How Much Evidence Is Needed to Trust Model Comparisons?

Anonymous authors
Paper under double-blind review

Abstract

Offline model comparisons depend on which candidates were allowed to compete. When historical membership is incomplete, a reconstructed list can turn an uncertain comparison into a seemingly precise winner. We study the evidence needed for a certified model choice. The key is shared uncertainty: membership changes can cancel between models. We prove that a comparison can require zero new records while fixing all selections requires linearly many. For single-query top-K evaluation, we characterize each feasible contrast through a single shared boundary block. This yields an exact algorithm whose exponential dependence is local, and a closed-form regret formula for disjoint adjacent swaps. A matching search separation shows that exact computation and minimal evidence acquisition are distinct problems. For general model orders and multiple queries, joint disagreement and cutoff constraints provide efficient, monotone regret bounds. These certificates support comparison-guided acquisition and stopping before full recovery. Seven trained stock scorers exhibit 42 winner reversals across 132 pool-date comparisons; 23 incur regret exceeding 0.01. Frozen-score experiments repair all 23 above-tolerance errors while leaving more than half the candidate memberships unresolved on average. In a controlled width-two stock reranking task, Block certifies 67/132 comparisons versus 5/132 for Cutoff and its block-span refinement at the same additional 25% record budget. Separate policy experiments save 53.1% of recommendation records at zero regret.

1 Introduction

A benchmark winner is a statement about both a model and an evaluation task. In ranking applications, the task includes the candidates eligible to compete: stocks in a historical pool, products available to a customer, or documents available at a query time. Changing this list can distort comparisons (Krichene & Rendle, 2020; Daniel et al., 2009), even when predictions and outcome labels stay fixed. A later list substituted for a missing historical list can therefore change the winner without any model improvement.

This work asks: *how much membership evidence is needed to trust a model comparison?* Complete reconstruction is sufficient, but often unnecessary. Even fixing every model's selected identities can demand more information than choosing a good model. A record is useful for model selection when it can resolve a disagreement between competing conclusions, not simply because its membership is unknown.

The distinction follows from shared uncertainty. If two models respond in the same way to an unknown part of the candidate state, their evaluation values can vary while their difference remains fixed. Bounding those values independently loses this cancellation. Recovering every selection also asks a stronger question than the comparison requires. Active evaluation already reduces the work needed to identify strong models (Zhou et al., 2025). Here, one unknown membership bit can change the selected candidates for several models at once. We exploit that shared state to certify their comparison directly.

Our contributions are threefold. First, frozen-score experiments establish membership-induced misselection, and a zero-versus-linear separation explains why model choice needs a different evidence target from recovery. Second, a count-feasible boundary-block characterization makes local reranking exactly computable, while joint disagreement and cutoff enclosures handle general

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.

orders. We separate local computational tractability from online evidence-search cost within the same ranking class. Third, frozen-score repair and constrained-reranking interventions test these mechanisms, with matched evidence and separate query-policy comparisons establishing where the record savings arise.

2 The Evaluation Problem

Let Ω be a reference catalog of n candidate identities, $\mathcal{M}$ a set of M models, and $\mathcal{U}$ a finite evaluation-query set. A membership vector $z \in \{0,1\}^n$ defines $C(z) = \{i \in \Omega : z_i = 1\}$. Query u has fixed eligibility $E_u \subseteq \Omega$ and integer cutoff $K_u \geq 1$. For each query, frozen prediction scores define each model's total order, with ties broken by ascending identity. Let TopK_{ua} return the first K_u candidates in that order. For finite gains g_{ui} and weights $w_u > 0$, $\sum_u w_u = 1$, define

$$S_{ua}(z) = \mathrm{TopK}_{ua}(C(z) \cap E_u), \qquad Q_a(z) = \sum_u w_u \sum_{i \in S_{ua}(z)} g_{ui}. \tag{1}$$

Gains and eligibility remain fixed across states and models. Recall uses $g_{ui} = \mathbf{1}\{i \in T_u\}/|T_u|$ for a fixed nonempty relevant-item set T_u, where $\mathbf{1}$ is the indicator. Mean bounded gain uses $g_{ui} = y_{ui}/K_u$ for fixed utilities $y_{ui} \in [0,1]$.

Evidence $\mathcal{E}$ supplies known members/nonmembers and integer counts c_g in G groups partitioning Ω. A total count gives one group; a snapshot O with known overlap gives two, O and $O^c = \Omega \setminus O$. The family $\mathcal{Z}(\mathcal{E})$ contains compatible memberships with at least K_u eligible members per query. We assume this family is nonempty; inconsistent records are rejected. All models and evaluation queries share one z. A membership query reveals one candidate's status for all evaluation queries. Joint's exact coordinate bounds assume that counts and known bits already imply query admissibility. Appendix D gives this check and the extension to a relaxed family.

A model-choice certificate. Define worst-case model-selection regret by

$$R(a \mid \mathcal{E}) = \max_{z \in \mathcal{Z}(\mathcal{E})} \left[\max_b Q_b(z) - Q_a(z) \right]. \tag{2}$$

For $\epsilon \geq 0$, a bound $\overline{R}(a) \leq \epsilon$ certifies that a is within ϵ of the best model for every remaining state. At $\epsilon = 0$, tied best models are allowed. This finite-evaluation guarantee, conditional on frozen predictions, labels and evidence, defines our stopping rule. The implementation uses tolerance $\tau = 10^{-8}$ and stops at $\epsilon + \tau$. Otherwise the choice remains unresolved. Realized regret is evaluated using the withheld membership only after the query prefix is fixed.

Empirical discovery with frozen predictions. We compare Ridge, LightGBM, MLP, GRU and common-input implementations of StockMixer, MASTER and SSPT (Fan & Shen, 2024; Li et al., 2024; Wang et al., 2025a). Training uses 2018–2022, validation 2023, and testing 2024–2026. The fixed cohort contains 5,072 stocks known by the training cutoff; inference context is unchanged when membership changes. Four capitalization pools and 33 complete-label dates give 132 pool-date comparisons. The primary metric is the mean forward-return percentile of the selected 60 stocks, with percentiles defined once on the common reference catalog.

Substituting terminal for point-in-time membership reverses the strict best model in 42/132 comparisons and 405/2,772 model pairs. Choosing the terminal winner incurs mean point-in-time regret 0.0061 on this $[0,1]$ metric; the mean among changed-winner cases is 0.0191. Recall@60 yields nine strict winner reversals, with unique winners in both states in 39 cases. The experiment isolates membership substitution as an argument of the frozen evaluator and motivates certifying the comparison before acting on it. Of the 132 comparisons, 23 incur regret above 0.01; the maximum is 0.0582. These quantities measure changes in normalized evaluation utility.

3 Why Comparing Can Be Easier Than Recovering

Write $D(z) = (S_{ua}(z))_{u,a}$ for the full selection vector. Fixing D fixes every bounded additive evaluation value, and the converse holds when equality is required for *all* gain assignments. This provides

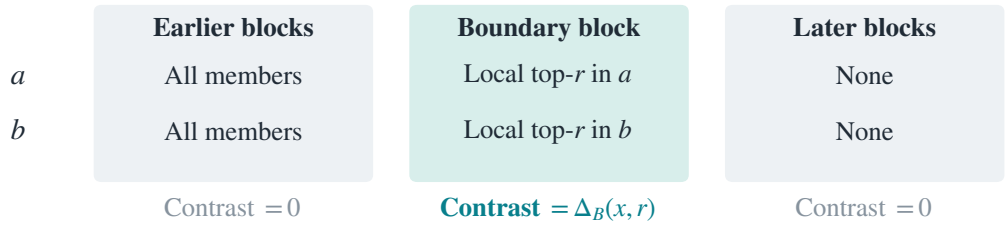


Figure 1: For one membership state, only one shared block can straddle the top-K boundary. Earlier blocks contribute equally to both models; later blocks contribute nothing. Across possible states, different boundary blocks are alternatives, not additive sources of contrast.

a reusable selection-recovery certificate, whose exact count-family cost is derived in Appendix F.1. Model choice instead requires certifying $R(a \mid \mathcal{E}) \leq \epsilon$ for some a.

Proposition 1 (Zero-versus-linear evidence separation)**.** *For every even $N \geq 2$, there is a count-feasible evaluation with N unknown memberships and two known anchors in which one model is a strict winner without new records, while a certificate fixing D requires $N/2$ records.*

Both models share an evaluation query selecting the $N/2$ actual members from N unknown identities. An anchor-only query contributes a fixed positive margin to A. The uncertain shared-query contributions cancel, so A always wins. Fixing the entire selected subset still requires $N/2$ records. The construction and independently checked instances appear in Appendix F.

Exact global optimization remains difficult. Maximizing a shared model contrast across a shared-membership query collection is NP-hard even for two binary-score top-1 models and one exact count: sets in Maximum Coverage become candidate identities, elements become queries, and a reference model selects a mandatory zero-gain anchor (Feige, 1998). This motivates efficient certification from exact local structure rather than exhaustive global state optimization.

4 Exact Comparison from a Shared Boundary Block

When models differ through local reranking, shared order structure can make exact comparison tractable. Assume a single evaluation query and models a, b; write $K = K_u$, $g_i = g_{ui}$ and $S_a = S_{ua}$, suppressing fixed query and evidence indices. Delete known-absent eligible candidates and split the retained orders wherever their prefix sets agree. This standard permutation block decomposition (Adin et al., 2017) gives a shared block sequence, with possibly different orders inside each block.

One boundary, one source of disagreement. All active identities in earlier blocks are selected by both models; none in later blocks is selected. If p active identities precede the boundary block B, both models select $r = K - p$ members inside it. Writing $x \subseteq B$ for its active identities and $\Delta_B(x, r)$ for the difference of the two local top-r gain sums, the entire model contrast equals $\Delta_B(x, r)$. If the cutoff ends after a complete block, the contrast is zero. Thus an arbitrary constant added to all gains in one block cancels from this pair's contrast in every state, even when the selected identities differ.

Which local states extend to a common catalog? Let P contain earlier blocks and $S = \Omega \backslash (P \cup B)$, including query-ineligible candidates. In group g, let m_{Pg}, a_{Pg} and m_{Sg}, a_{Sg} count mandatory (entailed present) and available (not entailed absent) candidates. For block membership x consistent with known bits, let x_g count its members in group g. The feasible prefix count is exactly

$$p_g^-(x) = \max\{m_{Pg}, c_g - x_g - a_{Sg}\}, \qquad p_g^+(x) = \min\{a_{Pg}, c_g - x_g - m_{Sg}\}. \tag{3}$$

Discard x if any interval is empty. Otherwise every integer between $\sum_g p_g^-$ and $\sum_g p_g^+$ is attainable. With $t = |x|$, a partial boundary has exactly the following feasible selected sizes:

$$\max\{1, K - \textstyle\sum_g p_g^+(x)\} \leq r \leq \min\{t - 1, K - \sum_g p_g^-(x)\}, \quad r \in \mathbb{Z}. \tag{4}$$

Theorem 2 (Exact boundary-block contrast)**.** *For a nonempty single-query count family restricted to admissible states, the minimum and maximum of $Q_a - Q_b$ equal the extrema of all $\Delta_B(x, r)$ satisfying Eqs.* (3)–(4)*, including zero when a common block boundary can have exactly K prefix members. If B contains b_B retained identities and u_B unknown bits, these endpoints take*

$$O\left(nG + \sum_B 2^{u_B}(b_B + G)\right) \tag{5}$$

work and $O(n + G + \max_B b_B)$ memory. Every endpoint has an attaining membership completion. For maximum block size b, the work is $O(nG2^b)$, and valid new evidence cannot widen the interval.

Every actual state supplies one feasible boundary configuration. Conversely, the group intervals complete every enumerated configuration to a prefix with $K - r$ members and a suffix satisfying all counts. This proves both containment and attainment. Local membership enumeration and the two ordered gain-prefix sums give the work bound; no complete-catalog state enumeration is needed when the blocks are small. We call this exact certificate Block. Appendix A gives the full proof.

A closed form for adjacent swaps. A swap (i, j) changes the selections exactly when both candidates are present and their earlier prefix has $K - 1$ members. Let $\mathcal{R}$ contain the reachable pairs and $\delta_{ij} = g_i - g_j$ in a's orientation. For models differing by disjoint adjacent swaps,

$$R(a) = \max(\{0\} \cup \{-\delta_{ij} : (i, j) \in \mathcal{R}\}), \quad R(b) = \max(\{0\} \cup \{\delta_{ij} : (i, j) \in \mathcal{R}\}). \tag{6}$$

The test costs $O(G)$ per pair, so exact regret takes $O(nG)$ work. The signs of the reachable exchanges determine whether a model is already certified; unknown memberships outside these alternatives need not be resolved.

A lossless regime for the general relaxation. Assume automatic admissibility, disjoint adjacent swaps, and at most one reachable swap under the current evidence. Then Joint and Cutoff have exact contrast endpoints. If no swap is reachable, the difference vector is zero. Otherwise it is either zero or $e_i - e_j$, where e_i is the i-th standard basis vector. Exact coordinate extrema and zero sum describe the segment between feasible alternatives, whose endpoints are attained by actual states. Neither relaxation adds an unattainable objective value.

With several reachable blocks, coordinate possibilities do not specify which differences belong to one common boundary state. Small blocks make exact reasoning inexpensive even when one-swap losslessness fails. Before enumeration, an $O(nG)$ count test identifies exactly which blocks can contain a partial boundary (Appendix A.3). The largest number of unknown bits in those blocks gives an observable tractability condition; other comparisons use the general enclosures.

Computation and evidence search are different costs. Exact evaluation and online evidence search have distinct complexity even at block size two. For integer $h \geq 1$ and $0 \leq \epsilon < 1$, there are $n = 4h$-candidate top-1 families whose minimum offline certificate cost is $C_* = 1$ at every state, but optimal worst-case online depth is $D_* = h = n/4$. Half the swap blocks favor each model; exactly two candidates are members. A positive answer isolates one possible nonzero exchange, whereas negative answers must eliminate every block of one sign. Appendix A proves the matching bounds.

An inexpensive block-span control. At most $\lfloor b_B/2 \rfloor$ exchanges occur inside the active block. Hence $|Q_a - Q_b|$ is at most $\max_B \lfloor b_B/2 \rfloor (\max_{i \in B} g_i - \min_{i \in B} g_i)$. This $O(n)$ control discards exchange direction and count feasibility. We intersect it with the general certificate below to isolate the value of exact boundary-state reasoning in the experiments.

5 Reliable Certificates for General Model Orders

Unrestricted model pairs can have a full-catalog block. We use polynomial-time outer bounds for that regime, including multiple evaluation queries. For one query, write $d_i = \mathbf{1}\{i \in S_a\} - \mathbf{1}\{i \in S_b\}$. The contrast is $\sum_i g_i d_i$, and equal selection sizes give $\sum_i d_i = 0$. Joint predecessor counts determine exact coordinate extrema ℓ_i, u_i under automatic admissibility (Theorem 4). Optimizing over this box and zero sum, then intersecting separate model-value bounds (Marginal), gives Joint.

