EXACT RANKING VERIFICATION UNDER SHARED MIN–MAX NORMALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Guaranteeing the stability of learned rankings is challenging when uncertain input features are jointly normalized: changing one candidate's input can alter every candidate's score. Under shared min–max normalization, independently attainable score extremes may be jointly infeasible, making independent bounds insufficient to determine ranking stability. We develop an exact verification framework for piecewise-constant scorers, including tree ensembles, with one uncertain feature per candidate, fixed remaining inputs, and uncertainty specified by pairwise difference constraints. Our central idea is to characterize which pairs of normalized feature values can occur jointly. We prove that this two-dimensional feasible set has boundary complexity linear in the number of candidates, together with a matching worst-case lower bound. This geometric structure enables a complete top-k verifier that either certifies an unchanged selected set or constructs a feasible counterexample, without enumerating all combinations of score-profile regions. Experiments on controlled workloads demonstrate the computational benefits of the geometric construction. Studies on NFCorpus and SciFact show that modeling shared-query dependencies enables ranking guarantees that are unavailable under independent uncertainty, with most certificates obtained through sufficient shared checks.

1 INTRODUCTION

Learning-to-rank systems score candidates using query-dependent features (Liu, 2009), often with tree ensembles such as LambdaMART (Burges, 2010), implemented in frameworks such as LightGBM (Ke et al., 2017). These features may be normalized within each query, as in LETOR (Qin et al., 2010; Qin & Liu, 2013). When an input is uncertain, a natural verification question is whether the same candidates remain in the top-k set throughout the allowed perturbations.

Shared min–max normalization couples this question across candidates: changing one raw feature can change the minimum or range used to normalize every candidate. This coupling, however, limits what independent score bounds can establish. The worst-case score of a selected candidate and the best-case score of an outsider may require incompatible configurations of the shared extrema, making separately attainable score extrema jointly infeasible. Consequently, even exact marginal bounds may fail to certify a ranking that is, in fact, invariant (Figure 2). Convex bounds also support robustness verification through shared softmax normalization (Wei et al., 2023); here we seek an exact characterization for min–max normalization and piecewise-constant scorers.

We study one uncertain scalar per candidate, with the candidate pool, all remaining inputs, scorer and tie priority fixed. Joint uncertainty is supplied as pairwise difference constraints $r_i - r_j \leq B_{ij}$, a standard difference-bound representation (Miné, 2001). In retrieval, such bounds can capture dependencies induced by quantizing one query vector shared by all documents. Fixing the other inputs reduces each tree ensemble to a one-dimensional piecewise-constant score profile, which need not be monotone. The verification task is to certify an unchanged top-k set for every feasible raw vector or return a feasible vector that changes it.

A change in top-k membership occurs exactly when an outsider outranks a selected candidate, with ties resolved by the fixed priority. We therefore characterize the attainable normalized pair image

$$\mathcal{Q}_{ij} = \{(z_i(r), z_j(r)) : r \in \mathcal{P}(B)\},$$

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.

