the realized world, computed exactly by revealing documents in decreasing order of $d_i - m_i$ (for a beneficial verdict) or $M_i - d_i$ (harmful). Moffat A/B/C are the targeted-judgment orders of Moffat et al. (2007); all orders share the same exact certificate. Random orders only nonzero-weight documents (over all union documents: median 0.82, mean 10.27).

every weighted document is relevant; all orders over them tie in the worst case, and what separates them is realized-path efficiency. Under RBP, $|c_i|$ is nearly determined by rank, so width order coincides with best-rank order on the 54% of instances without overlap, and targeted-judgment orders perform comparably.

Prior-guided reordering. Width order still leaves a sizeable hindsight gap to the realized oracle (median 0.60 versus 0.43; Appendix H). On the full 108,715-instance replay, a leave-one-topic-out prior barely changes simple policies: mean depth is 8.197 for width order, 8.208 for a static prior order and 8.200 for an adaptive greedy, against 5.959 for the oracle. Exact expected-cost optimization is feasible on a frozen 2,000-instance subset with at most 10 documents. On that same subset, width and oracle mean depths are 5.857 and 4.580; replaying the fitted-prior optimal adaptive policy on the realized labels closes 20.4% of this gap on means and none on medians. Separating prior information from reordering, the optimal adaptive policy saves 0.432 judgments in expected cost relative to the optimal fixed order under the same prior. A prior-strength sweep on 31,655 tractable cases shows a non-monotone pattern: the adaptive-over-fixed gain grows at intermediate information and falls to zero near perfect information, where a fixed order can already follow the oracle path (Appendix H).

Margins, graded judgments and incomplete pools. The ordering is stable under symmetric indifference bands $\tau \in \{0, 0.02, 0.10\}$, where width order needs median fractions 0.60, 0.64 and 0.75. Keeping all three TREC relevance grades likewise raises certification depth only modestly and leaves the ordering of oracle, width, rank-priority, Moffat and random policies unchanged on all 108,715 retained instances (Appendix I). Among incompletely judged unions, treating unjudged documents as permanently unknown still certifies 46.5% of excluded instances (Appendix H). Finally, a three-LLM replay on MMLU and ARC-Challenge confirms Corollary 1: labeling only disagreements gives the same verdict as labeling the whole pool in all 720 replays, with 4.5–32 times fewer labels.

Selection effects. The fully judged filter favors smaller unions: retention falls from 0.86 at union size 7 to 0.28 at size 14, and 1,704 of 10,153 run pairs contribute no retained instance. Within matched (topic, union size, overlap) strata, however, retained and excluded instances are close on every pre-label feature we measured (standardized mean differences ≤ 0.03). We therefore interpret the replay as evidence for judged-complete top-7 unions rather than as a prospective estimate for all TREC-COVID comparisons.

5 Related Work

Closest formal results. Evasiveness (Rivest & Vuillemin, 1976; Buhrman & de Wolf, 2002; Kahn et al., 1984; Björner & Kalai, 1988) and the desirability order of weighted threshold functions (Isbell, 1958; Taylor & Zwicker, 1999) give the binary unit-cost core of Theorem 1. The present result extends that structure to heterogeneous multivalued supports, arbitrary positive costs and monotone multi-tier links, and adds the coupled-support boundary. In the binary case, recognizing a relevant coordinate is the familiar dummy-player problem in weighted voting games (Elkind et al., 2009).

Setting	Information	Objective	Fixed versus adaptive order
Threshold evasiveness, desirability order	binary inputs, known weights	worst-case queries	binary unit-cost case of Thm. 1
Regular vs. oblivious decision trees	arbitrary decision tables	worst-case depth	general separations; no additive-sign characterization
Widest-first SUM (Khanna–Tan)	intervals, exact probes	numerical precision	stopping independent of values; no reordering gain
SBFE / stochastic score classification	known (usually independent) priors	expected query cost	adaptivity gap $O(1)$ to $\Omega(n/\log n)$
Active statistical evaluation	predictors, probabilities	estimation, testing	adaptive selection can reduce variance
This paper	finite feasible sets, no prior	exact verdict, worst-case cost	ratio exactly 1 on Cartesian supports (multivalued, any costs, monotone tiers); up to $\Theta(n/\log n)$ under coupling

Table 2: Closest formal settings. "Fixed order" permits outcome-dependent stopping. Citations are in the text of this section.

On general supports, the fixed-order optimum is the complexity of an optimal oblivious decision tree, equal to minimum reduct complexity (Ostonov & Moshkov, 2025); Proposition 3 realizes the near-linear regular-versus-oblivious separation inside additive sign certification. Table 2 places these settings side by side.

Computing with uncertainty. The interval-query model dates to Kahan (1991). Olston & Widom (2000) show that choosing an optimal oblivious query set for SUM to a prescribed precision has knapsack structure. In the unit-cost SUM setting of Khanna & Tan (2001), where widest-first is optimal, the stopping condition depends only on which items have been probed, so the observed values do not change the usefulness of reordering. In signed certification, stopping depends on revealed values: with $\mathcal{A}_1 = \{0, 2\}$, $\mathcal{A}_2 = \{-1, 1\}$ and the sign decision, revealing coordinate 1 certifies after $d_1 = 2$ but not after $d_1 = 0$, although the residual width is 2 in both cases; this is what makes the fixed–adaptive equality nontrivial under product support and what fails under correlation. Feder et al. (2003) study interval-uncertain selection, where adaptivity can retain value. The priced-information model of Charikar et al. (2002) is adversarial and competitive against the cheapest proof of the realized input, which is exactly our realized-path oracle. Explorable-uncertainty work mostly uses competitive analysis, under which even two unknown values force ratio n (Megow & Schlöter, 2023); in our setting every deterministic policy is n-competitive against this oracle ($\mathcal{A}_i = \{0, 1\}$, threshold $\frac{1}{2}$, adversary answering 0). Megow & Schlöter (2023) and Dürr et al. (2024) study width-based and mandatory queries for selection, without comparing adaptive and fixed orders.