Table 1: Structural regimes and their certificates. Work is per model pair and evaluation query; b is maximum retained block size and s counts candidates with potentially nonzero disagreement. Exact single-query endpoints and multi-query outer bounds serve different regimes.

Order structure	Certificate	Work
Disjoint adjacent swaps	Exact model regret	$O(nG)$
Shared blocks of size at most b	Exact contrast endpoints	$O(nG2^b)$
Unrestricted orders	Joint/Cutoff outer bounds	$O(Gn \log n + n + sK)$

Cutoff additionally enforces that all positive differences precede all negative ones in a's order, with at most K of each. It optimizes each one-order cutoff family and intersects the resulting bounds. Both certificates are sound and monotone. For either certificate, the aggregated upper bound U_{ab} on $Q_a - Q_b$ gives $\overline{R}(a) = \max(0, \max_b U_{ba})$. Theorems 5 and 6 give the derivations.

The methods preserve different structure: Joint retains coordinate disagreement, Cutoff adds compatible signs, and Block enforces one feasible boundary state. In the three-swap example of Appendix A, gains are constant within each block, so every true contrast is zero. The two order-specific optimizers nevertheless combine exchanges across different blocks, giving a positive upper endpoint. Block excludes those incompatible combinations. The span control also resolves this example; the real-score experiment tests the additional value of exact count and directional constraints.

The choice of certificate is determined by observed score-order structure and evidence, not by the hidden membership answers. For a single query, exact pairwise endpoints give exact model regret because the maximum over challengers commutes with the maximum over common states. With multiple queries, sums of query-wise maxima remain sound outer bounds, since their maximizing memberships may differ. The acquisition policy below can use such bounds without requiring each endpoint to be globally exact.

Certifying a choice does not require a precise metric estimate. If $U_{ab} \leq 0$, model b is at least as good as a in every compatible state, even if their contrast interval is wide. These bounds certify a model a when $U_{ba} \leq \epsilon$ for every challenger b. We therefore measure how many decisions can be certified from a given record budget, and how many records certification requires.

6 Comparison-Guided Record Acquisition

Contrast is our comparison-guided membership-query policy. Planning and certification use distinct incumbents. Choose $a_{\text{plan}} \in \arg\min_a \overline{R}_{\mathsf{J}}(a)$ and $a_{\text{out}} \in \arg\min_a \overline{R}_{\text{cut}}(a)$, breaking ties by a fixed model-name order. If a_{out} passes the requested threshold, return it with its evidence prefix and regret bound. Otherwise a_{plan} directs the next membership queries. This holds the query policy fixed while comparing stopping certificates on the same information; the returned model need not be the planning incumbent.

Matched-prefix experiments use Joint for planning and compare both certificates on the same evidence. Write $a = a_{\text{plan}}$; unresolved challengers satisfy $U^{\mathsf{J}}_{ba} > 0$. For each pair and evaluation query u, candidate j contributes $w_u|g_j - \lambda|(u_j - \ell_j)$, suppressing local query/pair indices. Here λ is the allocation threshold in Eq. (13), defined fully in Appendix E. A record influences j if it is j itself or precedes j in either eligible order. A weighted rank sweep sums these contributions, including zero-gain predecessors. Count effects enter through evidence updates. A zero-gain predecessor can displace a high-gain candidate near a model's cutoff. Revealing it as absent rules out those candidate states. Different predecessor orders across models make such records informative about the comparison.

Each optimistic top-K frontier contains the first K candidates not ruled out in that model's order. Query up to 64 unknown frontier candidates, ranked by comparison influence. Ties use frontier frequency, earliest rank and identity. Answers and counts imply further memberships, after which the certificate is recomputed. Gains are available to the policy; unrevealed memberships are accessed only after the query plan is fixed.

Table 2: Independent interventions isolate candidate-state, certificate, acquisition-policy and training-budget effects.

Question	Changed	Held fixed
Distortion	Candidate membership	Predictions, gains, metric
Certificate	Bound and stopping rule	Evidence and requested order
Acquisition	Next queried records	Certificate, initial facts, budget
Training	Validation-selected steps	Test catalog, queries and labels

Proposition 3 (Stopping and frontier support)**.** *The policy returns an ϵ-best model when its regret bound is at most ϵ. Each paid query concerns a truly selected candidate or a nonmember preceding a true selection cutoff, and is made at most once. Resolving these memberships fixes every model's evaluation value, ensuring termination; comparison certification can stop earlier.*

The influence rule is a tractable strategy derived from comparison support; correctness follows from the certificate. We therefore evaluate query-policy benefits separately from changing the certification target.

Fixed-order queries unknown memberships in deterministic identity-hash order. Frontier ranks optimistic-frontier candidates by frequency, earliest rank, then identity. Count-frontier prioritizes $p(1-p)$, where p is the required-member fraction among the group's unknown identities, breaking ties by Frontier.

7 Experiments

Section 2 isolates membership-induced model-choice errors. We now test their certified repair, the benefit of exact boundary structure, and the separate effects of stopping rules and acquisition policies.

Frozen-input protocol and precision. The stock study uses 264 mean-percentile instances: 33 dates, four pools, and two evidence conditions (known pool size, with or without snapshot overlap). Recall uses the 132 overlap instances; all use seven fixed scorers. Every policy starts with 25% known records in a fixed hash order and, for MovieLens, memberships implied by point-in-time-filtered positive labels. All four full recommendation catalogs from MovieLens 25M (Harper & Konstan, 2015; GroupLens, 2019) retain 32 evaluation queries and six scorers, including BPR-MF (Rendle et al., 2009) and LightGCN (He et al., 2020). We report $\epsilon \in \{0, .001, .0025, .005, .01, .025\}$. The primary stock tolerance is $\epsilon = .01$. Fixed-budget (60-step) recommendation best Recall is .014–.024, so the same numerical tolerance can permit 41–70% of that score; recommendation results therefore also foreground zero-regret certification ($\epsilon = 0$).

Repairing the discovered errors. Figure 2 shows that, at the Cutoff stopping point, none of the 132 stock cases incurs regret above .01, including all 23 originally above-tolerance errors. Mean cost is 957.8 records, with 2486.6 identities still unknown; the 23-case subgroup uses 889.0 records with 2574.0 unknown. These outcomes concern the same frozen evaluator used for discovery. Both Joint and Cutoff guarantee correct stopping; refinement reduces the evidence required for that guarantee.

Theory-predicted local reranking. Ridge supplies a base order and LightGBM reranks within consecutive width-2/4/8 buckets. Keeping all 33 dates, four pools and the original top-60 gains gives 396 constrained-reranking settings. The order intervention keeps the trained scorers and their raw predictions fixed. At the same additional 25% record budget, Block certifies 67, 28 and six of the 132 cases at widths 2, 4 and 8; Cutoff and Cutoff+span certify five, two and zero. At 50%, the counts are 123/93/60 versus 78/17/6. This isolates the value of feasible boundary states beyond a width-based gain cap. All 1,584 evidence states have independently rescored witnesses attaining both exact endpoints. Appendix A gives the protocol, all primary counts, synthetic exactness and work-scaling checks.

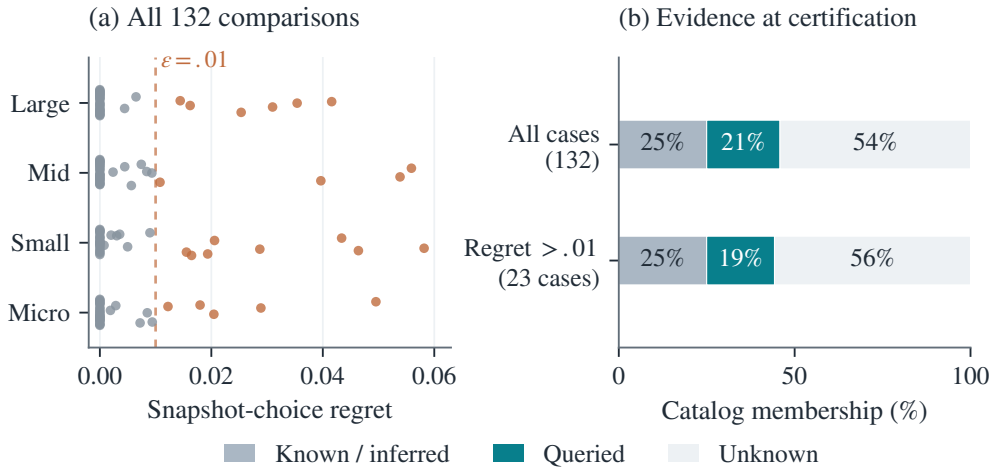


Figure 2: Discovery and repair use the same frozen stock evaluator. Left: all 33 dates in each pool; the dashed line is regret .01 and orange points exceed it. Right: average membership-information composition at the Cutoff stopping point, for all 132 cases and the 23 above-tolerance cases. More than half the identities remain unknown.

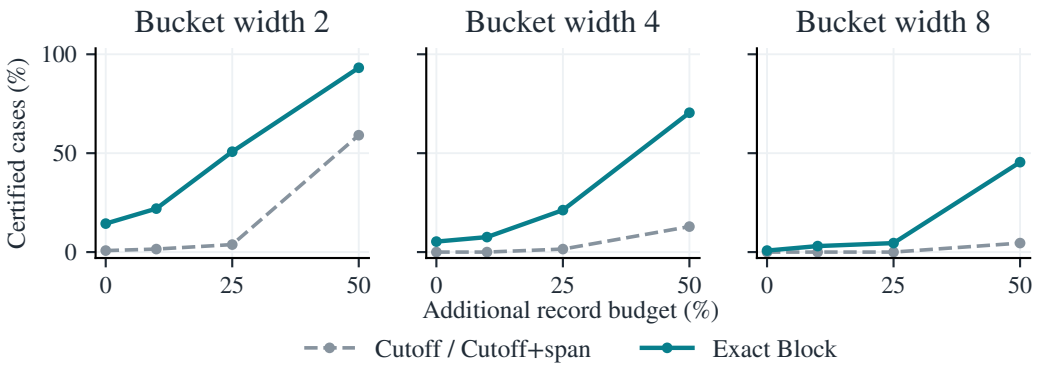


Figure 3: Constrained reranking at $\epsilon = .01$, with 132 pool-date comparisons per width. All methods receive the same records. Cutoff and Cutoff+span have identical certification counts here. Lines connect the four measured budgets, beyond the initial 25% known-record prefix.

Exactness, attainment and structural work. The independent boundary-block grid contains 120 families with signed gains, one or two count groups, and imposed block caps from one to six. Ranks and gains are generated without using realized memberships. Exhaustive enumeration agrees with all 360 endpoint pairs across three evidence levels, to maximum absolute error 1.12×10^{-16}. Thirty further checks cover evidence for which query admissibility is not implied by counts. The real-score checks construct complete membership witnesses attaining both endpoints and verify containment of the realized contrast.

At fixed widths two, four and eight, scaling from 64 to 4,096 identities preserves assignment work n, $2n$ and $16n$. A 27-case paired CPU comparison also checks both endpoints against a general exact MILP. At $n = 64$, median Block times are .040–.190 ms across the three widths, versus 751–781 ms for the MILP; all solutions agree (Appendix O.2). The natural-order diagnostic in Appendix O.1 determines where this local regime applies. Separately, exact decision trees verify $C_* = 1$, $D_* = n/4$ at four sizes, distinguishing computation from evidence search.

Table 3: Mean paid membership records by query policy and certificate. Frontier versus Contrast tests policy; Joint versus Cutoff uses the same Contrast sequence. Stock uses $\epsilon = .01$, MovieLens $\epsilon = 0$. Parentheses count evaluation instances; the MovieLens rows distinguish fixed and validation-selected training budgets.

Evaluation setting	Frontier / Joint	Contrast / Joint	Contrast / Cutoff
A-share mean percentile (264)	999.9	987.0	982.8
A-share Recall (132)	814.0	774.0	774.0
MovieLens, fixed-budget (4)	2528.0	1952.0	1952.0
MovieLens, validation-budget (4)	3104.0	1456.0	1456.0

Table 4: Evidence costs at $\epsilon = .01$. Exact stopping uses enumerated regret on the same fixed record order. Parentheses count instances; Hong Kong uses $K = 2$.

Small instances	Marginal	Joint	Cutoff	Exact stopping
Synthetic (144)	6.604	3.535	3.528	3.250
HK slices (34)	8.147	7.412	7.147	7.000

Same evidence, different certificates. We replay all 400 Contrast trajectories, comparing Marginal, Joint, and Cutoff at identical evidence states. All 6890 saved states were checked for containment of the realized evaluation-value differences and monotonicity; 800 model-pair/evaluation-query cases also used a separate cutoff optimizer, beyond the full small-instance checks. For stock mean percentile at .01, Cutoff reduces mean record cost by 0.43%, improving 24/264 cases under the original 64-record batches. Replaying the same requested order one record at a time, applying free count implications equally, gives 0.55% saving and 67/264 improvements. The matched query order isolates the contribution of the stopping certificate. For fixed-budget recommendation, record savings arise from acquisition policy; Joint and Cutoff have identical stopping costs. Local paired measurements give median bound-refresh time ratios of 1.39 for stock and 1.23 for recommendation; Appendix O.3 adds cumulative computation and a per-record break-even threshold. The order-gain control in Eq. (28), tested on all 400 original trajectories at all six tolerances, never changes Joint stopping record count. The observed Cutoff improvements therefore persist relative to this order-gain control.

Exact references and controlled mechanism. On 144 synthetic and 34 real-score Hong Kong small-problem trajectories, enumeration gives exact regret at every shared hash-order prefix. These are exact stopping references for that fixed record order (Table 4). A separate 32-case mechanism experiment holds catalog size, missing fraction and the realized model gap at 0.1 while changing order disagreement and known-record location. Known anchor contributions are explicitly calibrated for this synthetic control. Perturbed, bounded-gain constructions separately verify Proposition 7. An offline audit on all 288 settings of the original 72-family grid additionally separates intrinsic evidence cost from relaxation cost: the exact, Cutoff and Joint minima average 1.118, 1.312 and 1.396 records. Cutoff matches the exact minimum in 91/96 local-order and 52/96 independent-order settings. These hindsight optima quantify intrinsic evidence requirements alongside online acquisition costs (Appendix N).

Training quality and policy controls. Recommendation budgets are selected by Recall on the calendar month immediately before testing, using strictly earlier training data and a 60/120/240/480/960/1920-step grid. Models are refitted on all pre-test history using the selected budgets. All six-model catalogs are reevaluated under the four policies. At $\epsilon = 0$, Contrast saves 53.1% against Frontier; the 60-step model set gives 22.8%. Two further identity-hash initial-record sets give 51.3% and 52.6% savings on the same full catalogs, with lower cost in all 12 paired month/initial-set comparisons (Appendix O.4). With validation-selected training budgets, Contrast certifies all four MovieLens catalogs using an additional-record budget of 20% of catalog size; Frontier certifies none

(Figure 4). Costs are 7.5–17.5% for Contrast versus 26.0–29.4% for Frontier, with identical initial evidence and the Joint certificate.