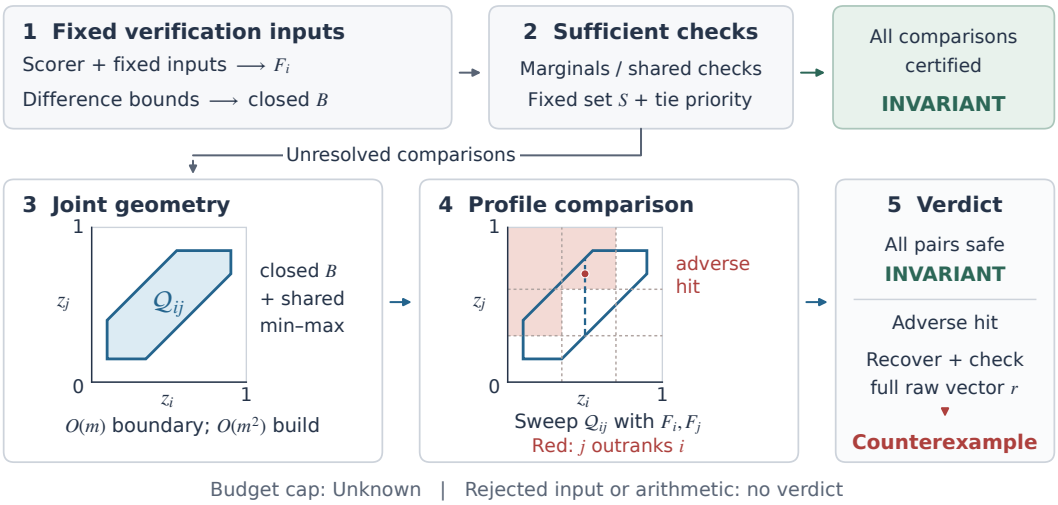


Figure 1: Overview of the signed-zone geometric verifier ($\min_{ij} B_{ij} < 0$). Profiles and difference bounds are separate inputs. If sufficient checks are inconclusive, the verifier intersects each attainable pair image with adverse score cells, respecting cell endpoints and tie priority. Invariance requires excluding every adverse selected–outsider pair; a hit is extended to a checked full raw-vector witness. Geometry is illustrative. Costs are per pair after closure; profile preparation, rational bit costs and witness checking are additional.

where $\mathcal{P}(B)$ is the raw difference zone. A ranking failure requires a selected and an outside candidate to reach adverse score cells together. Up to $m(m-1)$ extrema roles can contribute to this image, yet its boundary has only $O(m)$ straight-line complexity in a signed zone, where some difference bounds are negative. Projective coordinates expose factored caps: one depends only on the minimum candidate, another only on the maximum. Their union therefore has only $O(m)$ frontier pieces, even when the image is nonconvex. This builds on closed-zone geometry and fractional scaling (de la Puente, 2013; Charnes & Cooper, 1962) and separates uncertainty geometry from score-profile processing.

Our contributions are summarized as follows: First, we prove the linear boundary bound with matching full-dimensional examples (Theorem 2). Second, we construct the pair image in $O(m^2)$ arithmetic operations and process score profiles without enumerating their Cartesian product, recovering a full feasible raw vector whenever invariance fails. We also characterize when marginal checks are already complete and derive a faster specialization for interval uncertainty. Third, controlled workloads evaluate computational efficiency, while two retrieval tasks separate shared-constraint gains from the coverage of specified sufficient checks. Persistent-process measurements charge fresh query preparation and counterexample replay, distinguishing uncertainty-model strength from complete-verification cost.

2 Shared Normalization and Difference Uncertainty

There are m candidates. Candidate i has one uncertain raw scalar r_i; all its other model inputs are fixed. The supplied finite rational bounds $r_i - r_j \le B_{ij}$ define the joint uncertainty. We first derive the sharp marginal test, then retain the joint domain for complete verification. Attainment of the marginal bounds uses classical Kleene-star extremals (de la Puente, 2013; Johnson & Kambites, 2015). Figure 1 previews the geometric verification path developed in Section 3.

From contrasts to learned scores. The score profiles consume within-query min–max normalization,

$$z_i = \frac{r_i - \min_j r_j}{\max_j r_j - \min_j r_j}, \qquad z_i = 0 \text{ if all } r_j \text{ coincide.} \tag{1}$$

Separate numerator/range bounds can be loose even for tight contrasts. Instead, close the consistent difference constraints exactly, obtaining finite B with $B_{ii} = 0$ and $B_{ij} \leq B_{ik} + B_{kj}$. Write $\mathcal{P}(B) = \{r : r_i - r_j \leq B_{ij}\}$ and $\psi(d) = \max d/(\max d - \min d)$ for nonzero vectors containing zero, with $\psi(0) = 0$.

Proposition 1 (Attainable normalized coordinate bounds)**.** *The smallest interval containing* $\{z_i : r \in \mathcal{P}(B)\}$ *is*

$$\left[\psi\big((-B_{ji})_j\big),\ \psi\big((B_{ij})_j\big)\right]. \tag{2}$$

Both endpoints are attained, and all intervals cost $O(m^2)$ *arithmetic operations after closure.*

Proof sketch. The anchored contrasts satisfy $-B_{ji} \leq r_i - r_j \leq B_{ij}$. The function ψ is coordinatewise nondecreasing on this domain. Closure makes $r_j = B_{ji}$ and $r_j = -B_{ij}$ feasible, attaining the lower and upper endpoints respectively. Appendix A proves the degenerate cases and the once-rounded machine extension. □

Classical difference-bound closure costs $O(m^3)$ arithmetic operations (Miné, 2001); rational bit costs are additional. These are sharp coordinate hulls over the supplied constraints; they do not describe the exact joint normalized set. Every control receives this refinement.

For a supplied score profile F_i, let $[\alpha_i, \beta_i]$ be its exact score range over this coordinate hull. A proposed set S is invariant if $\alpha_i > \beta_j$, or $\alpha_i = \beta_j$ with i preferred by the fixed priority, for every $i \in S, j \notin S$. Failure is inconclusive because marginal extrema need not coexist. The scalar attainable set may itself have gaps, so a score extremum over its hull need not be attainable. Our interval study uses exact scalar unions instead of this hull relaxation (Appendix F). Section 3 retains the full joint domain for a complete decision.