Incomplete relevance judgments. For additive measures, $[L, U]$ is the signed minimum and maximum effectiveness difference over relevance assignments consistent with the judgments so far; $\max(|L|, |U|)$ is the MED of Tan & Clarke (2015). Minimal Test Collections combines a worst-case stopping rule with probabilistic document selection (Carterette et al., 2006); targeted-judgment, RBP-residual, move-to-front, bandit and active-sampling methods provide several ways to choose judgments (Moffat et al., 2007; Moffat & Zobel, 2008; Cormack et al., 1998; Losada et al., 2017; Li & Kanoulas, 2017). Our contribution is the fixed-versus-adaptive acquisition-order question over this uncertainty set. Concurrently, Kuboyama (2026) study exact finite-pool label certificates and budgets for selective classifiers; their focus is label complexity rather than adaptive ordering.

Stochastic testing, partial identification and monitoring. SBFE and stochastic score-classification query hidden inputs under known distributions and minimize expected cost (Deshpande et al., 2016; Gkenosis et al., 2018; Hellerstein et al., 2022; Ghuge et al., 2025; Nielsen et al., 2025; Ünlüyurt, 2004; 2025); correlated stochastic certification extends this view to joint distributions (Ghuge et al., 2026). These results optimize expected cost under a distribution, whereas Theorem 1 is a worst-case statement over feasible support. Decision-region identification, partial identification and targeted sequential experiments provide complementary Bayesian, approximation

and bound-tightening views (Cicalese et al., 2014; Javdani et al., 2014; Guillory & Bilmes, 2010; Yuan & Tang, 2023; Manski, 2003; Ailer et al., 2024). Active testing, active inference and deployment monitoring address statistical estimation, where adaptive selection can reduce variance (Kossen et al., 2021; Zrnic & Candès, 2024; Angelopoulos et al., 2025; Hara et al., 2024; Dorner & Hardt, 2024; Ginart et al., 2022; Koebler et al., 2025; Tonekaboni et al., 2022). Dorner et al. (2025) give a different evaluation limit, bounding how much an LLM judge can reduce ground-truth labels.

6 Discussion and Conclusion

Under Cartesian support and a monotone additive decision, outcome-dependent stopping remains useful but deterministic outcome-dependent reordering cannot improve worst-case cost. Width order reaches the minimax optimum without computing the relevant set, and the same list works across margins and decision tiers.

The boundary cases explain when more elaborate policies are worth the effort. Cross-case coupling can create near-linear gains, while a predictive prior or randomization can create smaller gains even on product support. The TREC-COVID replay shows the same split empirically: width order is strong without a prior, the fitted prior yields a modest adaptive-over-fixed expected-cost improvement, and that improvement disappears again as the prior approaches perfect information. Transitive preferences, hierarchical labels and known-count constraints are natural settings where stronger coupling may make adaptive reordering more valuable. Extending the analysis to noisy or batched judgments is another open direction.

AI use statement

We used large language models as general-purpose assistants for writing and editing the manuscript, for literature search, and as coding assistants for the verification scripts. All theoretical results, proofs and reported numbers were checked by the authors, who take full responsibility for the content of this paper.

Reproducibility statement

Complete proofs not given in the main text are in Appendices A–E, G, and J. The exact solvers and enumeration protocols are described in Section 4 and Appendix F; data protocols for TREC-COVID and the classifier replay are in Section 4 and Appendices H–J. All datasets used in the empirical replay are public.

References

Elisabeth Ailer, Niclas Dern, Jason Hartford, and Niki Kilbertus. Targeted sequential indirect experiment design. In *Advances in Neural Information Processing Systems*, volume 37, pp. 122029–122053, 2024.

Anastasios N. Angelopoulos, Jacob Eisenstein, Jonathan Berant, Alekh Agarwal, and Adam Fisch. Cost-optimal active AI model evaluation. arXiv preprint arXiv:2506.07949, 2025.

Anders Björner and Gil Kalai. An extended Euler–Poincaré theorem. *Acta Mathematica*, 161: 279–303, 1988.

Harry Buhrman and Ronald de Wolf. Complexity measures and decision tree complexity: A survey. *Theoretical Computer Science*, 288(1):21–43, 2002. doi: 10.1016/S0304-3975(01)00144-X.

Ben Carterette, James Allan, and Ramesh K. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 268–275, 2006. doi: 10.1145/1148170.1148219.

Moses Charikar, Ronald Fagin, Venkatesan Guruswami, Jon M. Kleinberg, Prabhakar Raghavan, and Amit Sahai. Query strategies for priced information. *Journal of Computer and System Sciences*, 64(4):785–819, 2002. doi: 10.1006/jcss.2002.1828.

Ferdinando Cicalese, Eduardo Laber, and Aline Medeiros Saettler. Diagnosis determination: Decision trees optimizing simultaneously worst and expected testing cost. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *PMLR*, pp. 414–422, 2014.

Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 282–289, 1998.

Amol Deshpande, Lisa Hellerstein, and Devorah Kletenik. Approximation algorithms for stochastic submodular set cover with applications to boolean function evaluation and min-knapsack. *ACM Transactions on Algorithms*, 12(3):42:1–42:28, 2016. doi: 10.1145/2876506.

Florian E. Dorner and Moritz Hardt. Don't label twice: Quantity beats quality when comparing binary classifiers on a budget. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11544–11572. PMLR, 2024.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025.