External event and evidence scope. A public September 2024 Hong Kong adjustment (China Investment Information Services Limited, 2024) defines a 66-stock cohort. On 17 evaluation dates, stale membership selects a suboptimal model on 0/5 pre-event and 10/12 post-event dates. On Contrast's 34 configurations, Cutoff saves 0.19% of records relative to Joint at $\epsilon = .01$. This event-selected test evaluates a documented membership transition. Further checks remove exact overlap or allow count intervals, deriving implications only from the weaker evidence. Matched selection-fraction comparisons retain all 33 stock dates and four pools. Appendix L gives the protocols; the appendices also report cost decomposition and bounded lookahead.

Pool differences depend on the selection task. At $K = 60$, Large/Mid/Small/Micro have 8/10/14/10 winner reversals; selecting 10% instead gives 8/11/8/4. Changing selection depth while holding predictions and labels fixed changes the pattern across pools. The task is therefore defined jointly by membership, model orders and selection cutoff, not by pool identity alone.

A fixed winner does not explain the repair. Motivated by the static-ranking control of Singh et al. (2026), we give each fixed-choice rule hindsight access to all 132 stock outcomes. Even the fixed model minimizing above-tolerance errors incurs 93/132; allowing a different fixed model per pool has the same minimum. Snapshot choice incurs 23/132, and the certified decisions incur zero. Thus neither a global nor a pool-specific fixed preference explains the repair. These hindsight minima isolate the static-choice explanation; equal-information costs are evaluated by the matched comparisons above. Appendix N reports all seven models.

8 Related Work

The PLENCH benchmark studies model selection with ambiguous validation labels (Wang et al., 2025b). Transferability benchmarking tests whether static model hierarchies explain apparent task-specific selection gains (Singh et al., 2026). Related protocol effects arise from candidate sampling (Krichene & Rendle, 2020; Cañamares & Castells, 2020), dynamic availability (Rappaz et al., 2021) and look-ahead benchmark bias (Daniel et al., 2009).

Acceptable model choices form overlapping decision regions (Javdani et al., 2014). Certificate and decision-tree complexity already distinguish verifying a realized answer from finding it online (Buhrman & de Wolf, 2002; Grossman et al., 2020); Bayesian active learning gives prior-dependent policy guarantees (Golovin et al., 2010). Our specialization is the shared-state top-K family, computable rank-based enclosures and their evidence costs.

Active Testing acquires labels to estimate model risk efficiently (Kossen et al., 2021); BanditEval allocates method–sample evaluations to identify strong configurations (Zhou et al., 2025). Our observations are shared membership bits, and the guarantee covers every compatible candidate state. Minimax-regret winner selection and regret-driven elicitation are established in voting (Lu & Boutilier, 2011); here shared memberships replace independently completed preference orders. Explorable uncertainty also studies costly revelation of input values (Erlebach et al., 2023). Uncertain-database methods address top-K query processing (Yi et al., 2008) and bounds on ranking-query answers (Feng et al., 2023). Our contribution is exact boundary-state completion and computable bounds on common-state model contrasts.

9 Discussion and Conclusion

Shared ranking structure can certify model choice before all candidate memberships are resolved. A single boundary block gives an exact characterization for local reranking; joint disagreement and compatible cutoffs give reliable bounds for general orders. The theory distinguishes computational tractability from evidence search, and frozen-score interventions show when preserving that structure certifies more decisions from the same records.

AI USE STATEMENT

We used GPT-5.6 Sol and GPT-6 Astra to assist with code development, including drafting and revising implementation code, debugging, and preparing validation scripts. The models also assisted with drafting and polishing selected passages of the manuscript. This writing support focused on grammar, clarity, sentence flow, and consistency of terminology. The authors are responsible for the final code and manuscript.

ETHICS STATEMENT

The financial results concern evaluation reliability, not investment advice. Dataset access and redistribution remain subject to their source terms. The study makes no attempt to identify MovieLens users.

REPRODUCIBILITY STATEMENT

The artifact preserves frozen predictions, fixed protocols, comparison trajectories and validation scripts. Checks include independent sorting-based evaluation, matched-query target comparisons and small-instance enumeration. The appendices give proofs, model configurations, input support and timing protocols.

REFERENCES

Ron M. Adin, Eli Bagno, and Yuval Roichman. Block decomposition of permutations and Schur-positivity. arXiv:1611.06979, version 2, 2017. URL https://arxiv.org/abs/1611.06979v2.

BaoStock. BaoStock financial data interface, n.d. URL https://www.baostock.com/. Data-provider website. Accessed September 11, 2026.

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.

Rocío Cañamares and Pablo Castells. On target item sampling in offline recommender system evaluation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 259–268, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3383313.3412259. URL https://doi.org/10.1145/3383313.3412259.

China Investment Information Services Limited. Notice on the adjustment of southbound eligible securities under shanghai connect. Notice dated September 9, 2024, hosted by Shanghai Stock Exchange. English translation of the Chinese title, 2024. URL https://www.sse.com.cn/services/hkexsc/disclo/announ/c/c_20240909_10761390.shtml.

Gilles Daniel, Didier Sornette, and Peter Woehrmann. Look-ahead benchmark bias in portfolio performance evaluation. *The Journal of Portfolio Management*, 36(1):121–130, 2009. doi: 10.3905/JPM.2009.36.1.121. URL https://doi.org/10.3905/JPM.2009.36.1.121.

Thomas Erlebach, Murilo de Lima, Nicole Megow, and Jens Schlöter. Sorting and hypergraph orientation under uncertainty with predictions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 5577–5585, 2023. doi: 10.24963/ijcai.2023/619. URL https://www.ijcai.org/proceedings/2023/619.

Jinyong Fan and Yanyan Shen. StockMixer: A simple yet strong MLP-based architecture for stock price forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8389–8397, 2024. doi: 10.1609/aaai.v38i8.28681. URL https://ojs.aaai.org/index.php/AAAI/article/view/28681.

Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998. doi: 10.1145/285055.285059. URL https://doi.org/10.1145/285055.285059.

Su Feng, Boris Glavic, and Oliver Kennedy. Efficient approximation of certain and possible answers for ranking and window queries over uncertain data. *Proceedings of the VLDB Endowment*, 16(6): 1346–1358, 2023. doi: 10.14778/3583140.3583151. URL https://www.vldb.org/pvldb/vol16/p1346-feng.pdf.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, volume 23, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/1e6e0a04d20f50967c64dac2d639a577-Abstract.html.

Tomer Grossman, Ilan Komargodski, and Moni Naor. Instance complexity and unlabeled certificates in the decision tree model. In *11th Innovations in Theoretical Computer Science Conference*, volume 151 of *Leibniz International Proceedings in Informatics*, pp. 56:1–56:38. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.ITCS.2020.56. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2020.56.

GroupLens. MovieLens 25M dataset. Stable dataset release, December 2019, 2019. URL https://grouplens.org/datasets/movielens/25m/.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020. doi: 10.1145/3397271.3401063. URL https://arxiv.org/abs/2002.02126.

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. URL https://proceedings.mlr.press/v33/javdani14.html.

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, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1748–1757, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3394486.3403226. URL https://doi.org/10.1145/3394486.3403226.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):162–170, 2024. doi: 10.1609/aaai.v38i1.27767. URL https://ojs.aaai.org/index.php/AAAI/article/view/27767.

Tyler Lu and Craig Boutilier. Robust approximation and incremental elicitation in voting protocols. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 287–293. AAAI Press, 2011. URL https://www.ijcai.org/Proceedings/11/Papers/058.pdf.

Jérémie Rappaz, Julian McAuley, and Karl Aberer. Recommendation on live-streaming platforms: Dynamic availability and repeat consumption. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 390–399, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3460231.3474267. URL https://doi.org/10.1145/3460231.3474267.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009. URL https://www.auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf.

Prabhant Singh, Sibylle Hess, and Joaquin Vanschoren. How NOT to benchmark your SITE metric: Beyond static leaderboards and towards realistic evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f0ebc318e2df08360b2df559e81602e5-Abstract-Conference.html.

Mengyu Wang, Tiejun Ma, and Shay B. Cohen. Pre-training time series models with stock data customization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3019–3030, New York, NY, USA, 2025a. Association for Computing Machinery. doi: 10.1145/3711896.3737005. URL https://doi.org/10.1145/3711896.3737005.

Wei Wang, Dong-Dong Wu, Jindong Wang, Gang Niu, Min-Ling Zhang, and Masashi Sugiyama. Realistic evaluation of deep partial-label learning algorithms. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d5b7afe24b3b891138105a5877b7decd-Abstract-Conference.html.

Ke Yi, Feifei Li, George Kollios, and Divesh Srivastava. Efficient processing of top-k queries in uncertain databases. In *Proceedings of the 24th IEEE International Conference on Data Engineering*, pp. 1406–1408. IEEE, 2008. doi: 10.1109/ICDE.2008.4497571. URL https://doi.org/10.1109/ICDE.2008.4497571.

Jin Zhou, Christian Belardi, Ruihan Wu, Travis Zhang, Carla Gomes, Wen Sun, and Kilian Weinberger. On speeding up language model evaluation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a38cfba4c717e5a37b200294615e546e-Abstract-Conference.html.

A Boundary-Block Proofs and Validation

A.1 Decomposition and exact completion

After deleting known-absent eligible candidates, both orders permute the same retained set. Map each candidate in order a to its one-based position in b. A prefix is shared exactly when its length equals the largest mapped position within it. Cutting there gives the standard direct-sum decomposition (Adin et al., 2017) in linear time. Singleton blocks are allowed, and the block sequence is shared.

Fix an admissible membership z. Let B be the first block whose cumulative member count reaches K, with $x = B \cap C(z)$ and $t = |x|$. Every member in an earlier block is selected under both orders; candidates in later blocks are unselected. Both models select r candidates inside B, possibly with different identities. If $r = t$, the contrast is zero; otherwise $1 \le r \le t-1$ and the contrast is $\Delta_B(x, r)$. This proves the forward direction of Theorem 2 and invariance to adding a common constant to a block's gains.

For the reverse direction, fix B, x and a feasible r. In group g, the remaining target $c_g - x_g$ must be distributed between P and S. Exactly p_g prefix members are possible iff $m_{Pg} \le p_g \le a_{Pg}$ and $m_{Sg} \le c_g - x_g - p_g \le a_{Sg}$. Their intersection is precisely Eq. (3). Mandatory identities are retained, and independent optional bits fill every integer in that interval. Disjoint groups imply that every integer between the sums of the lower and upper endpoints is attainable. Equation (4) therefore completes x to a state with $K - r$ prefix members and more than r block members. This state is automatically query-admissible and attains $\Delta_B(x, r)$. The suffix includes ineligible identities, which must still satisfy the membership counts.

Zero is added for a complete-block boundary exactly when its prefix can contain K members. Let $m_{\Omega g}, a_{\Omega g}$ count mandatory and available candidates in group g over the full catalog. Feasible prefix counts have endpoints $\max\{m_{Pg}, c_g - (a_{\Omega g} - a_{Pg})\}$ and $\min\{a_{Pg}, c_g - (m_{\Omega g} - m_{Pg})\}$. Their sums must bracket K. A partial block may also attain zero when its two selections have equal gain. Zero is not added unconditionally: two mandatory candidates with reversed top-1 orders and gains $(1, 0)$ have exact interval $[1, 1]$.

The preceding argument also applies when some count-feasible states have fewer than K eligible identities. Those states never produce a feasible partial or complete boundary. For this single-query result, automatic admissibility of the entire count family is therefore unnecessary.

A.2 Implementation, work and inexpensive control

The implementation scans group counts over the retained blocks. For each block it enumerates 2^{u_B} assignments, respecting mandatory candidates. Each assignment takes $O(b_B + G)$ work for group intervals, ordered gain-prefix sums and feasible selection sizes. The catalog scan and count updates give Eq. (5). Reusing arrays gives $O(n + G + \max_B b_B)$ memory. Blocks that cannot contain a partial boundary are skipped. Removing known-absent candidates can split blocks, but the exact interval remains monotone because the feasible membership family shrinks.

Every reported endpoint stores its boundary block, local assignment and selected size. To reconstruct a full witness, initialize each group's prefix count at its lower endpoint, distribute the remaining required prefix members within the upper capacities, and fill the suffix to its residual target. Mandatory bits are fixed throughout. Thus each endpoint is attained by a complete count-feasible membership state.

For the span control, if the two local selections differ in m identities on each side, then $m \le \min(r, t-r) \le \lfloor b_B/2 \rfloor$. Pairing the exchanged identities bounds the absolute contrast by $m(\max_{i \in B} g_i - \min_{i \in B} g_i)$. Maximizing over the possible boundary blocks gives the stated symmetric bound. Its intersection with Cutoff is the inexpensive baseline, denoted Cutoff+span.

A.3 Recognizing reachable boundary blocks before enumeration

The current evidence identifies the tractable regime without enumerating local assignments or observing new answers. For a block B, let m_{Bg}, a_{Bg} count its mandatory and available identities; keep the prefix/suffix notation of Eq. (3). Without fixing the block assignment, its group's feasible

prefix counts range from $L_g = \max\{m_{Pg}, c_g - a_{Bg} - a_{Sg}\}$ to $U_g = \min\{a_{Pg}, c_g - m_{Bg} - m_{Sg}\}$. Define

$$L = \sum_g L_g, \qquad H = \sum_g \min\{L_g + a_{Bg}, c_g - m_{Sg}\},$$
$$J = \sum_g \max\{0, \min(U_g, c_g - m_{Sg} - a_{Bg}) - L_g\}. \tag{7}$$

Then B can contain a partial boundary if and only if

$$L \le K - 1 \quad \text{and} \quad H + \min\{K - 1 - L, J\} \ge K + 1. \tag{8}$$

To prove the test, fix a feasible prefix count p_g. The largest combined prefix-plus-block count is $\min\{p_g + a_{Bg}, c_g - m_{Sg}\}$, attained by filling available block identities before optional suffix identities. This function has slopes one then zero, with $\max\{0, \min(U_g, c_g - m_{Sg} - a_{Bg}) - L_g\}$ positive increments. Starting from L_g in each group, at most $K - 1 - L$ additional prefix slots are allowed. Their optimal allocation gives the second expression in Eq. (8). A prefix with at most $K - 1$ members and a prefix-plus-block with at least $K + 1$ members is precisely a partial boundary. Every integer allocation is attainable under the disjoint count constraints, including when the suffix contains ineligible identities.

All tests take $O(nG)$ total work using prefix counts. Let $\mathcal{B}_\mathcal{E}$ contain the blocks passing Eq. (8) and $u_\mathcal{E} = \max_{B \in \mathcal{B}_\mathcal{E}} u_B$, with zero for an empty set. Skipping the other blocks preserves the exact endpoints and sharpens Eq. (5) to $O(nG + \sum_{B \in \mathcal{B}_\mathcal{E}} 2^{u_B}(b_B + G))$. Thus $u_\mathcal{E}$ provides a directly checkable computation budget: a prescribed cap permits local enumeration; larger values call for Joint/Cutoff. This computation choice preserves the stopping guarantee in both regimes.