3 Complete Verification of Normalized Rankings

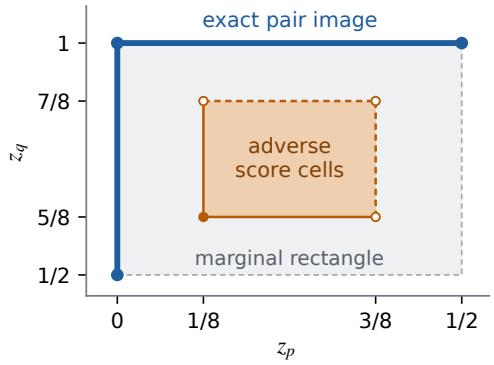
Sharp coordinate hulls can admit incompatible score extrema. We retain the joint domain, with one uncertain scalar per candidate and all other model inputs fixed.

Why joint extrema roles matter. Consider all translates of $(0, q, t_1, t_2)$, where $1 \leq q \leq 5/4$, $-1 \leq t_1, t_2 \leq 2$ and $|t_1 - t_2| \leq 1/4$; call the first candidate p. After normalization, p and q cannot both be interior: otherwise t_1, t_2 would attain both extrema, contradicting a range at least one. Give p score 0 when its once-rounded value lies in $[1/8, 3/8)$ and 1 otherwise; give q score $1/2$ on $[5/8, 7/8)$ and -1 otherwise. The other scores are -2. Thus p is always first, although its exact score range $[0, 1]$ overlaps q's $[-1, 1/2]$. Even the compatible-order, global-range, endpoint-role and coordinatewise role-envelope filters used below leave this decision unresolved (Appendix B.14). This analytical example proves filter incompleteness; computational performance is evaluated separately.

Theorem 2 (Sharp signed pair-image complexity)**.** *Let* B *be a finite, consistent, closed rational difference matrix with* $\min B < 0$*. For distinct candidates* i, j*, the exact normalized pair image* $\mathcal{Q}_{ij}$ *has* $O(m)$ *straight-line complexity, including lower-dimensional parts. It admits at most seven pieces of this total complexity, each with interval fibers in a possibly different coordinate direction. After closure, these pieces can be constructed in* $O(m^2)$ *arithmetic/comparison operations and working words; rational bit costs are additional. The worst-case boundary complexity is* $\Theta(m)$*, even for full-dimensional raw zones.*

The bound concerns geometric output size, not optimal running time (Propositions 9 and 10). The geometry is independent of score-profile cells. The same image can therefore serve different profiles or target sets on an identical indexed zone; Corollary 12 separates amortized construction from fresh profile processing and witness costs.

Supplied profiles and machine specialization. Candidate i supplies P_i ordered, nonempty real intervals forming a disjoint partition of $[0, 1]$, with rational endpoints, exact endpoint ownership, and a fixed exactly comparable score on each cell. Singleton cells and repeated scores are allowed. A binary64 machine profile instantiates this contract through its exact nearest-even real preimages, preserving the original evaluator and ties. The proof applies to the supplied real cells; the checked



Individually attainable, jointly impossible. The exact normalized pair image is

$$(\{0\} \times [1/2, 1]) \cup ([0, 1/2] \times \{1\}).$$

Its marginal rectangle permits p's score 0 and q's score $1/2$. The blue joint image never enters those adverse cells, so p always wins. The raw zone has interior and no fixed pairwise offsets; the obstruction comes from which candidates can attain the shared extrema.

Figure 2: Joint information can certify a ranking that exact marginal score ranges leave unresolved. Solid/dashed orange edges and filled/open corners indicate included/excluded cell boundaries. Rational ticks show machine thresholds; their exact nearest-even preimage shifts are below plot resolution. The exact image and preimages are proved in Appendix B.14.

implementations use this binary64 specialization. A successful marginal test can precede and avoid the complete joint test.