Christoph Dürr, Arturo Merino, José A. Soto, and José Verschae. Set selection with uncertain weights: Non-adaptive queries and thresholds. arXiv preprint arXiv:2404.17214, 2024.

Edith Elkind, Leslie Ann Goldberg, Paul W. Goldberg, and Michael Wooldridge. On the computational complexity of weighted voting games. *Annals of Mathematics and Artificial Intelligence*, 56:109–131, 2009. doi: 10.1007/s10472-009-9162-5.

Tomás Feder, Rajeev Motwani, Rina Panigrahy, Chris Olston, and Jennifer Widom. Computing the median with uncertainty. *SIAM Journal on Computing*, 32(2):538–547, 2003. doi: 10.1137/S0097539701395668.

Rohan Ghuge, Anupam Gupta, and Viswanath Nagarajan. Nonadaptive stochastic score classification and explainable half-space evaluation. *Operations Research*, 73(4):2204–2222, 2025. doi: 10.1287/opre.2023.0431.

Rohan Ghuge, Jai Moondra, and Mohit Singh. Stochastic function certification with correlations, 2026. arXiv:2604.02611.

Tony Ginart, Martin Jinye Zhang, and James Zou. MLDemon: Deployment monitoring for machine learning systems. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3962–3997. PMLR, 2022.

Dimitrios Gkenosis, Nathaniel Grammel, Lisa Hellerstein, and Devorah Kletenik. The stochastic score classification problem. In *26th Annual European Symposium on Algorithms (ESA 2018)*, volume 112 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 36:1–36:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPIcs.ESA.2018.36.

Andrew Guillory and Jeff A. Bilmes. Interactive submodular set cover. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 415–422, 2010.

Satoshi Hara, Mitsuru Matsuura, Junya Honda, and Shinji Ito. Active model selection: A variance minimization approach. *Machine Learning*, 113(11):8327–8345, 2024.

Lisa Hellerstein, Devorah Kletenik, Naifeng Liu, and R. Teal Witter. Adaptivity gaps for the stochastic boolean function evaluation problem. In *Approximation and Online Algorithms (WAOA 2022)*, volume 13538 of *LNCS*, pp. 190–210, 2022. doi: 10.1007/978-3-031-18367-6_10.

John R. Isbell. A class of simple games. *Duke Mathematical Journal*, 25(3):423–439, 1958.

Shervin Javdani, Yuxin Chen, Amin Karbasi, Andreas Krause, Drew Bagnell, and Siddhartha Srinivasa. Near optimal bayesian active learning for decision making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 430–438. PMLR, 2014.

Simon Kahan. A model for data in motion. In *Proceedings of the Twenty-Third Annual ACM Symposium on Theory of Computing*, pp. 265–277, 1991. doi: 10.1145/103418.103449.

Jeff Kahn, Michael Saks, and Dean Sturtevant. A topological approach to evasiveness. *Combinatorica*, 4(4):297–306, 1984.

Sanjeev Khanna and Wang Chiew Tan. On computing functions with uncertainty. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 171–182, 2001. doi: 10.1145/375551.375577.

Alexander Koebler, Thomas Decker, Ingo Thon, Volker Tresp, and Florian Buettner. Incremental uncertainty-aware performance monitoring with active labeling intervention. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2188–2196. PMLR, 2025.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763, 2021.

Tetsuji Kuboyama. How many labels does model choice need? certificates and budgets for selective prediction. arXiv preprint arXiv:2609.18622, 2026.

Dan Li and Evangelos Kanoulas. Active sampling for large-scale information retrieval evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 49–58. ACM, 2017. doi: 10.1145/3132847.3133015.

David E. Losada, Javier Parapar, and Álvaro Barreiro. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management*, 53(5):1005–1025, 2017.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Nicole Megow and Jens Schlöter. Set selection under explorable stochastic uncertainty via covering techniques. In *Integer Programming and Combinatorial Optimization (IPCO)*, 2023.

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):2:1–2:27, 2008.

Alistair Moffat, William Webber, and Justin Zobel. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 375–382, 2007.

Mads Anker Nielsen, Lars Rohwedder, and Kevin Schewior. Non-adaptive evaluation of k-of-n functions: Tight gap and a unit-cost ptas. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 353 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 26:1–26:18, 2025. doi: 10.4230/LIPIcs.APPROX/RANDOM.2025.26.

Chris Olston and Jennifer Widom. Offering a precision-performance tradeoff for aggregation queries over replicated data. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pp. 144–155. Morgan Kaufmann, 2000.

Azimkhon Ostonov and Mikhail Moshkov. Comparison of complexity of regular and oblivious decision trees for decision tables from closed classes. In *International Joint Conference on Rough Sets (IJCRS)*, Lecture Notes in Computer Science, pp. 3–17, 2025.

Ronald L. Rivest and Jean Vuillemin. On recognizing graph properties from adjacency matrices. *Theoretical Computer Science*, 3(3):371–384, 1976.

Luchen Tan and Charles L. A. Clarke. A family of rank similarity measures based on maximized effectiveness difference. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2865–2877, 2015. doi: 10.1109/TKDE.2015.2448541.

Alan D. Taylor and William S. Zwicker. *Simple Games: Desirability Relations, Trading, Pseudoweightings*. Princeton University Press, 1999.

Sana Tonekaboni, Gabriela Morgenshtern, Azadeh Assadi, Aslesha Pokhrel, Xi Huang, Anand Jayarajan, Robert Greer, Gennady Pekhimenko, Melissa McCradden, Fanny Chevalier, Mjaye Mazwi, and Anna Goldenberg. How to validate machine learning models prior to deployment: Silent trial protocol for evaluation of real-time models at icu. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 169–182, 2022.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. TREC-COVID: Constructing a pandemic information retrieval test collection. *ACM SIGIR Forum*, 54(1):1:1–1:12, 2020.