A.4 Adjacent swaps and matching search bounds

A nontrivial size-two block has orders (i, j) and (j, i). It can change selections only when both candidates are members and exactly one is selected. The earlier prefix must then contain $K - 1$ members; Eq. (3) with both bits fixed to one tests reachability. The resulting contrast is $g_i - g_j$, which is zero if the gains coincide. Taking the largest reachable directional gain gives Eq. (6). Self-comparison supplies zero in regret even when no state has zero pairwise contrast.

For the search separation, fix $0 \le \epsilon < 1$ and integer $h \ge 1$. Take $2h$ size-two swap blocks. The first h have gains $(1, 0)$ and the last h have gains $(0, 1)$ in order a. Exactly two of the $n = 4h$ identities are members and $K = 1$. If both belong to one block, its orientation determines the strict winner with gap one. If they belong to different blocks, both models select the same earlier member and tie. Initially both strict winners are possible.

At any realized state, querying one actual member rules out every complete-member block other than its own. The model favored by that block is acceptable on every compatible state, including cross-block ties. Hence the offline certificate minimum is exactly one. Online, query one representative from each of the h positive-direction blocks. A positive answer certifies a; h negative answers eliminate every positive-direction complete block and certify b. This gives depth at most h. After fewer than h all-negative answers, at least one untouched block in each direction remains. Each complete pair defines a compatible state, so no model is certified. This proves the matching lower bound $D_* = h = n/4$. Exact Bellman enumeration verifies $n = 4, 8, 12, 16$; the offline minimum is checked at every state of each family.

A.5 A cross-block incompatibility example

Let orders be $(0, 1, 2, 3, 4, 5, 6, 7)$ and $(0, 1, 3, 2, 5, 4, 7, 6)$, with $K = 3$ and total count four. For $\lambda \in [0, 1]$, set gains to $(0, 0, \lambda, \lambda, 0, 0, \lambda, \lambda)/3$. The three swap blocks have constant gains internally, so every real contrast is zero. Yet the first one-order envelope can exchange identity 2 against 5, while the second can exchange 6 against 5. Each gives $\lambda/3$, using different blocks in the two optimizers. Cutoff returns $[-\lambda/3, \lambda/3]$, whereas Block returns $[0, 0]$. Enumeration checks all 70 memberships for each $\lambda \in \{0, .25, .5, 1\}$. The inexpensive block-span control also returns zero here. This example isolates the single-block constraint; the real-score intervention tests exact local enumeration against the span control.

A.6 INDEPENDENT VALIDATION AND FROZEN-SCORE INTERVENTION

The prespecified validation grid contains 120 synthetic families: catalog sizes 8/12, imposed block caps 1/2/3/4/6, top-K values 1/3, one/two count groups and three generator instances distinct from the earlier cost grid. Both ranks and signed gains are generated independently of the realized membership. At 0/25/50% known records, exhaustive membership enumeration checks 360 exact intervals; maximum endpoint error is 1.12×10^{-16}. Thirty additional checks cover query admissibility not implied by counts. Unit tests cover endpoint witnesses, full selection, gain offsets, and explicit rejection when an active block exceeds the enumeration cap or the caller requests multi-query exactness.

The real-score intervention uses the first two models in the declared stock model list: Ridge defines the base order; LightGBM reranks only within its consecutive width-2/4/8 buckets. The intervention constrains the frozen orders while preserving the trained scorers. All 33 dates and four pools, original outcomes, top-60 metric, known count/overlap, and initial 25% hash-prefix records are retained. We then reveal fixed additional 0/10/25/50% catalog-size record batches, identically for all certificates. Counts propagate free implications after each batch. All 396 settings and 1,584 evidence states are retained. The full memberships attaining both Block endpoints are independently rescored at every state: 3,168 witness checks, alongside containment and monotonicity.

Table 5: Certified constrained-reranking comparisons at $\epsilon = .01$. Every row contains all 132 dates/pools. Budget is additional paid records as a fraction of catalog size, beyond the initial 25%.

Width	Budget (%)	Cutoff	Cutoff+span	Block
2	0	1	1	19
2	10	2	2	29
2	25	5	5	67
2	50	78	78	123
4	0	0	0	7
4	10	0	0	10
4	25	2	2	28
4	50	17	17	93
8	0	0	0	1
8	10	0	0	4
8	25	0	0	6
8	50	6	6	60

All five declared tolerances $0/.001/.005/.01/.025$ remain in the artifact. The primary table uses the original stock tolerance. These interventions test the structural regime predicted by the theorem; the original seven-model discovery and recommendation experiments retain their own unrestricted orders. Appendix O.1 reports their natural block structure and the resulting computation regime.

With reversed within-block orders and all memberships unknown, scaling checks use $n = 64, 256, 1024, 4096$ and widths 2/4/8. For these constructed families, local assignment counts are respectively n, $2n$ and $16n$, and every endpoint has an independently scored witness. This experiment isolates dependence on block width; Appendix O.2 adds a paired CPU comparison with a general exact single-query reference.

B CERTIFICATES FROM JOINT SELECTION DISAGREEMENTS

For one evaluation query and pair a, b, let $d_i(z) = \mathbf{1}\{i \in S_a(z)\} - \mathbf{1}\{i \in S_b(z)\}$. Then $Q_a - Q_b = \sum_i g_i d_i$, and equal selection sizes give

$$\sum_i d_i = 0. \tag{9}$$

The relevant uncertainty is whether an identity can be selected differently, not whether either model's selection is individually uncertain.

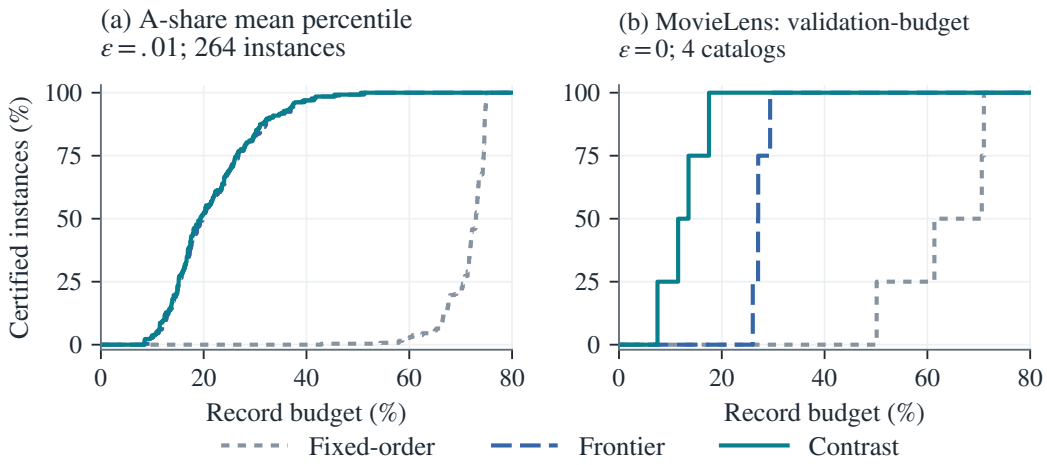


Figure 4: Membership-query policy comparison using the Joint certificate. Stock uses $\epsilon = .01$; MovieLens uses validation-selected training budgets and $\epsilon = 0$.

B.1 Exact local disagreement

Force candidate i to be a member and partition each count group by precedence in the two eligible orders: before i in both (11), only a (10), only b (01), or neither (00). Candidate i itself and query-ineligible candidates belong to cell 00. Let M_{rs}, A_{rs} count mandatory and available candidates after forcing i present. Available nonmandatory candidates are optional. Suppressing the group subscript, for target count c the number t of member a-predecessors lies in

$$L = \max\{M_{11} + M_{10}, c - A_{01} - A_{00}\},$$
$$U = \min\{A_{11} + A_{10}, c - M_{01} - M_{00}\}. \tag{10}$$

For feasible t, its largest possible b-predecessor count is

$$f(t) = \min\{A_{11}, t - M_{10}\} + \min\{A_{01}, c - t - M_{00}\}. \tag{11}$$

Theorem 4 (Exact directional disagreement)**.** *Let* $I_g = \max\{0, \min(U_g, A_{11,g} + M_{10,g}, c_g - M_{00,g} - A_{01,g}) - L_g\}$. *Assume automatic query admissibility, that i is eligible and can be a member, that all group ranges are nonempty, and that* $\beta = K - 1 - \sum_g L_g \geq 0$. *Here* L_g, U_g, f_g *denote the preceding quantities for group g. Then* $d_i = +1$ *is reachable exactly when*

$$\sum_g f_g(L_g) + \min\{\beta, \sum_g I_g\} \geq K. \tag{12}$$

Swapping the orders gives reachability of $d_i = -1$.

The profile f_g is discretely concave with slopes $+1, 0, -1$. From L_g, its number of positive increments is I_g. Allocating at most β remaining predecessor slots gives Eq. (12). Together with forced selection, these directions give exact coordinate extrema ℓ_i, u_i of d_i. Rank sweeps compute predecessor intersections in $O(Gn \log n)$ time per pair. Identical selection behavior gives $\ell_i = u_i = 0$, eliminating shared uncertainty before aggregation.

B.2 Conservation-aware aggregation and certified choice

Compute the extrema of

$$\sum_i g_i d_i \quad \text{subject to} \quad \ell_i \leq d_i \leq u_i, \qquad \sum_i d_i = 0. \tag{13}$$

Write $\ell = (\ell_i)_i$ and $u = (u_i)_i$. Starting at ℓ, allocate $-\sum_i \ell_i$ units of capacity $u_i - \ell_i$ in descending gain order for the maximum and ascending order for the minimum. This solves the relaxation exactly. Intersect it with the separate model-value bounds (Marginal) and aggregate across queries using the common weights to obtain $[L_{ab}, U_{ab}]$. This certificate is Joint (J).

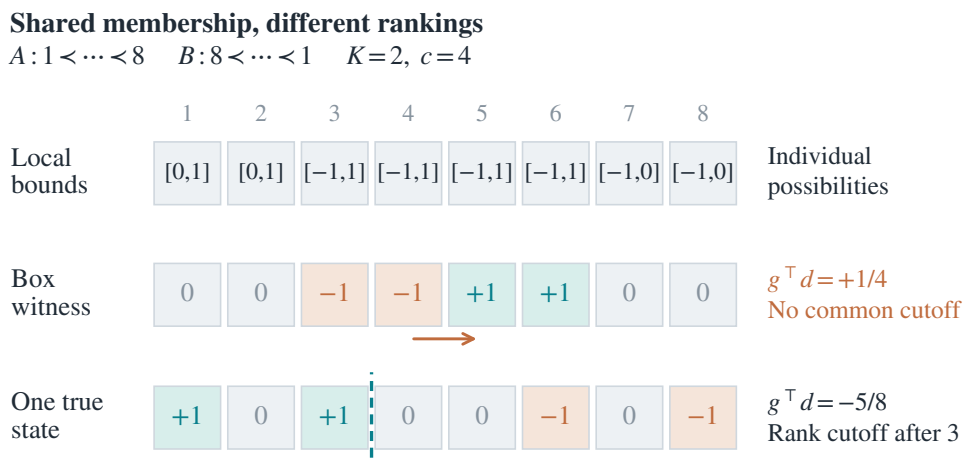


Figure 5: A local possibility need not belong to one common ranking state. For candidates $i = 1, \ldots, 8$, cutoff $K = 2$, total count $c = 4$ and gains $g_i = i/16$, the box witness satisfies all coordinate ranges and zero sum, yet puts negative signs before positive signs in A's order. No shared cutoff realizes it. A true state and its compatible signs are shown below. The illustration explains the relaxation gap.

Theorem 5 (Sound, monotone comparison certificates)**.** *Every feasible common-state contrast lies in* $[L_{ab}, U_{ab}]$. *The enclosure is no wider than the independent-model cardinality enclosure and cannot widen when valid records are added. Consequently,*

$$\overline{R}(a) = \max\{0, \max_b U_{ba}\} \tag{14}$$

is a valid bound in Eq. (2)*.*

These are computable global enclosures built from exact local reachability. The local relaxation and query-wise aggregation provide outer bounds on global endpoints. Their benefit is sound model choice with polynomial-time updates, including zero-record certification in Proposition 1.

C Enforcing a Shared Ranking Cutoff

Exact coordinate extrema need not be jointly realizable. The box can choose signs that contradict the models' orders: if $d_i = 1$ and $d_j = -1$, then i precedes j in order a, and j precedes i in order b. Thus one cutoff in a separates all positive differences from all negative ones. Also $\sum_i (d_i)_+ \le K$, where $(x)_+ = \max\{x, 0\}$, in addition to $\sum_i d_i = 0$.

A concrete obstruction. In Figure 5, the coordinate box accepts $d = (0, 0, -1, -1, 1, 1, 0, 0)$ and assigns $g^\top d = 1/4$. But A selects the two smallest active identities and B the two largest. Among all $\binom{8}{4} = 70$ states, $Q_A - Q_B$ is between $-3/4$ and $-1/4$. The positive endpoint combines locally feasible signs that no common membership can realize. Cutoff's upper bound is zero and certifies B, even though neither model's evaluation value is identified.

The cutoff family. For one evaluation query, let $r_a(i) \in \{1, \ldots, |E_u|\}$ be i's position in a's eligible order. Ineligible coordinates have $\ell_i = u_i = 0$ and can be omitted. For cutoff $t \in \{0, \ldots, |E_u|\}$, define the following family, applying rank restrictions only to eligible coordinates:

$$\begin{aligned} \mathcal{P}_a(t) = \{d \in \mathbb{Z}^n : \ell \le d \le u, \quad \textstyle\sum_i d_i = 0, \quad \sum_i (d_i)_+ \le K, \\ d_i \ge 0 \text{ if } r_a(i) \le t, \quad d_i \le 0 \text{ if } r_a(i) > t\}. \end{aligned} \tag{15}$$

Let $\mathcal{P}_a = \bigcup_t \mathcal{P}_a(t)$. Empty components are skipped; consistency of the original evidence guarantees at least one nonempty component. Every true difference belongs to $\mathcal{P}_a$; its negative belongs to the analogous $\mathcal{P}_b$ with reversed bounds.

Theorem 6 (Exact one-order cutoff envelope)**.** *At a compatible cutoff, let p, q count forced positive and negative coordinates, and C be their signed gain sum. Let A_j sum the j largest optional positive gains in the prefix, and B_j the j smallest optional negative gains in the suffix. The exact maximum at that cutoff is*

$$C + \max_m \{A_{m-p} - B_{m-q}\}, \qquad \max(p, q) \le m \le \min(K, p + n_+, q + n_-), \tag{16}$$

where n_+, n_- count optional positive-prefix and negative-suffix coordinates, respectively, and $A_0 = B_0 = 0$. The maximum ranges over integers; an empty range means an incompatible cutoff. After removing coordinates with $\ell_i = u_i = 0$, let s candidates remain. Maximizing over cutoffs takes $O(n + sK)$ time and $O(n + (s+1)K)$ working memory, including the support filter and best-K lists. If $s = 0$, the contrast is zero.

Each allocation pairs m positive and negative differences. Forced signs leave $m - p$ positive and $m - q$ negative slots; gain ordering optimizes the optional exchanges. Maximizing over feasible cardinalities and cutoffs covers the one-order union.