Theorem 3 (Complete ranking verification)**.** *Let B be a finite, consistent, closed rational difference matrix, and let S be a proposed size-k set, $0 < k < m$, with a fixed total tie priority. Put $g = \min B$, $L(P) = 1 + \lceil \log_2(P+1) \rceil$, and $C_{ij} = \min\{P_i + P_j, P_i L(P_j), P_j L(P_i)\}$. Put $D_{ij} = m + P_i + P_j$ and define*

$$W_{ij} = \begin{cases} \min\{m^2 C_{ij},\ m^2 + P_i P_j,\ m^2 + D_{ij} L(D_{ij})\}, & g < 0, \\ P_i + P_j, & g = 0. \end{cases} \tag{3}$$

For the supplied profiles under exact min–max normalization and the zero-vector convention for constants, uniform top-k identity on $\mathcal{P}(B)$ can be decided in

$$O\left(m^2 + \sum_i P_i + \sum_{i \in S,\, j \notin S} W_{ij} \right) \tag{4}$$

exact arithmetic/comparison and word-RAM operations after closure and complete profile construction. Streaming candidate pairs uses $O(m^2 + \sum_i P_i)$ working storage. Failure of invariance yields a full rational potential in $\mathcal{P}(B)$ selecting a different set. Closure, rational bit costs, model-to-profile construction, witness replay and construction of the supplied uncertainty bounds are additional costs.

Corollary 4 (Verification from interval endpoints)**.** *For nonempty closed raw intervals $[L_v, U_v]$ with finite rational endpoints and the supplied-profile contract of Theorem 3, at most eight cell-independent representative extrema pairs suffice, including candidate aliases. The polygon sweep therefore costs $O(P_i + P_j)$ per compared pair even for signed derived zones, after linear endpoint preparation. Complete top-k verification costs $O(m + \sum_v P_v + \sum_{i \in S, j \notin S}(P_i + P_j))$ with $O(m + \sum_v P_v)$ working storage; endpoint input requires no dense matrix or closure (Appendix B.13).*

Thus interval uncertainty admits a simpler complete test than a general signed difference zone.

One scale interval for all extrema roles. Write two nonempty real cells as $I_i, I_j \subseteq [0, 1]$, with endpoints l_i, u_i, l_j, u_j and exclusion flags $s_{li}, s_{ui}, s_{lj}, s_{uj}$. Put $b_u = \min_v B_{uv}$ and $c_v = \min_u B_{uv}$. For reciprocal range $\lambda > 0$, classical closed-zone box consistency (Krivulin, 2014) reduces feasibility before fixing extrema to seven inequalities:

$$\begin{aligned} -1 &\le \lambda g, \\ l_i - 1 &\le \lambda b_i, & -u_i &\le \lambda c_i, \\ l_j - 1 &\le \lambda b_j, & -u_j &\le \lambda c_j, \\ l_i - u_j &\le \lambda B_{ij}, & l_j - u_i &\le \lambda B_{ji}. \end{aligned} \tag{5}$$

A comparison is strict when any participating candidate endpoint is excluded; the first is closed. Intersect these half-lines with $\lambda > 0$ to obtain Λ. If $g < 0$ and Λ is nonempty, its supremum L is finite and positive. Let η indicate that $L \in \Lambda$.

For $g < 0$, every extrema role adds five inequalities monotone in positive scale. Feasibility therefore reduces to the same common supremum, with exact rules for excluded endpoints and zero coefficients (Lemma 8). At closed point cells, a feasible role remains feasible at the attained upper scale.

From extrema roles to a small pair image. Apply Eq. (5) to closed singleton cells $x = z_i, y = z_j$. The common upper scale is attained. Its limiting constraint is global, a candidate row or column, or a cross constraint. For each of these at most seven types, substitute its affine upper scale into all seven point inequalities. Add the redundant lower bound $\lambda \geq 1/R$, where $R = \max B > 0$ bounds every feasible raw range. This defines a closed, constant-complexity convex active patch, with ties allowed between patches. The global patch already forces both extrema.

For a row-i patch, put $d = -b_i > 0$, so $\lambda = (1 - x)/d$, and transform

$$u = \frac{y}{1 - x}, \qquad v = \frac{1 - y}{1 - x}.$$

Closure makes one of the five role inequalities redundant. Each remaining role becomes the downward-closed polygon

$$u, v \geq 0, \quad u \leq A_a, \quad v \leq C_h, \quad u + v \leq V_{ha}, \tag{6}$$

where $A_a = B_{ja}/d$, $C_h = B_{hj}/d$ and $V_{ha} = \min\{B_{ha}, B_{ia} + d\}/d$. Although there can be $m(m-1)$ roles, A_a depends only on the minimum identity and C_h only on the maximum identity. Its upper frontier is $\min\{C_h, V_{ha} - u\}$ on $[0, \min\{A_a, V_{ha}\}]$. The union frontier is nonincreasing. Every positive-length horizontal plateau has one of the at most m heights C_h, and each height occurs on only one connected plateau. A positive downward jump can occur only at one of the at most m cutoffs A_a: a role ending at V_{ha} has height zero. Between plateaus and jumps, all edges have slope -1 and consecutive collinear edges merge. The maximal frontier therefore has $O(m)$ edges, including its closed event fibers and degenerate axis pieces.