Jing Yuan and Shaojie Tang. Worst-case adaptive submodular cover. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1915–1922, 2023.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Tonguç Ünlüyurt. Sequential testing of complex systems: A review. *Discrete Applied Mathematics*, 142(1–3):189–205, 2004. doi: 10.1016/j.dam.2002.08.001.

Tonguç Ünlüyurt. Sequential testing problem: A follow-up review. *Discrete Applied Mathematics*, 377:356–369, 2025. doi: 10.1016/j.dam.2025.07.039.

A RESIDUAL STATES AND ELEMENTARY PROPERTIES

Appendices A–C give an independent constructive proof of the unit-cost, three-way-sign case of Theorem 1. The width-exchange argument converts any depth-k adaptive tree into the widest-k prefix, which is also the form checked by the exact verifier in Appendix F.

For each query i, let $\mathcal{A}_i \subset \mathbb{R}$ be a nonempty finite contribution set, with

$$m_i = \min \mathcal{A}_i, \qquad M_i = \max \mathcal{A}_i, \qquad w_i = M_i - m_i.$$

The feasible support is Cartesian, $\Gamma = \prod_i \mathcal{A}_i$, and the target is the sign class of $\sum_i d_i$; the factor $1/n$ in the main paper does not change that class. A residual state is a pair (s, R), where s is the sum of revealed contributions and R is the unrevealed index set. Its sharp bounds are

$$\ell(s, R) = s + \sum_{i \in R} m_i, \qquad u(s, R) = s + \sum_{i \in R} M_i.$$

The state is resolved if $\ell > 0$, or $u < 0$, or $\ell = u = 0$; otherwise it is unresolved.

Lemma A.0 (removing determined coordinates). If $w_i = 0$, then $\mathcal{A}_i = \{m_i\}$ and the contribution is known before reveal. Replacing s by $s + m_i$ and removing i from R leaves the decision problem and every policy cost unchanged.

Proof. Every feasible world has the same contribution at coordinate i, so revealing its outcome provides no information about the aggregate sign. □

Hence the remaining proof assumes $w_i > 0$ for every unrevealed coordinate.

Proposition A.1 (sharp residual bounds). At every residual state,

$$\min_{d_R \in \prod_{i \in R} \mathcal{A}_i} \left(s + \sum_{i \in R} d_i \right) = \ell(s, R),$$

$$\max_{d_R \in \prod_{i \in R} \mathcal{A}_i} \left(s + \sum_{i \in R} d_i \right) = u(s, R),$$

and both extrema are attainable.

Proof. The objective is additive and the feasible set is a Cartesian product, so all coordinate minima, or all coordinate maxima, can be attained simultaneously. □

Lemma A.2 (support monotonicity). Let $\Gamma' \subseteq \Gamma$ be any nonempty joint support. At every reveal history h that is consistent with at least one world in Γ', let $[\ell_{\Gamma'}(h), u_{\Gamma'}(h)]$ be the exact attainable aggregate range over worlds in Γ' consistent with h. Then

$$\ell_{\Gamma'}(h) \geq \ell_{\Gamma}(h), \qquad u_{\Gamma'}(h) \leq u_{\Gamma}(h).$$

Consequently, every sign certificate valid for the Cartesian envelope is valid for Γ', and along a common reveal sequence support-aware resolution occurs no later than envelope-based resolution.

Proof. The worlds consistent with h in Γ' form a subset of those consistent with h in Γ. Minimizing over a subset cannot lower the minimum, and maximizing over a subset cannot raise the maximum. The certificate regions therefore persist. This argument concerns validity and stopping along a fixed sequence; it does not extend the width-order theorem to correlated support. □

Lemma A.3 (certificate persistence). Once a residual state is resolved, every extension of that reveal history has the same sign class.

Proof. A reveal replaces one uncertain contribution by an attainable exact value. Therefore the lower bound cannot decrease and the upper bound cannot increase. Positive and negative certificates persist. If $\ell = u = 0$, then $u - \ell = \sum_{i \in R} w_i = 0$; all unrevealed contributions are already determined, so every extension remains null. □

B UNIVERSAL RESOLUTION AND WIDTH DOMINANCE

Fix an unresolved residual state (s, R). A set $S \subseteq R$ is *universally resolving* if every assignment $d_S \in \prod_{i \in S} \mathcal{A}_i$ leaves a resolved residual state. By Lemma A.3, every superset of a universally resolving set is universally resolving.

For $i \in R$, define the normalized outcome set

$$B_i = \{a - m_i : a \in \mathcal{A}_i\} \subseteq [0, w_i].$$

Both endpoints 0 and w_i belong to B_i. Write

$$\alpha = -\ell(s, R) \geq 0, \qquad \beta = u(s, R) \geq 0,$$
$$\alpha + \beta = \sum_{i \in R} w_i.$$

For $S \subseteq R$, let $W_S = \sum_{i \in S} w_i$. If the normalized revealed values sum to z, the residual interval is

$$[-\alpha + z,\ \beta - W_S + z].$$

Lemma A.4 (unresolved normalized interval). If $S \subsetneq R$, then the residual state after revealing S is unresolved exactly when

$$W_S - \beta \leq z \leq \alpha. \tag{1}$$

Proof. The residual is beneficial iff $z > \alpha$ and harmful iff $z < W_S - \beta$. Because at least one positive-width coordinate remains, the residual interval has positive width and cannot be an exact-null singleton. The complement of the two strict certificate regions is therefore (1). □

Lemma A.5 (fixed-order prefix characterization). For a fixed reveal order σ, its worst-case reveal depth equals the smallest k for which the prefix $S_k = \{\sigma_1, \ldots, \sigma_k\}$ is universally resolving.