Write $F_a(g; \ell, u) = \max_{d \in \mathcal{P}_a} g^\top d$, with $g = (g_i)_i$. Negating gains gives the lower endpoint. Restoring query indices, let $g_u = (g_{ui})_i$, and let ℓ_u^{ab}, u_u^{ab} be the coordinate-bound vectors for query u and pair (a, b). Define

$$\begin{aligned} H_u^{ab} &= \min\{F_{ua}(g_u; \ell_u^{ab}, u_u^{ab}), F_{ub}(-g_u; -u_u^{ab}, -\ell_u^{ab})\}, \\ U_{ab}^{\mathrm{cut}} &= \min\{U_{ab}^{\mathsf{J}}, \textstyle\sum_u w_u H_u^{ab}\}, \qquad L_{ab}^{\mathrm{cut}} = -U_{ba}^{\mathrm{cut}}. \end{aligned} \tag{17}$$

The shared-cutoff certificate (Cutoff) is sound, never wider than Joint, and monotone under valid new evidence: the coordinate boxes shrink while the cutoff families remain fixed. Its regret bound is $\overline{R}_{\mathrm{cut}}(a) = \max(0, \max_b U_{ba}^{\mathrm{cut}})$. The two one-order optimizations form an outer relaxation of common-membership optimization; their optima need not use the same pair of cutoffs.

Decision-relevant contraction. For a fixed proposed model a, only challengers in $\mathcal{C}_a = \{b \neq a : U_{ba}^{\mathsf{J}} > \epsilon\}$ can prevent certification. Since $U_{ba}^{\mathrm{cut}} \le U_{ba}^{\mathsf{J}}$,

$$\overline{R}_{\mathrm{cut}}(a) \le \epsilon \quad \Longleftrightarrow \quad \max\big(\{0\} \cup \{U_{ba}^{\mathrm{cut}} : b \in \mathcal{C}_a\}\big) \le \epsilon. \tag{18}$$

Thus reducing many nonbinding pair intervals can leave the decision unchanged. The criterion also separates two useful diagnostics: how much uncertainty a relaxation removes, and whether it resolves the comparisons that currently prevent a valid choice.

Proposition 7 (Robust separation without identical orders)**.** *Index $n = 4K$ candidates by $i = 1, \ldots, n$, with K divisible by four. All memberships are unknown and exactly $2K$ candidates are members. Models A, B use ascending and descending identity orders. For $\eta \ge 0$ and*

$$g_i = \frac{i/n + \eta + \xi_i}{K(1 + 2\eta)}, \qquad |\xi_i| \le \eta,$$

gains lie in $[0, 1/K]$, and $Q_B - Q_A \ge (1/4 - 2\eta)/(1 + 2\eta)$ in every state. The cutoff certificate needs zero records when $\epsilon \ge 2\eta/(1 + 2\eta)$. For $(\eta, \epsilon) = (0, 0)$ or $(0.01, 0.025)$, the Joint bound, even intersected with marginal cardinality bounds, needs more than $\lfloor K/64 \rfloor$ records.

The gap is fixed rather than vanishing with catalog size, and the gain perturbation need not preserve its monotone order. Thus the separation does not require identical model orders or exact cross-query cancellation. Appendix K proves the result, including the linear lower bound. The structural property is a compatible sign cutoff, not low missingness alone.

An inexpensive order-gain control. An $O(n)$ control bounds each positive–negative exchange by its gain and rank compatibility (Appendix M). It also certifies the monotone/near-monotone families above. Cutoff additionally retains the number, identity and coexistence of exchanges, and is no weaker after both bounds are intersected with Joint. Testing this control on identical real-data prefixes isolates the contribution of cutoff compatibility.

D LOCAL DISAGREEMENT: ASSUMPTIONS AND PROOF

D.1 COUNT CLOSURE AND ADMISSIBILITY

Within a group with target c, known positive count h and v unknown bits, feasibility requires $0 \le c - h \le v$. If $c - h = 0$, every remaining bit is negative; if $c - h = v$, every remaining bit is positive. Otherwise each unknown bit can take either value under the independent count constraint.

For $A \subseteq \Omega$, let $M_g(A), A_g(A)$ count mandatory and available candidates in the intersection of A with group g. Complements are taken in Ω. The smallest eligible count in group g is $\max\{M_g(E_u), c_g - A_g(E_u^c)\}$: mandatory eligible candidates must be included, and any deficit after filling from ineligible candidates must also be eligible. Optional candidates attain this minimum. Query admissibility therefore adds no information exactly when

$$\sum_g \max\{M_g(E_u), c_g - A_g(E_u^c)\} \ge K_u \quad \text{for every } u. \tag{19}$$

The Joint/Cutoff benchmark cases satisfy this condition. If it fails, the local calculations describe a relaxed count family and provide safe outer bounds for the original admissible family. Exactness of individual coordinate extrema is then asserted only for that relaxation.

D.2 DIRECTIONAL RANK FEASIBILITY

For a forced-present, query-eligible identity i, each group splits into four disjoint predecessor cells. Exactly t selected a-predecessors means choosing t members from cells 11/10 and $c - t$ from 01/00. Their mandatory and available counts give Eq. (10). Every integer in this interval is attainable because individual optional memberships are independent within the group.

Among the t predecessors, maximize membership in cell 11 after satisfying mandatory cell-10 identities; this gives $\min(A_{11}, t - M_{10})$. Among the $c - t$ non-predecessors, maximize membership in cell 01 after satisfying mandatory cell-00 identities; this gives $\min(A_{01}, c - t - M_{00})$. Both choices can be realized simultaneously, proving Eq. (11).

The first term of $f(t)$ increases with slope one until it saturates, while the second is initially constant and subsequently decreases with slope one. Therefore f is discretely concave. Positive increments occur while $t < A_{11} + M_{10}$ and $t < c - M_{00} - A_{01}$, giving the stated I_g after restriction to $[L_g, U_g]$. To keep total a-predecessors below K, begin at all L_g and allocate at most $\beta = K - 1 - \sum_g L_g$ extra units. Every positive increment has value one, so the exact largest total number of b-predecessors is Eq. (12)'s left side.

A forced-present identity is selected by a but not b exactly when its a-predecessor count is at most $K - 1$ and its b-predecessor count is at least K. This proves Theorem 4; reversing orders gives the negative direction. A constant $d_i = 1$ requires i to be entailed present, always selected by a, and never selected by b. The reverse characterizes $d_i = -1$. Other coordinate extrema are -1 when the negative direction is possible, $+1$ when the positive direction is possible, and zero otherwise.

D.3 COMPUTATIONAL REALIZATION

Sort eligible identities in order a and sweep them. A Fenwick tree indexed by rank in order b maintains mandatory/available counts in each group. Querying the strict prefix before inserting the current identity gives the intersection of its two predecessor sets. Single-order prefix counts then give the remaining three cells. There are $2G$ count channels, yielding $O(Gn \log n)$ work for one pair and $O(Gn)$ working space. Count complements include query-ineligible identities from the complete catalog.

For M models and multiple evaluation queries, the per-round bound cost is $O(\sum_u M^2 Gn \log n)$, with fixed score and gain orders reusable across rounds. Implementations use a compiled rank sweep; no mixed-integer solve is needed for this certificate.

E AGGREGATION AND ACQUISITION GUARANTEES

Every actual difference vector lies inside its coordinate bounds and sums to zero, hence it is feasible for Eq. (13). Starting at the lower bounds leaves an integer budget $-\sum_i \ell_i$ and integer capacities $u_i - \ell_i$. If a maximizing allocation uses a lower-gain capacity while a higher-gain capacity remains available, exchanging a unit cannot decrease the objective. Repeated exchanges prove the greedy descending-gain optimum; the ascending order proves the minimum. This is the exact solution to the relaxation.

Independently, bound each model's selection indicators by their always- and possibly-selected status and impose their sum K. The same allocation argument gives valid marginal enclosures. Intersecting two valid contrast enclosures retains all actual contrasts. Valid new records shrink the count family, hence shrink exact coordinate ranges and both relaxations. Their intersection cannot widen. Summation with fixed nonnegative query weights preserves validity for the shared multi-query state. Finally, $R(a) = \max_b \max_z (Q_b(z) - Q_a(z)) \leq \max\{0, \max_b U_{ba}\}$, proving Theorem 5.

For acquisition, a record's influence sums contributions from candidates whose uncertain selection differences it can affect. A record i affects candidate j directly when $i = j$ or i precedes j in either model order. The complement consists of candidates preceding i in both orders. A weighted rank sweep therefore computes influence without a dense candidate-by-candidate matrix. Orient each pair by the fixed model-name order. Its centering threshold λ is the gain at the last coordinate receiving positive capacity in the descending-gain maximization of Eq. (13); if no capacity is allocated, use the median gain. This defines the policy also in the zero-allocation case.

The actual state remains feasible after every truthful response. Consequently, an optimistic top-K cutoff occurs no later than its actual cutoff. An unknown queried on that frontier is either actually selected or an actually excluded predecessor. Each is queried once. If a frontier is unresolved, it contains another unknown record; otherwise all relevant selections are fixed. This proves finite termination and a frontier-support query bound. The regret certificate validates each stopping decision for any truthful query sequence. Numerical certificates use an explicit 10^{-8} tolerance in addition to the reported ϵ.

F TARGET SEPARATION AND STRONG CERTIFICATES

Take N even, exactly $N/2$ positive memberships among N unknown identities, and two mandatory anchors, giving catalog size $n = N + 2$. The shared query is eligible only on the unknown identities, has $K = N/2$, and gives both models the same order and gains. An anchor-only top-1 query gives gain one to the anchor selected by A and zero to the anchor selected by B. Use weights $19/20$ and $1/20$, respectively. Then $Q_A(z) - Q_B(z) = 1/20$ for all feasible states: the shared query's contribution cancels. The proposed enclosure reproduces this fixed margin.

The complete selection vector identifies all $N/2$ positive unknown identities. Before $N/2$ records have been queried, there remains an unknown positive and an unknown negative that can be exchanged while preserving the count. Thus a universal selection certificate needs at least $N/2$ queries. Querying all actual positives attains the bound. This proves Proposition 1. Sizes 8, 32, 128 and 512 are evaluated in the artifact; separate small-state tests also enumerate memberships.

F.1 CERTIFICATES FOR COMPLETE SELECTION RECOVERY

With positive query weights, the full selection vector is fixed iff the joint model metric vector is fixed for every bounded additive gain assignment. For the converse, if one query/model selects identity i in one state but not another, put unit gain on that query-identity coordinate and zero elsewhere. Its metric changes by the positive query weight. This establishes equivalence.

For a realized state, let P contain all selected members and N_- all nonmembers preceding a selected cutoff in at least one eligible order. The full selection vector is fixed exactly when evidence entails presence for every candidate in P and absence for every candidate in N_-. In each group of initially unknown candidates, let p_g, n_g count realized members/nonmembers and r_g^+, r_g^- count the required

positive/negative membership facts. Under automatic admissibility, the exact offline recovery cost is

$$C_{\text{univ}} = \sum_g \min\{r_g^+ + r_g^-, p_g, n_g\}. \tag{20}$$

Achievability follows by querying required literals, all actual positives, or all actual negatives in each group. If a required positive is unqueried, every actual unknown negative must have been queried; otherwise an exchange preserves evidence and changes the decision. The symmetric argument holds for a required negative. This proves the lower bound. The formula is an offline verification benchmark; the attaining identities may be unknown to an online auditor. Comparison certification is a weaker target and can have a smaller cost, as the separation shows.

F.2 Global contrast hardness

For a Maximum Coverage instance, let candidate identities represent sets and queries represent elements. Add a mandatory zero-gain dummy and require exactly c set identities plus the dummy. Model A scores covering sets one and all other identities zero; gains are the same coverage indicators. Model B scores only the dummy one. With top-1 and equal query weights, $Q_A - Q_B$ is the fraction of elements covered by the c chosen sets. Maximizing it solves Maximum Coverage. All states are query-admissible, all scores are binary, and the count family has one group. The reduction is checked on eight independently enumerated incidence matrices.

G Datasets, Models and Evaluation Protocol

G.1 A-share inputs

The price/volume panel and security-status fields originate from BaoStock (BaoStock, n.d.). Pools are reconstructed using ranks of the market capitalization of circulating shares. The fixed cohort contains 5,072 non-Beijing stocks observed by the end of 2022. Models share a 20-trading-day window and eight causal price/volume-derived inputs. The target is twice the within-date percentile of $\text{VWAP}(t+21)/\text{VWAP}(t+1) - 1$ minus one, where VWAP is volume-weighted average price. Samples whose forward-label intervals cross a split boundary are removed. Training uses 60 dates and 207,512 labeled sequences; validation uses 12 dates and 54,177 labels. Neural models train for six supervised epochs, selecting checkpoints by validation mean squared error (MSE). SSPT first uses one pretraining epoch of Stock Code Classification (SCC) and Moving Average Prediction (MAP). Stock Sector Classification (SSC) is omitted because point-in-time sector labels are unavailable.

The four reconstructed capitalization-rank pools are Large 1–300, Mid 301–800, Small 801–1800, and Micro 1801 onward. Predictions and cross-sectional inference context stay fixed under candidate replacement. Of 34 scheduled test dates, 33 have full forward-label horizons. Their common finite-score, finite-outcome reference catalogs contain 4,460–4,727 stocks. Truth for Recall is the top 60 forward returns over the reference catalog, not a pool-specific relabeling. Evaluation uses the fixed cohort and observed outcomes; the cohort excludes the 448 later-listed stocks.

Input channels and preprocessing. Write $C_t, O_t, H_t, L_t, V_t, A_t, T_t, W_t$ for close, open, high, low, volume, traded amount, turnover and VWAP. The eight channels are $\log(C_t/C_{t-1})$, $\log(O_t/C_{t-1})$, $\log(H_t/C_{t-1})$, $\log(L_t/C_{t-1})$, $\log[(V_t+1)/(\bar{V}_t+1)]$, $\log[(A_t+1)/(\bar{A}_t+1)]$, $\log(1+T_t)$ and $\log(C_t/W_t)$. The bars denote trailing 20-session means including t. Each channel is centered and scaled using input windows from the purged training dates, then clipped to $[-10, 10]$. Missing observations are not forward-filled. Cross-stock models retain the fixed 5,072-identity axis, with unavailable inputs zero-padded and an observation mask; the candidate pool never changes that inference context.

Neural training uses Adam with learning rate 10^{-3}, one full-date context per update, and gradient norm clipping at one. MLP, GRU and MASTER minimize MSE. StockMixer and SSPT add 0.1 times the mean all-pairs penalty $\max\{0, -(s_i - s_j)(y_i - y_j)\}$, where s_i, y_i are the predicted and supervised target scores for stock i on that date. Only observed targets enter the loss. Selected epochs are 1/6/3/6/6 for MLP/GRU/StockMixer/MASTER/SSPT. SSPT pretraining masks 30% of time positions per channel for historical-price MAP and freezes its input linear projection during

Table 6: Configurations of the seven frozen common-input stock scorers. All receive the same 20×8 history. Neural checkpoint selection uses validation MSE.