For construction, sort the A_a groups in decreasing order and maintain $M_h = \max_{a \text{ active}} V_{ha}$. Between consecutive cutoffs the frontier is $\max_h \min\{C_h, M_h - u\}$. In decreasing C_h order, discard dominated pairs (C_h, M_h); the remaining intercepts strictly increase and their capped-line envelope takes $O(m)$ work. Processing at most m cutoff intervals, with all role updates charged, constructs the union in $O(m^2)$ work. Closed boundaries retain roles that disappear at the next cutoff.

Column and cross patches have analogous projective forms. The scale cutoff keeps their denominators positive. Clipping each union to its active patch and mapping back preserves linear boundary complexity; different active patches meet only on their affine tie lines. The whole pair image can be nonconvex, but retains only $O(m)$ straight-line pieces, proving the upper bound of Theorem 2 (Appendix B.11). Its matching staircase family also consists of full-dimensional signed zones.

Processing profiles without their Cartesian product. A row-i piece has interval fibers in x at fixed y: increasing x increases both transformed coordinates, so a downward-closed union admits a prefix. Column pieces give suffixes; cross pieces permit either direction. Intersecting with the convex active patch preserves the appropriate interval fiber. We use only the direction justified by this argument. Sort the piece's vertex coordinates along its valid scan direction. Closed fibers at these coordinates and open strips between them form $O(m)$ disjoint strata, each a constant-complexity convex set, including lines and points. Merge their ordered scan intervals with one profile. There are $O(m + P_{\text{scan}})$ intersections. Project each onto the other coordinate with its exact endpoint ownership, then query the minimum or maximum score over the intersecting cells (Bender & Farach-Colton, 2000). Binary boundary searches give $O(m^2 + (m + P_i + P_j)\log(2 + m + P_i + P_j))$ work per pair, including geometry builds. The adverse cell intersection supplies a rational point; one final role scan and closed-zone extension recover its full raw witness.

Complementary methods and full witnesses. Active-scale substitution also gives six suffix-maximum tables costing $O(m^2 + P_i P_j)$; this can be favorable when one profile is short (Appendix B.10). Alternatively, each fixed extrema role has a constant-complexity pair polygon. Classical slab queries and range extrema process it in $O(C_{ij})$ work, including both binary-search orientations.

Table 1: Per-pair complexity under the common profile and tie-priority contract. Here $N \leq m(m-1)$ counts eligible extrema roles; the polyhedral method enumerates all extrema pairs and retains the cost of each LP call. Active-scale and grouped-image bounds include pair-dependent construction. Shared closure, profile extraction, rational bit costs and witness replay are additional. The interval method uses at most eight role polygons. These are upper bounds for the stated procedures. The nonnegative order branch remains analytical.

Complete procedure	Per candidate pair, over all roles
General polyhedral cell-pair tests	$O(m^2 P_i P_j)$ exact LP calls
Closed-zone cell-pair enumeration	$O(N P_i P_j)$ operations
Polygon slabs, binary queries	$O(N \min\{P_i L(P_j), P_j L(P_i)\})$
Polygon slabs, batched queries	$O(N(P_i + P_j))$
Active scale, batched signed zones	$O(m^2 + P_i P_j)$
Grouped pair image, signed zones	$O(m^2 + D_{ij} L(D_{ij}))$
Order regions, nonnegative zones	$O(P_i + P_j)$
Interval-derived zones, either sign	$O(P_i + P_j)$

For nonnegative zones, at most eleven fixed endpoint/order regions replace role enumeration and cost $O(P_i + P_j)$ (Appendix B.12); constants add $(0, 0)$. Choosing by supplied sizes gives the hybrid bound, whose automatic dispatch remains analytical. Every positive-range adverse hit supplies a rational scale and extrema; closed-zone extension then recovers a full raw potential. A changed top-k set has an adverse cross pair, including priority equality. The appendix proves endpoint handling and the once-returned $O(m^2)$ witness check.