Proof. If S_k is universally resolving, every world is certified by depth k. If it is not, some assignment to S_k leaves an unresolved state. Cartesian support extends that assignment to a full feasible world. Along that world the policy cannot have stopped at an earlier prefix, because any earlier certificate would persist by Lemma A.3. □

Lemma A.6 (width-dominance exchange). Let $S = T \cup \{j\}$ be universally resolving and let $i \in R \setminus S$ satisfy $w_i \geq w_j > 0$. Then $S' = T \cup \{i\}$ is universally resolving.

Proof. Assume S' is not universally resolving. Then there are normalized outcomes for T with sum y and an $x_i \in B_i$ such that

$$W_T + w_i - \beta \leq y + x_i \leq \alpha. \tag{2}$$

Since $0 \leq x_i \leq w_i$, (2) implies

$$W_T - \beta \leq y \leq \alpha. \tag{3}$$

If $y \geq W_T + w_j - \beta$, choose the attainable endpoint $x_j = 0$; then (1) holds for S, contradicting universal resolution.

Otherwise $y < W_T + w_j - \beta$. We claim $y \leq \alpha - w_j$. If $y > \alpha - w_j$, then $\alpha + \beta < W_T + 2w_j$. But with $Q = R \setminus (T \cup \{i, j\})$,

$$\alpha + \beta = W_T + w_i + w_j + \sum_{q \in Q} w_q \geq W_T + 2w_j,$$

a contradiction. Thus $y \leq \alpha - w_j$. Choose the other attainable endpoint $x_j = w_j$. Using (3),

$$W_T + w_j - \beta \leq y + w_j \leq \alpha,$$

so (1) again holds for S, a contradiction. □

Fix any deterministic total order $\prec$ that sorts coordinates by nonincreasing width: $w_i > w_j$ implies $i \prec j$. Let $\mathrm{Top}_k^{\prec}(R)$ denote the first k coordinates of R under this order.

Corollary A.7 (prescribed widest-set dominance). If $0 \le k \le |R|$ and a universally resolving set of size at most k exists, then $\mathrm{Top}_k^{\prec}(R)$ is universally resolving for every width-refining tie rule $\prec$.

Proof. Enlarge the original set to size k using monotonicity. If it differs from the target set, choose i in the target but not the current set and j in the current set but not the target. By definition of the top-k set, $w_i \ge w_j$. Lemma A.6 permits the replacement. Repeating this finite process reaches the prescribed target set. □

C ADAPTIVE-TO-FIXED REDUCTION

Let $V(s, R)$ be the optimal worst-case number of additional reveals among all deterministic adaptive policies from residual state (s, R).

Lemma A.8 (adaptive depth implies a universal widest prefix). If $0 \le k \le |R|$ and $V(s, R) \le k$, then $\mathrm{Top}_k^{\prec}(R)$ is universally resolving for every width-refining tie rule $\prec$.

Proof. We induct on k. For $k = 0$, the state is resolved and the empty set is universally resolving. Suppose $k \ge 1$. If the current state is already resolved, Lemma A.3 makes every reveal set universally resolving. Otherwise, an adaptive policy of depth at most k reveals some coordinate j first. For each attainable outcome $a \in \mathcal{A}_j$, the residual branch has adaptive value at most $k - 1$. If a branch is already resolved, monotonicity makes any set universally resolving there. Otherwise, the induction hypothesis says that the same set

$$T = \mathrm{Top}_{k-1}^{\prec}(R \setminus \{j\})$$

is universally resolving in that branch. The set is branch-independent because contribution widths and the tie rule do not depend on the observed value of j. Hence $T \cup \{j\}$ is universally resolving in the original state. Corollary A.7 converts it to $\mathrm{Top}_k^{\prec}(R)$. □

Unit-cost ordering (special case of Theorem 1). Assume Cartesian feasible support, additive aggregate utility, exact reveals, unit reveal costs, and a deterministic minimax worst-case objective. Then every deterministic reveal order that is nonincreasing in width, with arbitrary fixed tie-breaking and outcome-dependent stopping, attains the optimal worst-case reveal depth among all adaptive policies.

Proof. Apply Lemma A.0 exhaustively to the initial state, obtaining an equivalent state (s_0, R_0) in which every remaining coordinate has positive width. Let $k = V(s_0, R_0)$ and fix an arbitrary deterministic nonincreasing-width order σ. Let $\prec_\sigma$ be the total width-refining tie rule induced by σ among equal-width coordinates. Because full revelation always resolves the state, $0 \le k \le |R_0|$; Lemma A.8 makes $\mathrm{Top}_k^{\prec_\sigma}(R_0)$ universally resolving. These are exactly the first k positive-width coordinates of σ. Lemma A.5 therefore gives worst-case depth at most k. Append the removed zero-width coordinates at the end; the stopping rule never needs to reveal them. Conversely, every fixed-order policy is an adaptive policy, so its worst-case depth cannot be below k. Since σ was arbitrary, every nonincreasing-width order is minimax-optimal. □

The theorem compares minimax values, not realized paths. Stopping remains outcome-dependent; only reordering is fixed. Arbitrary fixed reveal costs and monotone multi-class links are covered by the relevance argument of Section 3; outcome-dependent costs are not.

D STRICT CORRELATED-SUPPORT GAP

Consider

$$\Gamma = \{(-2, -2, 1), (-2, 2, -2), (-2, 2, 1), (1, 2, -2)\},$$

with sign class determined by the coordinate sum. The four signs are $-, -, +, +$. Reveal x_3 first. If $x_3 = 1$, revealing x_2 separates the two remaining signs; if $x_3 = -2$, revealing x_1 does so. Thus the adaptive policy has worst-case depth two.