Model	Configuration
Ridge	Flattened 160-dimensional input; ℓ_2 coefficient 100.
LightGBM	31 leaves, learning rate .03, minimum 100 samples per leaf, at most 250 rounds, 30-round validation early stopping; selected round 57.
MLP	Fully connected widths $160 \to 64 \to 32 \to 1$, ReLU hidden activations.
GRU	One recurrent layer of width 32 and a linear head on the last hidden state.
StockMixer	Time resolutions 20 and 10, stock-mixing hidden width 20, fixed identity axis and a scalar score head.
MASTER	Width 32, four attention heads, zero dropout, gate temperature one; context summaries replace external index features.
SSPT	Two width-32, four-head attention layers, zero dropout, learned position embeddings; SCC and historical-price MAP pretraining.

supervised fitting. These common-input configurations define the frozen model set used to study evaluation.

G.2 MovieLens and trained recommendation models

We use the stable MovieLens 25M release (Harper & Konstan, 2015; GroupLens, 2019). The four evaluation months are June 2010, 2012, 2014 and 2016. History-only activity ranks select 600 users; four rank strata contribute the first eight evaluable users each. This yields 32 users per month, 128 user-month cases and 98 distinct users. Reference catalogs contain 10,013, 10,881, 11,107 and 13,523 history-eligible, history-scorable items. Historical seen items are excluded per query. Recall@20 truth is evaluation-month interactions intersected with the point-in-time popularity pool and excluding history-seen items. This fixed truth contract is retained throughout membership changes. Because positive labels were filtered through the point-in-time pool, their union logically implies candidate membership. These positive facts are mandatory initial evidence for every acquisition policy and certificate. This information closure is absent for stock labels, which are defined on the reference catalog independently of pool membership.

The base scorers are item-kNN, popularity, random and a fixed hybrid. BPR-MF (Rendle et al., 2009) and LightGCN (He et al., 2020) are additionally trained on the same 600 users' complete distinct pre-month interactions: 1,074,151; 1,147,242; 1,195,948; and 1,308,895 positive user-item edges. Both use 32-dimensional embeddings, Adam with learning rate 0.01, 60 fixed gradient steps and 32,768 sampled triplets per step, with regularization 10^{-4}. Negatives are unobserved historical items for the sampled user. LightGCN averages the initial embedding and two normalized bipartite-graph propagation layers. The final fixed-step checkpoint is used; evaluation-month feedback is never used for training or selection. Both scorers use the same reference catalog and evaluation queries.

G.3 Budgets and controls

Fixed-order queries unknown memberships in a deterministic identity-hash order; the first quarter supplies initial evidence. Frontier ranks unknown optimistic-frontier members by frequency, earliest rank, then identity. Every method receives the same correct count, optional exact snapshot overlap and fixed evaluation gains. Additional budgets are 0, 5, 10, 20, 40, 60 and 75% of catalog size. Trajectories use batches of at most 64 and continue until exact model-choice certification, with first certificates also reported for $\epsilon = 0.01$ and 0.025. A budget uses only completed query batches; an already issued epsilon certificate is retained. Count-frontier uses $p(1-p)$ for the remaining-positive fraction in each group, with the Frontier order breaking ties. All methods apply free count closure.

The main recommendation results use the six-scorer model set. Costs count individual membership facts. The target ablation uses identical query prefixes and thus separates a changed stopping objective from a changed acquisition policy.

The controlled mechanism study uses four fixed instances for each combination of 0/8/32/128 permuted model-order positions, initial evidence chosen by hash order or from model A's lowest-scored candidates, and anchor weights 0.01/0.2. Catalog size is 128 uncertain identities plus two known anchors, with half the uncertain identities positive. It uses eight-record batches. The anchor changes a known component of model performance; it does not hold the entire realized model gap fixed.

G.4 Artifact checks

Protocols, sources, frozen scores and output hashes are recorded separately. Validation independently sorts the original scores, verifies every observed answer and issued regret certificate, checks interval refinement, and recomputes comparison-guided query plans. Unit tests enumerate coordinate disagreements and feasible memberships; controlled examples verify the target separation. Cached traces preserve all cases and are not fresh runtime measurements.

H Additional Evidence-Cost Separations

Evidence cost decomposes into intrinsic, relaxation and search components. For a realized state z, let $C_*(z, \epsilon)$ be the minimum number of paid records certifying some model as ϵ-best on every compatible state. Let C_{J} be the same minimum using Joint, and C_π the realized cost of a policy π stopped by that bound. Suppressing fixed z, ϵ, we have

$$C_* \le C_{\mathsf{J}} \le C_\pi, \qquad C_\pi - C_* = (C_{\mathsf{J}} - C_*) + (C_\pi - C_{\mathsf{J}}). \tag{21}$$

Both set minima are offline: their attaining records may depend on z. For a proposed answer a, form all bad alternative states where its regret exceeds ϵ. A certificate hits the coordinate-difference set of every bad alternative. Minimizing this hitting set over a gives an exact finite-family reference, checked by independent enumeration and integer optimization. This instantiates established certificate concepts in the shared-ranking family.

For an online reference, let $D_*(V)$ be optimal worst-case membership-query depth on evidence family V. A terminal node has a model acceptable on every state in V, and $D_*(V) = 0$ there. Write $V_{i,v} = \{z \in V : z_i = v\}$ and $I(V) = \{i : V_{i,0} \neq \varnothing,\ V_{i,1} \neq \varnothing\}$ for informative, non-entailed records. Otherwise its Bellman recurrence is

$$D_*(V) = 1 + \min_{i \in I(V)} \max_{v \in \{0,1\}} D_*(V_{i,v}). \tag{22}$$

The analogous enclosure-stopped optimum D_{J} satisfies $D_* \le D_{\mathsf{J}} \le \max_z C_\pi(z)$. These online quantities use a worst-case objective.

Proposition 8 (Search can remain expensive)**.** *For every even $N \ge 4$ and $0 \le \epsilon < 1/2$, a two-model top-1 family with N unknown memberships, one known anchor and one exact count has $C_* = C_{\mathsf{J}} = 1$ at every state but $D_* = D_{\mathsf{J}} = N/2$, at a fixed model gap of $1/2$.*

One of the N unknown candidates is a member. Its public gain is zero or one according to a balanced partition; the mandatory anchor has gain $1/2$. Thus the full catalog has $n = N + 1$ candidates. Model A selects the unknown member and B the anchor. An offline certificate queries that member, whereas an all-negative online path must eliminate one whole class. This separates online search from offline verification cost.

Proposition 9 (A fixed-margin enclosure penalty)**.** *For $N = 2K$ unknown memberships, $0 < \delta < 1$ and $\epsilon \ge 0$, there is a family with three evaluation queries whose true model gap is $\delta/(2+\delta)$ and $C_* = 0$, but the query-wise Joint bound has exact offline cost*

$$C_{\mathsf{J}} = \max\{0, \lceil K[1 - \delta - (2+\delta)\epsilon] \rceil\}. \tag{23}$$

Two evaluation queries have swapped model orders and identical gains, so their uncertain contrasts cancel. An anchor-only query supplies the fixed margin. Bounding the queries separately loses this cancellation. At $\delta = 0.1$, $N = 128$ and $\epsilon = 0.01$, Joint needs 57 records although the comparison needs none. This isolates a cost of the comparison bound itself, distinct from the cost of recovering selections.

Preserving cancellation across queries. For a model pair, queries with identical eligible orders and cutoff share the same selection-difference function, possibly with reversed sign $s_u \in \{-1, 1\}$. Within each such group, combine gains as $\widetilde{g}_i = \sum_u w_u s_u g_{ui}$ before applying the zero-sum bound. Intersecting with the original bound preserves soundness and monotonicity. In the construction, all combined uncertain gains vanish, permitting certification without records. Appendix I proves this exact combination rule.

I Evidence-Cost Proofs and Conditional Computation

I.1 Offline certificates and online decision regions

Write $W_\epsilon(z) = \{a : \max_b Q_b(z) - Q_a(z) \le \epsilon\}$. An evidence family V is terminal iff $\bigcap_{z \in V} W_\epsilon(z)$ is nonempty. Pairwise intersection is insufficient: acceptable sets $\{a, b\}, \{b, c\}, \{a, c\}$ overlap pairwise but have no common answer. The exact online dynamic program uses the full intersection condition.

Fix a realized z and a proposed model a. Let $\mathcal{F}_a$ contain initial feasible states z' where selecting a has regret greater than ϵ. A record set S certifies a exactly when it intersects every difference set $\{i : z_i \neq z'_i\}$ for $z' \in \mathcal{F}_a$. An unexcluded alternative agrees with all queried answers and remains a counterexample; intersecting every difference set excludes all counterexamples. Initially known records are not paid again, and the feasible family already includes count implications. Minimizing the record count over a gives C_*. A sound bound certifies an acceptable model, and a stopped policy provides such a certificate, proving Eq. (21).

An online decision tree either stops with a common acceptable model or selects an informative first test. A test whose answer is already entailed can be removed without altering the decision, so the minimum may be restricted to $I(V)$; every resulting branch strictly shrinks V. Its worst-case remaining depth is the maximum over its nonempty response branches. Minimizing that first test proves the Bellman recurrence; joining optimal child trees proves attainability. Replacing the terminal condition by enclosure certification gives D_J. These definitions require no prior. The recurrence optimizes worst-case query depth; Bayesian expected-cost criteria use a prior-weighted objective.

I.2 Fixed-gap search separation

Partition N unknown candidates into equal public gain-zero and gain-one classes, with exactly one member. Add a mandatory gain-$1/2$ anchor, giving catalog size $n = N + 1$. All candidates are eligible, $K = 1$, model A ranks unknowns before the anchor, and B ranks the anchor first. The strict winner depends on the unknown member's class and always has gap $1/2$. Querying that member costs one and fixes the state by count closure; without records both winners remain possible, so $C_* = C_\mathsf{J} = 1$.

After fewer than $N/2$ tests on an all-negative response path, an unqueried candidate remains in each class. Either can be the sole unknown member and yields a different winner. Thus any deterministic policy has worst-case cost at least $N/2$. Querying one entire class attains this bound: stop at a positive answer, or infer the opposite class after all negative answers. The lower bound concerns deterministic instancewise guarantees relative to the offline certificate; the optimal online tree is a different comparator.

I.3 Fixed-gap query-wise enclosure penalty

Let $N = 2K$ unknown candidates contain exactly K members, and add K mandatory zero-gain anchors and two auxiliary anchors, giving $n = 3K + 2$. The first two evaluation queries are eligible on the unknowns and zero-gain anchors. In the first, A ranks unknowns first and B anchors first; the second swaps these orders. The same public half H of unknowns has gain $1/K$ in both queries; other gains are zero. A third, auxiliary-anchor-only query contributes one to A and zero to B. The three normalized weights are $(1, 1, \delta)/(2 + \delta)$. The first two contrasts cancel at every state, giving gap $\delta/(2 + \delta)$ and $C_* = 0$.

Let $x = |H \cap C(z)|$ and let W be the width of its feasible interval divided by K. The individual query bounds are exact in this construction, but separate aggregation gives contrast interval $[(\delta - W)/(2 + \delta), (\delta + W)/(2 + \delta)]$. The smaller regret bound is $\max\{0, (W - \delta)/(2 + \delta)\}$.

After t explicit records, let h_1, h_0 count unqueried candidates in H and its complement within the N unknown candidates, and let r be K minus the queried positive count. The width of the feasible interval for x is

$$\min(h_1, r) - \max(0, r - h_0) = \min(h_1, h_0, r, h_1 + h_0 - r) \tag{24}$$
$$\geq \max(0, K - t). \tag{25}$$

Each of the four terms starts at K and decreases by at most one per paid record. Querying $t \leq K$ identities exclusively in H attains range $K - t$ for every response pattern. Solving $W \leq \delta + (2 + \delta)\epsilon$ proves Proposition 9, including its zero-cost regime.

I.4 Combining queries with identical selection rules

For a model pair, two evaluation queries with the same full eligible orders and cutoff have the same selection-difference vector, or its negative when orders swap. Within a group of such queries, choose a canonical orientation $d_i(z)$ and signs $s_u \in \{-1, 1\}$. The contrast identity is

$$\sum_u w_u \sum_i g_{ui} s_u d_i(z) = \sum_i \left(\sum_u w_u s_u g_{ui} \right) d_i(z).$$

Signed coefficients are allowed by the linear box-plus-sum solver. Apply the sound enclosure to each group and intersect the aggregate with the uncombined query-wise bound. Fixed geometry groups and contracting coordinate ranges give monotonicity; intersection never widens the uncombined result. The uncertain group in the construction has zero combined coefficients and thus costs zero records. Similar orders or equality on the realized state alone do not justify this combination.

I.5 One-step planning and exact conditional count updates

For each candidate membership query i, simulate feasible answers $v \in \{0, 1\}$. Let $\mathcal{E}_{i,v}$ be the evidence after adding $z_i = v$, and calculate $\rho_{iv} = \min_a \overline{R}(a \mid \mathcal{E}_{i,v})$. Minimizing $\max_v \rho_{iv}$ finds a query that guarantees certification after one answer whenever such a query is available in the considered set. Its worst-child bound is then at most ϵ. The realized answer is read only after the plan is fixed.

For efficient simulation, consider an unknown j in a count group with current mandatory/available totals m, a, target c, remaining count $r = c - m$ and unknown count $u = a - m$. Close entailed counts first, so $0 < r < u$. In any predecessor region P, let $p = \mathbf{1}\{j \in P\}$ and current counts be M_P, A_P. Conditional counts are given in Table 7. The same rules apply to total counts and

Table 7: Exact conditional predecessor counts after one membership answer, including all free count implications.

Answer	Condition	New mandatory	New available
Present	$r > 1$	$M_P + p$	A_P
Present	$r = 1$	$M_P + p$	$M_P + p$
Absent	$r < u - 1$	M_P	$A_P - p$
Absent	$r = u - 1$	$A_P - p$	$A_P - p$

singleton status; other groups are unchanged. They include all free count implications, even when one answer determines many other records. Substitute these integers in the existing directional reachability and allocation formulas. Current-state rank sweeps cost $O(\sum_u M^2 G n \log n)$, followed by $O(\sum_u M^2 G n)$ arithmetic per simulated answer with sorted gains reused. The conditional implementation is compared with the original bounds at all distinct decision states in the full-catalog refinement experiment, and for all alternatives at each instance's first one-step lookahead decision. Exhaustive small tests additionally cover count-closure branches, signed gains and query-ineligible records.

J EXPERIMENTAL SETTINGS

The synthetic exact-cost grid varies the unknown count (8, 10 or 12), evaluation-query count (one or two), model orders within each evaluation query (identical, single-swap perturbations, or independently generated), count groups (one or two) and two fixed seeds. Two realized states and two tolerances give 288 settings. Online exact references use the 48 size-eight settings. The Hong Kong exact-cost slices use fixed-hash 8/12-stock subsets of all 17 dates, seven unchanged model scores, $K = 2$ and three tolerances, giving 102 settings. Subset enumeration and proven-optimal hitting-set integer programs agree. The large fixed-margin constructions are verified against the analytic lower bound and an attaining query sequence.