Charged costs and precision scope. For T threshold-tree nodes, each candidate has at most $T + 1$ cells before merging; extracting and evaluating them is separately charged (Appendix B). Let $N \leq m(m-1)$ count eligible roles. With $N = \Theta(m^2)$ and $P_i = P_j = \Theta(m^2)$, the pair-image bound is $O(m^2 \log m)$ per pair, versus $O(m^4)$ for the polygon and signed-table procedures. For $P_i = P_j = \Theta(m)$, both grouped images and tables cost $O(m^2)$; sparse roles or short profiles can favor polygon sweeps. These bounds include pair-dependent preparation but exclude closure, profile extraction and operand bit costs. Unbounded profile-size regimes use supplied rational partitions or increasing precision; fixed binary64 has finitely many inputs. These are procedure upper bounds, not lower bounds on other verifiers or measured speedups.

Marginal and joint checks in a diagnostic workload. A fixed bank gives 206 queries under paired storage orders, agreeing with an independent raw-range oracle: 114 invariants and 92 checked adverse potentials. Of 80 candidate queries, 34 are invariant; marginal verification resolves 26 and the role envelope resolves all 34. Thus no candidate invariant requires the full index beyond these filters. This workload therefore provides no evidence of incremental certification from complete verification (Appendix B.15).

When marginal verification is complete. There is also an exact no-gain regime: if $m \geq 4$ and all off-diagonal entries of B are positive, every normalized candidate pair in $[0, 1]^2$ is attainable. Exact independent scalar-profile separation is then necessary and sufficient for uniform top-k identity (Appendix A.2). The joint algorithm cannot improve that test there. Completeness concerns the supplied uncertainty domain; an adverse raw-feature vector need not have an application-level realization.

More generally, exact marginals are complete for every score assignment on fixed cells exactly when every selected/outsider cell-reachability relation factors into its marginals (Corollary 7). This finite separability criterion does not imply a gain for a particular trained scorer when factorization fails.

Machine profiles and implementation scope. Reachable scalar splits define binary64 machine profiles with original branch ownership and accumulation order (Törnblom & Nadjm-Tehrani, 2020), charging construction separately. The LightGBM adapter preserves native small-input zero folding and checks complete row profiles against the frozen float64 raw-score interface. Grouped counts coalesced frontier spans within patches, not maximal union edges. The nonnegative order sweep and automatic hybrid remain analytical; contracts, finite-check scope and full raw-vector replay

are in Appendices B and F. Two specified floating preprocessing paths preserve native scores at all 2,838 tested query–representation points, despite feature differences. This finite agreement does not imply uniform equivalence (Appendix K.1). A separate conditional transfer to a specified sequential binary32 pipeline retains 1,654/1,657 reference invariants over all 623 queries (Appendix L).

4 Experiments: solver efficiency and shared uncertainty

We evaluate efficiency on 52 prospectively specified constructed inputs and shared uncertainty on full retrieval populations. Table 2 also retains earlier development comparisons, with their separate protocols in Appendix C.

Workloads and evaluation protocol. The graph-metric cohort supplies 16 signed zones from rings and two-block bridge graphs, with $m \in \{8, 16, 32, 64\}$ and $P \in \{32, 128\}$ cells per profile. Hashed scores determine the nominal top-two set without protected cells or preassigned robustness labels. The controlled-scaling cohort supplies 36 frontier cases spanning core size, profile length and selected-group size: 18 analytically specified invariants and 18 paired counterexamples. Its positive-width offsets share a core; changing group size also changes m. These are supplied-profile inputs. The MiniLM development study uses a fixed NFCorpus (Boteva et al., 2016) LambdaRank model and 16 DEV queries with 32 candidates, 8/16-bit query quantization and $k = 1, 3$; all 512 profiles pass native checks.

We compare grouped pair images with clipped images, Cartesian role indexing, active-scale tables and a Z3 linear-real-arithmetic encoding (de Moura & Bjørner, 2008). Every method pays for common closure, profile preparation and marginal/hull screens, followed by its solver and full witness replay. Each condition has three fresh-process repetitions with rotated order and a five-second adapter CPU cap, including imports. All capped calls remain in the denominators. Appendices C and G retain full protocols, phase costs, failures and the separate offline cost of constructing the MiniLM study's common supplied zones.

Table 2: Conditions completed in all three repetitions under each study's fixed budget. The development column uses the common signed support domain; the MiniLM DEV study's 64 conditions share 16 queries. Completion requires a valid decision within the budget in every repetition.