Every choice of two fixed coordinates leaves opposite-sign worlds indistinguishable:

Observed pair	Omitted	Opposite-sign worlds
$\{1,2\}$	3	γ_2, γ_3
$\{1,3\}$	2	γ_1, γ_3
$\{2,3\}$	1	γ_2, γ_4

where $\gamma_1 = (-2,-2,1)$, $\gamma_2 = (-2,2,-2)$, $\gamma_3 = (-2,2,1)$, and $\gamma_4 = (1,2,-2)$. Thus all six fixed orders have worst-case depth three.

Envelope certificates on a coupled support. A coupling model may be trusted enough to decide what to label next but not to certify a release: a verdict computed from the marginal sets alone stays valid for an auditor who rejects the coupling model. We therefore recomputed both optima with exact dynamic programming when the policy knows Γ and the adversary is restricted to Γ, but a history stops only when the Cartesian envelope $\prod_i \mathcal{A}_i$ of the marginal sets certifies. On the support above the gap disappears ($\mathrm{OPT}_{\mathrm{ad}} = \mathrm{OPT}_{\mathrm{fix}} = 3$): the adaptive tree of Proposition 2 relies on support-aware certification. Coupling can nevertheless create a gap under envelope certificates: on Γ_3 of Proposition 3(a) ($n = 7$), $\mathrm{OPT}_{\mathrm{ad}} = 5 < 6 = \mathrm{OPT}_{\mathrm{fix}}$, while support-aware certification gives $3 < 4$. On 400 random product supports both variants agree with Theorem 1. How often a random coupled support shows a gap under either rule depends strongly on how supports are sampled (two sampling schemes we tried ranked the two rules in opposite orders), so we report only these deterministic cases, which an independent reimplementation reproduced exactly.

E Randomized policies: proof of Proposition 4

Write worlds as (d_1, d_2, d_3). The class is $+$ exactly on $(2,-1,2)$, $(2,0,0)$ and $(2,0,2)$. All three coordinates are relevant, so every deterministic policy needs three reveals in the worst case (Theorem 1).

Fixed orders. The uniform mixture of the orders $(1,2,3)$, $(1,3,2)$ and $(2,3,1)$ has expected cost at most $8/3$ on every world. Conversely, under the uniform distribution on $\{(2,-1,0),(2,-1,2),(2,0,0)\}$ each of the six fixed orders has expected cost exactly $8/3$; by Yao's principle no mixture of fixed orders does better.

Adaptive policies. Policy P reveals d_1, stops if $d_1 = -2$, and otherwise reveals d_2 and then, if needed, d_3; its costs on the eight worlds are $(1,1,1,1,3,3,2,2)$ in lexicographic order. Policy Q reveals d_3; if $d_3 = 2$ it reveals d_1, and if $d_3 = 0$ it reveals d_2 and then, if needed, d_1; its costs are $(2,2,3,2,2,2,3,2)$. The equal mixture costs at most $5/2$ on every world. Conversely, under the uniform distribution on $\{(2,-1,2),(2,0,0)\}$, both worlds need their first coordinate and a different second one (d_3 and d_2 respectively); whichever coordinate a deterministic policy reveals first, it pays 3 on at least one of them, so its expected cost is at least $5/2$.

Exploratory scan. A floating-point scan over 1,220 random product supports with three or four coordinates, value sets of size two or three in $\{-2,\dots,2\}$ and half-integer thresholds, suggested larger randomized gaps. The largest ratio observed in this sample was $7/6$, on $\mathcal{A} = (\{-1,1\},\{-2,2\},\{1,2\},\{-2,-1\})$ with threshold $\frac{1}{2}$; we then verified this witness exactly with rational certificates. For randomized fixed orders, the mixture assigning probabilities $1/2, 1/4, 1/4$ to orders $(1,3,4,2)$, $(2,3,1,4)$ and $(2,4,1,3)$ has worst-world expected cost $7/2$, while the Yao distribution assigning mass $1/2$ to $(-1,2,1,-2)$ and mass $1/6$ to each of $(-1,2,1,-1)$, $(-1,2,2,-2)$ and $(1,2,1,-2)$ gives the matching lower bound. For randomized adaptive policies, an equal mixture of two exact decision trees has worst-world cost 3. Policy P queries coordinate 1; if $d_1 = 1$ it queries coordinate 2, while if $d_1 = -1$ it queries coordinate 3, then coordinate 2 when $d_3 = 2$, and when $d_3 = 1$ it queries coordinate 4 and queries coordinate 2 iff $d_4 = -1$. Policy Q queries coordinate 4; if $d_4 = -1$ it queries coordinate 2, while if $d_4 = -2$ it queries coordinate 3, then coordinate 2 when $d_3 = 2$, and when $d_3 = 1$ it queries coordinate 1 and queries coordinate 2 iff $d_1 = 1$. In every branch the policy stops as soon as the class is certified. The uniform Yao distribution on three of the seven positive worlds, $(-1,2,1,-1)$, $(-1,2,2,-2)$ and $(1,2,1,-2)$, gives the matching lower bound 3 over all deterministic decision trees. Thus $R_{\mathrm{fix}} = 7/2$ and $R_{\mathrm{ad}} = 3$ exactly. We report this only as an existence witness; the exploratory scan is not a bound on the largest possible randomized gap.

F INDEPENDENT EXACT VERIFICATION

The accompanying verifier uses exact rational arithmetic. For a residual state, it computes

$$\begin{aligned} V(s,R) &= 0 && \text{if } (s,R) \text{ is resolved,} \\ V(s,R) &= 1 + \min_{i \in R} Q_i(s,R) && \text{otherwise,} \\ Q_i(s,R) &= \max_{a \in \mathcal{A}_i} V(s+a, R \setminus \{i\}). \end{aligned}$$

It compares this fully adaptive recurrence with the analogous recurrence restricted to each fixed order. The implementation uses the sharp residual bounds from Proposition A.1, so it does not approximate the attainable sign class.