The Hong Kong event experiment uses a single September 2024 announcement cohort. Models receive five causal price/volume channels over twenty actual observations. Training is January 2022–June 2023, validation July–December 2023, and test dates are fixed in September 2024. Fifty-three identities have observed training history and thirteen are cold-start in that window; all 66 are retained. Labels are next-common-market-session open-to-close return percentiles, averaged over the selected five. They are an evaluation utility, not executable Stock Connect P&L. Current ISINs are not backfilled into historical identities. Query costs count individual membership facts.

The full-catalog refinement experiment uses the first, middle and last of 33 stock dates, all four pools, and known pool size with or without snapshot overlap, plus all four full recommendation catalogs. Stock reference catalogs contain 4,474–4,706 identities; recommendation catalogs contain 10,013–13,523. Initial evidence and label-implied memberships are unchanged. Refinement has a common bound-based entry threshold, a 64-record cap, and identical fallback. Schedule refinement and query-rule comparisons are reported separately, since schedule changes can affect both later plans and stopping. Numerical rounding is used only for ranking ties at 10^{-8}; reported and stopping bounds remain unrounded. Numerical agreement tolerance is 10^{-10}.

K SHARED-CUTOFF PROOFS

K.1 CUTOFF CHARACTERIZATION AND EXACT OPTIMIZATION

If $d_i = 1$, identity i is selected by a but not b. If $d_j = -1$, j is selected by b but not a. Both are members. Hence i precedes j in a's total order and j precedes i in b's order. The actual K-th selected position of a separates all positive and negative signs. There are at most K of each, since the underlying selected sets have size K. Eligibility is common across models within a query; ineligible coordinates are zero and play no role in this argument.

Fix one proposed cutoff. A forced positive after it, or a forced negative before it, makes this cutoff infeasible. Otherwise all optional nonzero signs are positive in the prefix or negative in the suffix. If a vector has m positive signs, zero sum requires m negatives. Exactly $m - p$ optional positives and $m - q$ optional negatives must be chosen, yielding the feasible range in Eq. (16). Their largest possible objective takes the largest prefix gains and smallest suffix gains. These choices attain the formula because the two optional sets are disjoint and no further constraints define this relaxation.

For each suffix, maintain its K smallest optional negative gains and their prefix sums. Sweep the cutoff while maintaining the K largest optional positive gains in its prefix. Inserting a gain into a sorted length-K list takes $O(K)$ work; evaluating each cutoff takes at most K possible cardinalities. Zero coordinates with $\ell_i = u_i = 0$ change neither signs nor objective, so deleting them only merges duplicate cutoff positions. Keeping the s nonzero-range coordinates, including fixed nonzero signs, gives $O(n + sK)$ work and $O((s + 1)K)$ stored suffix sums. The implementation's support mask adds $O(n)$ working space. The $s = 0$ case has zero optimum and is covered by the same time bound since $K \le n$. A negative-gain pass gives the minimum. No nonnegativity assumption on gains is needed for the optimization.

Each true difference lies in the first cutoff union, and its negative lies in the second. Thus intersecting their objective intervals preserves all true contrasts. Nonnegative query weights preserve enclosure under summation; intersection with Joint cannot widen it. Under additional evidence, the current coordinate ranges contract. Both cutoff unions therefore contract, proving monotonicity. The

construction optimizes each one-order cutoff union independently, yielding an outer relaxation of the common-membership family.

K.2 Fixed-gap and perturbation guarantees

First consider unnormalized gains $h_i = i/(Kn) + \zeta_i$, where $|\zeta_i| \le \eta/K$, and $n = 4K$. For any active set with ordered identities $j_1 < \cdots < j_{2K}$, model A selects the first K and B the last K. Since $j_{r+K} - j_r \ge K$,

$$Q_B - Q_A \ge \sum_{r=1}^{K} \frac{K}{Kn} - 2\eta = \frac{1}{4} - 2\eta.$$

In every feasible A-cutoff vector, pair its $m \le K$ positive coordinates with its negative coordinates. Every paired positive index i is smaller than its negative index j, so $h_i - h_j \le 2\eta/K$. Its objective is at most 2η. The shared-cutoff bound therefore certifies B at tolerance at least 2η without any membership records.

For bounded gains use $g_i = (h_i + \eta/K)/(1 + 2\eta)$, which is exactly the formula in Proposition 7 with $\xi_i = K\zeta_i$. Every g_i lies in $[0, 1/K]$. The constant shift cancels between equal-size selections and in every zero-sum relaxation, and division scales all contrast endpoints by $1 + 2\eta$. This proves the stated true gap and Cutoff threshold without clipping the perturbations.

K.3 Linear evidence lower bound for the coordinate enclosure

Consider any $t \le \lfloor K/64 \rfloor$ truthful queried identities under the exact $2K$ count. No free count closure has yet fixed an additional membership. Any unqueried identity in $[K + t + 1, 3K - t]$ can realize either sign: for $+1$, choose $K - 1$ earlier members and K later members; for -1, choose K earlier and $K - 1$ later members. The interior position leaves sufficient capacity after at most t fixed bits, and fewer than K compulsory members occur on either side. Therefore its coordinate range contains both -1 and $+1$.

The low band $[K + t + 1, 3K/2]$ and high band $[5K/2 + 1, 3K - t]$ each contain at least $K/2 - 2t$ unqueried identities. Choose $m = K/2 - 3t$ positive high-band and negative low-band coordinates. Every such pair contributes at least $(1/4 - 2\eta)/K$ in the unnormalized gains. There are at most t forced nonzero coordinates; use at most t further available band coordinates to balance their signed sum. Forced and balancing contributions lose at most $2t(1 + \eta)/K$. Set all other coordinates to zero. This is feasible for the coordinate box and zero sum, even if an additional K-positive cap is imposed. Its objective is at least

$$\left(\frac{1}{2} - \frac{3t}{K}\right)\left(\frac{1}{4} - 2\eta\right) - \frac{2t(1 + \eta)}{K}. \tag{26}$$

A corresponding lower bound holds for the marginal-cardinality enclosure. With all upper $2K$ identities active, A's utility exceeds B's utility with all lower $2K$ active by at least $1/4 - 2\eta$. Each of these two different states can satisfy the t records by at most t membership exchanges. A top-K sum changes by at most $(1 + 2\eta)/K$ per exchange. Thus the marginal contrast upper bound is at least

$$\frac{1}{4} - 2\eta - \frac{2t(1 + 2\eta)}{K}. \tag{27}$$

The coordinate enclosure's upper bound is at least the smaller of Eqs. (26) and (27). Normalize both expressions by $1 + 2\eta$. For $(\eta, \epsilon) = (0, 0)$ or $(.01, .025)$, both remain strictly above ϵ throughout the stated range of t. The reverse contrast upper bound is at least the actual fixed positive gap. Neither model can therefore pass the Joint regret certificate. The Joint cost is greater than $\lfloor K/64 \rfloor$, whereas the cutoff cost is zero.

L Protocol, Training and Evidence Conditions

L.1 Matched-prefix certificate comparison

The full audit uses all 396 original stock and four fixed-budget recommendation Contrast trajectories. Predictions, gains, initial facts, records and batch boundaries are unchanged. All 6,890 recorded

states are checked against their realized evaluation-value differences and monotone refinement; the Joint bound matrices match the frozen implementation. A separate sort-and-scan reference checks one model pair/evaluation query at the initial state and the Cutoff stopping state ($\epsilon = .01$) in every instance: 800 comparisons, with maximum absolute discrepancy below 4×10^{-16}. Small-instance enumeration additionally checks uniform validity over the complete declared state family.

The single-record comparison walks the same saved requested order, dropping records already entailed by prior answers for both certificates. Monotone binary search locates each certificate's first passing prefix; it is an offline evaluation procedure on the fixed requested order. The preceding prefix is explicitly checked to fail. Marginal bounds that remain unresolved at the end of the saved Joint-certified sequence are right-censored. At .01 the fixed-budget recommendation has identical stopping costs under Joint and Cutoff; the stock Recall fine-grained audit improves one of 132 cases by four records. The mean-percentile improvements are reported in the main text.

L.2 Connecting discovery and repair

Table 8 retains all original discovery cases and reports the above-tolerance subgroup separately. "After" counts regret above .01 at each case's own certificate. Cases not certified at a fixed budget are unresolved. Remaining unknown identities refer to memberships after free count closure; outcome labels stay fixed.

Table 8: Same-case repair at $\epsilon = .01$. Before/after are counts of above-tolerance errors; records and remaining unknown memberships are means.

Subset	Cases	Before	After	Records	Unknown
All	132	23	0	957.8	2486.6
Above tolerance	23	23	0	889.0	2574.0

L.3 Validation-selected training budgets

The 600-user cohort, evaluation queries, full item catalogs, gains and membership evidence remain fixed. For each test month, the immediately preceding calendar month is validation; embedding training uses strictly earlier interactions. Validation Recall@20 excludes training-seen items and averages users with at least one supported unseen positive, among the fixed 600 users. Budget ties choose the earlier checkpoint. The final validation grid contains 60, 120, 240, 480, 960 and 1920 steps. The selected budget is then refitted from initialization using all pre-test history; no test interactions select steps. Table 9 reports all eight model-month results; four selected budgets equal the 1920-step upper limit.

Table 9: History-validation budget selection. Validation and test columns are Recall@20 on their separately declared populations. Test columns compare 60-step and validation-selected training budgets on the same test month.

Month	Model	Selected steps	Validation	Test (60 steps)	Test (selected budget)
2010-06	BPR-MF	960	0.0151	0.0078	0.0232
2010-06	LightGCN	1920	0.0180	0.0073	0.0313
2012-06	BPR-MF	480	0.0242	0.0000	0.0803
2012-06	LightGCN	1920	0.0224	0.0000	0.0566
2014-06	BPR-MF	1920	0.0257	0.0078	0.0516
2014-06	LightGCN	1920	0.0268	0.0000	0.0294
2016-06	BPR-MF	960	0.0195	0.0156	0.0030
2016-06	LightGCN	480	0.0146	0.0195	0.0308

Architecture, optimizer, batch size and embedding dimension follow the original fixed-budget implementation in Appendix G.2. This check addresses training-budget sensitivity with the common-input implementation. The retrained six-model study reruns all four policies on each full catalog; these comparisons are reported separately from the stock discovery and fixed-budget recommendation results.

L.4 Certification with count intervals

For an integer count interval $[c_-, c_+]$, the evidence family is the union of its feasible exact-count components. Taking the minimum lower bound and maximum upper bound over these components is sound for the union. Infeasible components are removed using the *explicit* records; each component derives its own free count entailments. Entailments from the original exact count or overlap must not be imported into a weaker family. We apply the standard union construction.

We use the first, middle and last of 33 stock dates, all four pools, and the initial and Joint stopping records ($\epsilon = .01$). The nested levels are exact count plus overlap, exact count only, and count intervals of radius 1 or 5 without overlap. All 12 Joint stopping points remain certifiable in this local check. The intervals are controlled perturbations around known truth, with the declared radii defining the metadata uncertainty evaluated here.

L.5 Pool size and controlled disagreement

All 33 dates and four pools are also evaluated with $K = 60$ versus a prespecified $K = \lfloor .1c + .5 \rfloor$ rule, where c is pool size. Frozen model scores and outcome percentiles defined on the reference catalog are unchanged, with mean gain renormalized by the declared K. The resulting 264 rows retain both rules and all cases. This tests selected-fraction sensitivity.

Table 10: Counts of realized snapshot-choice regret exceeding each threshold over 33 dates per pool. Group headers specify selection size: $K = 60$ or approximately 10% of pool size. Model scores stay fixed.

	$K = 60$		$K \approx 10\%$ of pool	
Pool	Regret > 0	Regret $> .01$	Regret > 0	Regret $> .01$
Large	8	6	8	5
Mid	10	4	11	7
Small	14	8	8	5
Micro	10	5	4	2

The synthetic fixed-gap experiment has 128 uncertain identities, 64 active, $K = 8$, 25% of uncertain identities initially verified, four fixed instances, 0/8/32/128 permuted model-order positions and two initial-evidence patterns: hash order or model A's lowest scores. A known anchor contribution is calibrated to keep the *actual* model gap at 0.1 in every instance. All methods receive the same resulting problem and use the same hash order. The robust-separation construction checks $K = 4, 16, 64, 128$, perturbations 0/.01, and random, increasing-i and decreasing-i record prefixes (construction index i) against the analytic lower bounds.

M Order-Gain Control and Certificate Guarantees

M.1 Proof and implementation of the linear-time control

Let $P = \{i : u_i = 1\}$ and $N_- = \{j : \ell_j = -1\}$. Write $i \prec_a j$ when i precedes j in model a's order. Every positive–negative pair obeys $i \prec_a j$ and $j \prec_b i$. Consequently,

$$Q_a - Q_b \le K \min \left\{ \max_{\substack{i \in P, j \in N_- \\ i \prec_a j}} (g_i - g_j)_+, \max_{\substack{i \in P, j \in N_- \\ j \prec_b i}} (g_i - g_j)_+ \right\}, \tag{28}$$

Table 11: Linear-time order-gain control at $\epsilon = .01$: mean paid records on the same Contrast prefixes. Counts in parentheses are evaluation instances. The complete six-tolerance check leaves all 400 stopping counts unchanged relative to Joint; the table uses a common absolute tolerance whose relative scale is task-dependent.

Task	Joint	Order-gain	Cutoff
A-share mean percentile (264)	987.03	987.03	982.81
A-share Recall (132)	773.98	773.98	773.98
MovieLens, fixed-budget (4)	1472.00	1472.00	1472.00

with an empty maximum defined as zero. Running extrema evaluate this control in $O(n)$ time once the local ranges are known.

For a feasible contrast vector, pair each positive identity i with one negative identity j. Zero sum makes such a pairing possible, with at most K pairs. Every pair has $i \prec_a j$ and $j \prec_b i$; therefore its gain is bounded by each maximum in Eq. (28). Summing the bounds proves that equation. Both maxima use the positive part, so the inequality remains valid for negative-gain exchanges and for an empty exchange set. A forward scan keeps the largest prior positive gain; at each possible negative coordinate it tests the resulting difference. The reversed orientation uses negated gains and exchanged sign roles. Strict precedence is enforced by reading the running extremum before inserting the current coordinate. Total additional work is linear in the eligible order length.

Each vector in the corresponding one-order cutoff family admits the same pairing. Its exact optimum is therefore at most the one-order gain bound. Taking the two-order minimum, nonnegative weighted query sum and intersection with Joint preserves this dominance. The control isolates the contribution of cutoff compatibility. In particular, it certifies the special family in Proposition 7; that proposition establishes the evidence-cost separation between Cutoff and Joint.

The matched-prefix comparison uses all 400 Contrast trajectories at $\epsilon = 0, .001, .0025, .005, .01, .025$. It checks every realized contrast, monotonicity, and the ordering of the three bounds. The order-gain intersection has the same stopping count as Joint in every case on this grid. Independent small-instance tests verify the bound and the illustration in Figure 5. Model scores remain fixed across this control.