Method	Dev.	Graph	Controlled inv.	Controlled adverse	MiniLM
Grouped image	58/60	16/16	18/18	18/18	63/64
Clipped image	54/60	16/16	14/18	18/18	62/64
Cartesian index	46/60	16/16	6/18	16/18	63/64
Active tables	30/60	15/16	2/18	4/18	61/64
Z3 (QF-LRA)	51/60	12/16	8/18	18/18	44/64

Results and scope. Grouped completes all 52 constructed cases, versus 48 for clipping and 38 each for Cartesian and Z3 (Figure 3). On the 14 controlled invariants jointly completed with clipping, grouped is faster in every case; the median within-case clipped/grouped CPU ratio is 2.54, including common preparation and imports. All 16 graph-metric cases yield replayed counterexamples; grouped and Cartesian each win eight matched timing comparisons. All 18 controlled invariants agree with their analytical construction, but only eight also receive completed corroboration from another solver family.

External search comparison. A comparison on the same 36 controlled cases evaluates Grouped, Cartesian and VERITAS (Devos et al., 2021) with an exact role-feasibility oracle. They complete 36, 22 and 21 cases in all three repetitions under the same five-second adapter CPU cap. On the 21 jointly completed cases, the median paired VERITAS/Grouped CPU ratio is 5.55. This is adapted native search over supplied profiles, sharing Cartesian's geometric oracle, rather than an independent geometric proof family. On a fixed 128-condition real-query subset all three complete 81 invariants and 47 counterexamples; import-excluded timing gives no general Grouped advantage (Appendix J.2).

MiniLM/Qwen3 DEV have 36/39 invariants among 64 conditions each, all marginal-certified (Appendices G–H).

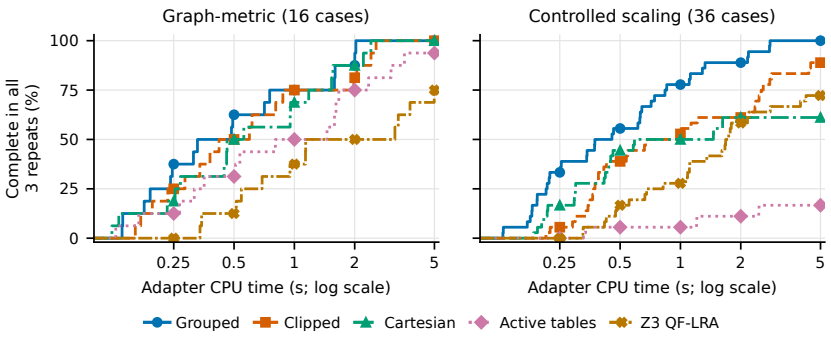


Figure 3: Observed completion under a CPU allowance. Each step occurs at the largest adapter CPU time among three valid repetitions of a case, using the fixed five-second runs. Capped cases remain incomplete; these curves do not represent separately executed timeout experiments.

Shared uncertainty in two retrieval tasks. Separate frozen seven-feature LambdaRank models use BM25 top-32 candidate pools and a frozen MiniLM encoder. NFCorpus and SciFact (Wadden et al., 2020) have 323 and 300 TEST queries; 8/16-bit query quantization and $k \in \{1, 3\}$ give 1,292 and 1,200 conditions. NDCG@10 is 0.330787/0.719613 versus BM25's 0.308057/0.660464. These quality scores characterize the rankers, not an effect of verification. Table 3 compares independent uncertainty with shared constraints and the original sufficient-check baseline. Each of the 89/86 coupling gains has a replayed independent-domain counterexample and a shared-domain invariant; these differences are not caused only by timeouts. Our stronger role-envelope checks also certify the two invariants beyond the original baseline (Appendix K.2). These cases therefore establish shared-constraint gains without requiring complete solving beyond those stronger checks.

Table 3: Invariant conditions for the original frozen models. Independent and shared complete columns use their respective full domains. Shared screens denotes the union of marginal, global-range and joint-hull checks in the original protocol, not all sufficient checks.

Dataset	Conditions	Independent	Shared screens	Shared complete
NFCorpus	1,292	735	823	824
SciFact	1,200	747	832	833

A bounded common hull-screen policy lets both Grouped and Cartesian finish all 1,292 NFCorpus conditions (824 invariants, 468 counterexamples); the two recovered outcomes are counterexamples, not additional certificates. All three image/index backends finish SciFact's 1,200 conditions with 833 invariants and 367 counterexamples. Timing and preparation contracts are separated in Appendix J.

Persistent verification. The constructed completion gaps persist in warm processes. A separate three-round, full-population comparison includes fresh query preparation and native counterexample replay. All four methods complete every condition. Grouped's total request wall time is $1.010\times$ and $1.028\times$ that of the fastest comparator, adapted VERITAS, on NFCorpus and SciFact. Thus the controlled geometric gains do not translate into an overall advantage on these $m = 32$ workloads (Appendix K.2).