The recorded verifier run passes three test suites: (i) all 20,475 four-query multisets constructed from support sets of size at most three over $\{-2,-1,0,1,2\}$ for a deterministic width order; (ii) all 1,708 instances with two to six queries over the nonempty subsets of $\{-1,0,1\}$, checking every width-compatible tie order; and (iii) 3,000 random multivalued instances with two to six queries. The second test stresses equal-width tie choices, singleton sets, exact-null states, and early resolution. For the general statement of Theorem 1, on 4,000 random product supports with one to three cut points and random positive costs, the adaptive optimum, the worst-case cost of the width order and $c(\mathcal{R})$ coincide in every case. These runs serve as regression checks for the implementation and proof edge cases.

G PROOF OF PROPOSITION 3

Proof. (a) Let $\Gamma_1 = \{(-1),(+1)\}$. Given Γ_{d-1} on n_{d-1} coordinates with no zero sum, let $\mu > 0$ be its minimum absolute coordinate sum, fix $0 < \varepsilon < \mu$, let $\mathbf{0}$ be the zero vector of length n_{d-1}, and define

$$\Gamma_d = \{(-\varepsilon, x, \mathbf{0}) : x \in \Gamma_{d-1}\} \cup \{(+\varepsilon, \mathbf{0}, x) : x \in \Gamma_{d-1}\}.$$

Since $|\sum_k x_k| \geq \mu > \varepsilon$, every world keeps the sign of its embedded x and no sum is zero; by induction each coordinate takes at most two nonzero values (the new first coordinate takes $\pm\varepsilon$, and embedding only adds the value 0), so $|\mathcal{A}_i| \leq 3$. *Upper bound:* revealing the first coordinate identifies the active block, on which the residual support is a copy of Γ_{d-1} while all other coordinates are constant; recursing gives depth at most d. *Lower bound:* by Lemma 3, $\mathrm{OPT}_{\mathrm{fix}}$ is the minimum size of a determining set S. Worlds in the left branch agree on the first coordinate and on the right block, so S restricted to the left block must determine λ on a copy of Γ_{d-1}; the same holds for the disjoint right block. Hence $\mathrm{OPT}_{\mathrm{fix}}(\Gamma_d) \geq 2\ \mathrm{OPT}_{\mathrm{fix}}(\Gamma_{d-1})$, with $\mathrm{OPT}_{\mathrm{fix}}(\Gamma_1) = 1$. (b) Let U be the set of coordinates queried anywhere in an optimal adaptive tree of depth $a = \mathrm{OPT}_{\mathrm{ad}}$. Two worlds that agree on U follow the same path to the same leaf, where the residual support is class-homogeneous; so U is a determining set, and $|U|$ is at most the number of internal nodes, $\sum_{t<a} k^t$. With $f = \mathrm{OPT}_{\mathrm{fix}} \leq n$, this gives $a > \log_k f$ and $f/a < f/\log_k f \leq n/\log_k n$ for $3 \leq f \leq n$; if $f \leq 2$, then $a \geq 1$ unless λ is constant, so $f/a \leq 2$. (c) The class-1 worlds are $110, 001, 011, 111$ and the class-0 worlds are $000, 100, 010, 101$. Reveal x_1: if $x_1 = 1$ the class equals x_2, and if $x_1 = 0$ it equals x_3; so $\mathrm{OPT}_{\mathrm{ad}} = 2$. Fixing any two coordinates leaves the pairs $(000, 001)$, $(100, 110)$ or $(001, 101)$ unresolved, so $\mathrm{OPT}_{\mathrm{fix}} = 3$ by Lemma 3. (d) The address function reads k address bits and one of 2^k data bits (Buhrman & de Wolf, 2002); every coordinate is relevant, and a determining set on the full product must contain every relevant coordinate. Exact dynamic programming confirms $(\mathrm{OPT}_{\mathrm{ad}}, \mathrm{OPT}_{\mathrm{fix}}) = (3,4)$ for $n = 7$ and $(4,8)$ for $n = 15$ in (a), and on the full cube $\{0,1\}^4$ exhaustive enumeration finds a strict gap for 7,560 of 65,534 nonconstant labelings and for none of the 1,880 threshold labelings. □

H ADDITIONAL TREC-COVID ANALYSES

Distribution of certification depth. Table 3 reports quantiles of depth divided by the union size n (same denominator for all orders; random restricted to nonzero-weight documents, per-instance median over 100 permutations).

Order	p10	p25	p50	p75	p90
Width	0.36	0.44	0.60	0.79	0.93
Rank priority	0.36	0.46	0.64	0.83	1.00
Moffat B	0.36	0.50	0.64	0.85	1.00
Random (nonzero weight)	0.58	0.69	0.79	0.92	1.00
Realized oracle	0.36	0.36	0.43	0.50	0.57
Gap (width − oracle)/n	0.00	0.07	0.15	0.29	0.36

Table 3: TREC-COVID quantiles of the fraction of the top-7 union judged before certification.

Expected-case adaptive policies. A leave-one-topic-out prior $P(\text{relevant} \mid \text{best rank})$ ranges only from 0.65 to 0.50, against a 3.8-fold spread of $|c_i|$, so the static prior order $p_i|c_i|$ coincides with width order on 73.4% of instances. On all 108,715 retained instances, mean depths are 8.197 (width), 8.208 (static prior order) and 8.200 (adaptive direction-aware greedy), against 5.959 for the realized oracle. Exact optimization is feasible on a frozen 2,000-instance subset with at most 10 documents: width and oracle mean depths there are 5.857 and 4.580, and the fitted-prior optimal adaptive policy closes 20.4% of that realized gap on means and none on medians. To isolate reordering from the value of the prior itself, we also solve the optimal fixed order under the same prior; the adaptive policy lowers expected cost by 0.432 judgments on average. A nine-point prior-strength sweep over 31,655 tractable cases gives a non-monotone adaptive-over-fixed gain: it grows at intermediate information and returns to zero near perfect information, when a fixed order can already follow the oracle path.