M.2 Interpretation of the mathematical guarantees

Local coordinate extrema are exact only for the stated count family under automatic query admissibility. The cutoff optimization is exact for its union relaxation, which encloses the common-membership family. The NP-hardness reduction applies to multiple evaluation queries sharing a membership vector. The fixed-gap lower bound is relative to a named enclosure, while the offline and online comparators in the supporting cost analysis have different information. Numerical tests use the explicitly declared 10^{-8} stopping tolerance.

N Context-Specific Choice and Evidence-Cost References

N.1 Static-choice controls

Transferability benchmarking uses a static ranking to test whether a model-selection task rewards a pre-existing model hierarchy (Singh et al., 2026). We adapt that diagnostic to the fixed stock evaluator, keeping all 33 dates, four pools, seven models and the primary $\epsilon = .01$ tolerance. For each model, Table 12 counts cases where its realized point-in-time regret exceeds $.01 + 10^{-8}$. Training and membership-query trajectories are held fixed.

The hindsight static choice minimizes the error count over seven models, breaking ties by mean regret and then canonical model order. SSPT achieves the minimum of 93 errors and also minimizes mean regret (0.05650). A pool-specific static rule may choose a separate model for each pool but cannot change it across dates. Exhaustive evaluation of all $7^4 = 2,401$ assignments confirms that its

Table 12: Errors of fixed model choices on all 132 stock comparisons. Each pool contains 33 dates. Mean regret averages all cases. StockMixer, MASTER and SSPT retain the common-input implementations specified in the model protocol.

Fixed model	Large	Mid	Small	Micro	All	Mean regret
Ridge	29	25	25	24	103	0.06716
LightGBM	29	29	27	27	112	0.07262
MLP	30	31	31	28	120	0.07365
GRU	23	25	26	28	102	0.06736
StockMixer	31	25	25	31	112	0.06993
MASTER	28	26	25	24	103	0.06429
SSPT	23	23	24	23	93	0.05650

minimum is also 93 errors. The respective per-pool minima are 23, 23, 24 and 23; the mean-regret tie break selects SSPT in each pool. This rules out these two static policy classes as an explanation of the all-case repair.

Snapshot selection incurs 23 errors with mean regret 0.00609; the certified decisions incur zero errors with mean regret 0.00041. The certificate procedure acquires membership evidence, whereas the hindsight controls use the complete realized evaluation to choose their fixed preferences. They isolate whether fixed preferences explain the repair. Equal-information efficiency is measured by the matched-certificate and matched-policy experiments in the main text.

N.2 Intrinsic cost, relaxation cost and search cost

Fix an initial evidence family, a realized state z and a tolerance ϵ. Let C_* be the minimum number of additional records certifying an acceptable model using exact common-state regret; define C_{cut} and C_{J} analogously using Cutoff and Joint. Each minimum is offline: the minimizing records may depend on z. Soundness and pointwise dominance imply

$$C_* \le C_{\mathrm{cut}} \le C_{\mathsf{J}}, \qquad C_{\pi,\mathrm{cut}} - C_* = (C_{\mathrm{cut}} - C_*) + (C_{\pi,\mathrm{cut}} - C_{\mathrm{cut}}), \tag{29}$$

Here $C_{\pi,\mathrm{cut}}$ is the cost of policy π stopped by Cutoff; $C_{\pi,\mathsf{J}}$ is defined analogously. Every Cutoff certificate is valid under exact regret, and every Joint certificate is valid under Cutoff. Minimizing over these nested classes of record sets gives the first inequalities. A policy's stopping records belong to the Cutoff class, so both terms in the decomposition are nonnegative. On one fixed record order, pointwise dominance also gives $C_{\pi,\mathrm{cut}} \le C_{\pi,\mathsf{J}}$. The offline minima may use different record sets.

We use the full factorial grid: $n \in \{8, 10, 12\}$, one or two evaluation queries, shared/local/independent model orders, one or two count groups and the two fixed generator instances. Two realized states and $\epsilon \in \{0, .05\}$ give 288 settings in 72 families. Membership enumeration independently evaluates all models; subset search computes C_{cut} and rechecks C_* against the original independent hitting-set optimum. Every evaluated Cutoff bound contains the exact common-state regret and is no larger than Joint.

Table 13: Offline minima and matched-order stopping costs on all 288 small-family settings. The policy columns follow the shared Contrast order; the offline columns minimize over record sets.

Order regime	C_*	C_{cut}	C_{J}	$C_{\pi,\mathrm{cut}}$	$C_{\pi,\mathsf{J}}$
Shared (96)	0.000	0.000	0.000	0.000	0.000
Local (96)	0.812	0.865	1.021	2.010	2.302
Independent (96)	2.542	3.073	3.167	4.906	4.948
All (288)	1.118	1.312	1.396	2.306	2.417

Cutoff attains the exact offline minimum in 239/288 settings and reduces the Joint minimum in 23. On the fixed Contrast order it reduces stopping cost in 15 settings. The mean relaxation cost

$C_{\mathrm{cut}} - C_*$ is 0.194 records, and the mean search cost $C_{\pi,\mathrm{cut}} - C_{\mathrm{cut}}$ is 0.993. This finite-grid decomposition quantifies the contributions of relaxation and search. The shared-order settings certify at zero cost under every certificate. Among nonshared orders, Cutoff attains the exact minimum in 91/96 local-order settings and 52/96 independent-order settings; their respective numbers of offline improvements over Joint are 14 and nine. This stratification separates exact cancellation from local perturbations and independent rankings.

N.3 When interval contraction changes the decision

Cutoff narrows the mean pair interval in 4,808 saved real-data states; in 2,166 of them the minimum regret upper bound remains unchanged. Equation (18) identifies why: certification depends on the challengers preventing an acceptable choice, rather than average interval width. The fine-grained audit improves 67/264 stock mean-percentile stopping costs. Both fixed- and validation-budget recommendation model sets have identical Joint and Cutoff stopping costs; their main record saving comes from comparison-guided acquisition.

Exact evidence-cost references. The audit uses 72 small ranking families and 288 realized-state/tolerance settings. Mean costs are 1.118 for C_*, 1.396 for C_{J}, and 2.417 for Contrast. In 48 small online settings, mean worst-case depths are 1.958 for D_*, 2.146 for D_{J}, and 3.562 for Contrast. Thus the offline and online comparisons have distinct objectives. Another 102 small problems retain seven-model scores from subsets of all 17 Hong Kong dates, providing exact references on real predictions.

External Hong Kong event check. A public September 2024 adjustment (China Investment Information Services Limited, 2024) defines a retrospective cohort of 66 stocks: 33 additions and 33 removals. All remain in the experiment, including two absent from the current exchange list. Seven common-input models are trained on historical prices, and all 17 fixed dates have complete inputs and outcomes. Before the adjustment, the old membership does not misselect a model on any of five dates; afterward it misselects on 10/12 dates. This event-selected stress test evaluates the effect of a documented membership transition.

Bounded lookahead, with scheduling controlled. One-step lookahead simulates feasible answers to sixteen membership queries prioritized by Contrast and minimizes the worst-child regret bound. If one of these queries guarantees certification after either answer, lookahead finds such a query. On the Hong Kong cohort, mean cost falls from 31.53 to 28.85 records (8.49%), with 27 lower-cost and seven higher-cost configurations. Considering all unknown records gives mean cost 32.82. This tests the query subset prioritized by Contrast in this cohort.

For a full-catalog check, we fix three evenly spaced stock dates, all four pools and both information regimes (24 cases), plus all four recommendation catalogs. Every refinement policy shares 64-record Contrast batches until the observable regret bound reaches 0.02, then has at most 64 single-record queries and the same 64-record fallback, targeting $\epsilon = 0.01$. The query-rule comparison therefore holds the schedule fixed.

Table 14: Mean paid records in full-catalog refinement. Mixed schedules share a coarse prefix, at most 64 single-record queries, and the same fallback. Lookahead is one step over 16 candidate membership queries; Batch-only uses 64-record batches throughout.

Schedule / query policy	A-share 24 instances	MovieLens 4 catalogs
Batch-only	979.38	1472.00
Mixed / Frontier	982.17	1520.00
Mixed / Contrast	968.62	1479.25
Mixed / One-step lookahead	967.62	1479.25

The additional one-step-lookahead saving is 0.10% on the sampled stock cases and zero on the four recommendation catalogs. Seventeen stock cases finish during the common batch stage. The

comparison evaluates bounded lookahead within its computation budget, with Contrast as the default policy. Combining queries with identical selection rules certifies the corresponding cancellation construction without records. Stock instances have one evaluation query; the four recommendation catalogs contain no repeated eligible-order/cutoff pair.

O Observable Structure and Operational Costs

O.1 Natural shared-order structure

We apply Eq. (8) to all 21 unordered pairs of the seven original stock scorers, all 33 dates, four pools and both count regimes. The models' orders are unchanged. Every case starts with the original 25% identity-hash prefix; additional fractions 0/.10/.25/.50 reveal the same fixed hash-ordered records to all pairs, with free count closure. This gives 22,176 pair/evidence states. For each state we report the largest number $u_{\mathcal{E}}$ of unknown bits in a reachable partial-boundary block. The screening test is independently checked against full membership enumeration in 75 small families, including one/two count groups and query-ineligible identities.

Table 15: Natural-order tractability diagnostic. Each budget has 2,772 model-pair states per count regime. The median $u_{\mathcal{E}}$ is the same in the two regimes; the last columns count states satisfying an unknown-bit cap of 16. Caps of eight give the same counts.

Additional records/n	Median $u_{\mathcal{E}}$	Count only	Count and overlap
0	3,457	0	0
.10	2,997	0	0
.25	2,305	0	0
.50	1,153	0	21

The 21 cap-satisfying states at the final budget are the pairs of one pool-date case whose memberships become fully entailed by count closure. The diagnostic therefore assigns the original unrestricted scorer comparisons to the general-enclosure regime. The width-constrained reranking study evaluates a different, explicitly imposed local-order regime, where the same count test exposes small exact subproblems. These experiments identify the appropriate solver regime directly from the observable boundary structure.

O.2 Exact single-query runtime reference

The runtime grid fixes $n \in \{16, 32, 64\}$, imposed bucket widths 2/4/8 and three generator instances, for 27 problems. A random base permutation is independently reranked inside each bucket; gains are independent uniform values divided by $K = n/8$. All memberships are initially unknown, the total count is $n/2$, and there is one evaluation query. Block and the general exact mixed-integer formulation optimize both endpoints of the same contrast. The latter retains its exact aggregation of never-selected candidates into background counts and automatic encoding selection. The solver has a 30-second limit per endpoint; all 81 repeated interval solves reach optimal status, and every endpoint agrees with Block within 10^{-6}.

We use an Apple M5 CPU, Python 3.13.9, NumPy 2.3.5 and SciPy 1.16.3. JIT compilation and common problem construction precede timing; there are no concurrent experiment processes. Each case has three repetitions with alternating method order. Table 16 reports the median of the three per-case medians in each size/width cell. Times include computing both endpoints, not training or acquiring records. This comparison measures the runtime benefit of local enumeration against a general exact solver, complementing the structural work bound.

O.3 Record cost and certificate computation

We measure each bound refresh up to certification at $\epsilon = .01$ on all 264 original mean-percentile stock Contrast trajectories. The saved 64-record requests and answers are replayed unchanged. At each state, fresh Joint and Cutoff evaluators are timed three times, alternating order; input construction, JIT,

Table 16: Wall-clock milliseconds for a complete exact contrast interval on CPU. Each cell aggregates three generated problems and three repetitions per problem. Both methods return the same endpoints.

Catalog n	Bucket width	Block	General exact MILP
16	2	.031	21.618
16	4	.033	12.770
16	8	.061	9.410
32	2	.036	169.932
32	4	.038	128.671
32	8	.102	132.738
64	2	.040	766.068
64	4	.044	751.319
64	8	.190	781.140

policy planning and record access are excluded. Every Joint bound reproduces the saved bound, and both stopping-record counts reproduce the prior experiment. Summing each state's median refresh time up to the respective stopping point gives mean cumulative costs of .0790 seconds for Joint and .1348 seconds for Cutoff. Their mean record counts are 987.03 and 982.81, a difference of 4.22 records. The fixed query sequence isolates the computation required by each stopping module.

Let $\sigma \in \{\mathsf{J}, \mathrm{cut}\}$ index the certificate. Write N_σ for its paid record count and B_σ for cumulative bound-refresh time in seconds. If common initial evidence costs T_0 seconds and each additional record costs κ seconds, the stopping-module cost and break-even unit cost are

$$T_\sigma = T_0 + B_\sigma + \kappa N_\sigma, \qquad \kappa_{\mathrm{break}} = \frac{\max\{0, \bar{B}_{\mathrm{cut}} - \bar{B}_{\mathsf{J}}\}}{\bar{N}_{\mathsf{J}} - \bar{N}_{\mathrm{cut}}}. \tag{30}$$

Bars denote means over the same 264 cases; the denominator is positive. The measurements give $\kappa_{\mathrm{break}} = .0132$ seconds per record. Above this unit cost, the mean record saving offsets additional certificate computation in this model. A record may require verifying a dated inclusion/exclusion entry, whose retrieval cost differs from a bulk local lookup. Initial records, counts and overlap are shared between certificates and their acquisition belongs in T_0. The threshold translates measured computation into an acquisition-cost condition, with application-specific record-access latency supplied as a parameter. The count-interval experiment separately tests uncertainty in the supplied metadata.

O.4 INITIAL-EVIDENCE SENSITIVITY OF ACQUISITION

We retain the four full MovieLens catalogs, all original evaluation users, the six frozen validation-budget scorers, and the Joint stopping certificate. Only the initial 25% identity-hash sample changes. Three prespecified identity-hash salts define the initial evidence samples independently of membership answers. Positive-label-implied memberships remain initial facts for both policies. Frontier and Contrast are then rerun to $\epsilon = 0$ using the same 64-record batch rule. All 24 trajectories check realized-contrast containment and monotonicity at every state. Replaying the original sample reproduces its complete requested-record sequences and stopping costs exactly.

Table 17: Additional membership records at zero regret under three initial evidence samples. Entries are Frontier/Contrast. Each column uses the same four catalogs and frozen users/models; the columns vary initial records, not the evaluation population.

Catalog	Original sample	Alternative 1	Alternative 2
June 2010	2,944 / 1,152	3,072 / 1,280	2,944 / 1,344
June 2012	2,944 / 1,472	2,944 / 1,600	2,816 / 1,408
June 2014	3,008 / 832	3,008 / 640	2,944 / 576
June 2016	3,520 / 2,368	3,456 / 2,560	3,584 / 2,496
Mean	3,104 / 1,456	3,120 / 1,520	3,072 / 1,456
Saving in mean cost	53.1%	51.3%	52.6%

Contrast uses fewer records in all 12 paired catalog/sample comparisons; within-pair savings range from 25.9% to 80.4%. All three sample-level savings exceed 51%; these are ratios of mean record counts, rather than means of within-pair percentages. The fixed four-catalog panel isolates sensitivity to initial evidence: evaluation users, trained scores, labels and catalog support are unchanged.