Indifference band. With the three-tier rule $g(\Delta) = \operatorname{sgn}(\Delta - \tau) + \operatorname{sgn}(\Delta + \tau)$ and $\tau \in \{0, 0.02, 0.10\}$, width order has mean depth 8.20 / 8.54 / 9.35 and median fraction judged 0.60 / 0.64 / 0.75, and the share of negligible verdicts is 7.3% / 10.3% / 29.8%. A negligible verdict is two-sided and costs 11.4 judgments on average (full reveal in 70.7% of such cases). The ranking of orders (width, rank priority, Moffat B, A, C, random) is unchanged at every τ.

Genuinely incomplete instances. Among the 195,875 excluded instances, whose unions contain unjudged documents, revealing only judged documents in width order and treating unjudged ones as permanently unknown certifies 46.5% at $\tau = 0$ and 43.7% at $\tau = 0.02$ (median 7 judged documents), from 83.1% with one unjudged document to 0.2% with nine or more.

I TREC-COVID RELEVANCE MAPPING

TREC-COVID judgments are graded (0 not relevant, 1 partially relevant, 2 fully relevant). The main analysis uses $y = \mathbf{1}[\text{qrel} \geq 1]$.

Three-level relevance. As a multivalued robustness check, we retain all three grades and map them to gains 0, $1/2$ and 1, so each document contributes one of $\{0, c_i/2, c_i\}$. The replay is valid on all 108,715 retained instances; exact oracle checks agree with exhaustive search on 117 sampled instances. Certification becomes slightly deeper than in the binary replay, but the ordering of the realized oracle, width, rank-priority, Moffat baselines and random order is unchanged.

Strict binary mapping. Under $y = \mathbf{1}[\text{qrel} = 2]$ (537,530 instead of 789,705 relevant document slots), the kept set and the initially unresolved share (99.74%) are unchanged by construction; the median fraction judged is 0.62 (width), 0.83 (random) and 0.43 (oracle); width order requires every judgment in 13.1% of instances; and mean depths are 8.40 (width), 8.66 (rank priority), 8.76 (Moffat A and B), 9.26 (Moffat C) and 10.46 (random). The ordering of policies is identical under both binary mappings. Zero-weight documents are 1.34% of document slots; restricting the random order to nonzero-weight documents lowers its mean depth from 10.27 to 10.15 (qrel ≥ 1) and from 10.46 to 10.35 (qrel $= 2$).

J Proof of Corollary 1 and two-classifier replay

Proof. With $d_i = \mathbf{1}[f^1(x_i) = y_i] - \mathbf{1}[f^0(x_i) = y_i]$, agreements have $\mathcal{A}_i = \{0\}$ and disagreements have $\{-1, +1\} \subseteq \mathcal{A}_i \subseteq \{-1, 0, +1\}$, so all disagreements have width 2. Write $t = n\tau$; the verdict is undetermined iff $-|D| \le t \le |D|$. For $j \in D$, the other disagreements attain every sum $b \in \{-(|D|-1), -(|D|-1)+2, \ldots, |D|-1\}$, and the intervals $[b-1, b+1]$ cover $[-|D|, |D|]$; choosing b with $b - 1 \le t \le b + 1$ gives $\operatorname{sgn}(b - 1 - t) < \operatorname{sgn}(b + 1 - t)$, so j is relevant. Apply Theorem 1. □

We use a two-classifier replay to illustrate Corollary 1 in a modern model-comparison setting. With four answer options, a disagreement on which both models may be wrong has $\mathcal{A}_i = \{-1, 0, +1\}$, giving a multivalued instance of Theorem 1. We queried three Azure OpenAI deployments (gpt-4o, gpt-4.1-mini, gpt-4o-mini) at temperature 0 on 1,000 fixed items (seed 0) each from MMLU (test) and ARC-Challenge, asking for the option letter only. Seventeen MMLU prompts were rejected by the provider's content filter and are excluded for all models, leaving a common pool of 983. Accuracies are 0.850 / 0.788 / 0.746 on MMLU and 0.965 / 0.952 / 0.923 on ARC-Challenge. For every model pair and margin $\tau \in \{0, 0.01, 0.02\}$ (three-tier rule, exact integer arithmetic), we hid the labels and replayed (a) labeling disagreements only in random order, (b) labeling the whole pool in random order (20 seeds each), and (c) the realized oracle. Table 4 reports ranges over pairs and margins. Labeling only disagreements certified the same verdict as labeling the whole pool in all 720 replays; (b)/(a) tracks $n/|D|$, as Corollary 1 predicts (every disagreement has width 2 and every agreement width 0, so (a) is also the width order and a width-ordered pass over the whole pool coincides with it); and random order over D still needed a median of 59–94% of $|D|$, while the oracle needed 58–89% of that, so the worst case $|D|$ is not loose in practice.

Benchmark	pool	$\lvert D\rvert$	(a) disagreements only	(b) whole pool	(c) oracle	saving (b)/(a)
MMLU	983	141–215	100–196	662–899	71–125	4.5–6.8×
ARC-Challenge	1,000	32–63	22.5–57	571–925	13–42	15–32×

Table 4: Labels needed to certify the three-tier accuracy verdict (medians over 20 seeds; ranges over 3 model pairs and 3 margins).