Query witnesses and quality targets. Across an expanded five-precision NFCorpus study, 559 of 3,230 conditions have audited common-query vector counterexamples; SciFact has 186 of 1,200. These are witnesses in the quantizer cells, not text perturbations; unconfirmed searches remain unknown (Appendix J.4). For validation-selected rankers with known offline relevance labels, fixed-anchor verification certifies positive NDCG@1 lower bounds for 116/323 NFCorpus queries versus 114 for sufficient checks. A separate full relevance-set checker adds two certificates, allowing the good winner to vary with the raw input; SciFact remains at 159/300. These guarantees preserve the midpoint grade without changing predictions (Appendix J.7).

Interval specialization. Complete enumeration on 480 LETOR TRAIN-development conditions finds 430 invariants (355 from exact marginal checks and 75 trivial full-set conditions) and 50 counterexamples. Joint reasoning adds none; the study checks the interval application and native profiles (Appendix F).

5 Related work

Tree-ensemble verification. Exact adversarial optimization, abstract interpretation and constrained search provide complementary approaches to verifying tree ensembles (Kantchelian et al., 2016; Ranzato & Zanella, 2020; Chen et al., 2019). VERITAS already optimizes coupled model-output differences using complete constrained search (Devos et al., 2021, Eq. 2, Sections 3.4–3.7). Our comparison supplies scalar profiles and an exact normalized role-feasibility oracle, charging both the oracle and common preparation (Appendix J). Our contribution is the specialized geometric representation and processing bound, rather than the ability to express coupled feasibility. Feature-sensitivity verification uses pseudo-Boolean and data-aware MILP/SMT formulations to search over input pairs constrained to agree outside a designated feature set (Ahmad et al., 2025; Varshney et al., 2026). Those shared coordinates remain search variables; our setting fixes all other inputs and studies scalar profiles coupled by normalization and difference constraints.

Certified ranking and shared normalization. Certified ranking methods address related stability questions for neural rankers (Wu et al., 2022a), while tree-ensemble verification provides exact optimization and search procedures (Kantchelian et al., 2016; Chen et al., 2019; Devos et al., 2021). Ranking preservation through smoothing, quality-aware retrieval guarantees, and multiclass tree verification are established (Wu et al., 2022a;b; Devos et al., 2024); convex softmax bounds address shared-normalization verification (Wei et al., 2023). We study exact top-k set identity under a shared min–max normalizer. The uncertainty acts on numerical features; the guarantees do not assume that every allowed vector corresponds to a textual perturbation.

Difference constraints and geometric algorithms. Closed-DBM saturation and exact raw-coordinate projection are classical (Miné, 2001, Theorems 2 and 4). Kleene-star representations describe raw difference polytopes (de la Puente, 2013, Theorem 2.1); tropical and Euclidean convexity are related by Johnson & Kambites (2015, Theorems A–B). Coordinate range is the tropical norm modulo common translation, with unit-ball facets indexed by ordered extrema (Criado et al., 2022). Our pair image divides by a varying common range. Fractional scaling and box feasibility provide the role reductions (Charnes & Cooper, 1962; Krivulin, 2014). The additional structural argument is that factored role caps restrict the entire signed pair image to $O(m)$ boundary complexity, with a matching lower bound and coordinate directions suitable for sweeping. Envelope and range-query methods then process the profiles (Hershberger, 1989; Lueker, 1978; Bender & Farach-Colton, 2000). The interval specialization needs only a constant number of role polygons.

6 Conclusion and limitations

Shared min–max normalization yields pair images of tight linear boundary complexity and a complete verifier with raw witnesses for signed difference zones. Appendix M gives affine-profile reductions and separate exact feasibility formulations for softmax and nonnegative ℓ_1 normalization; our sharp complexity bounds and implemented verifier remain specific to min–max. Controlled workloads show computational gains. On retrieval tasks, shared constraints improve certification, but our role-envelope checks recover both identity increments over the original baseline, with no overall speed advantage over the fastest complete comparator.

Guarantees require complete scalar profiles, fixed inputs and stated arithmetic. Outer-zone witnesses need not lift to common queries or text; failed lifting remains unresolved. Preparation, rational arithmetic and replay can dominate cost. Fixed pools limit relevance coverage. Offline quality diagnostics use labels and do not improve predictions. Later TRAIN/DEV-selected studies followed original TEST outcomes and are not untouched confirmations. LETOR uses TRAIN-development queries; TEST was checked only for identity and schema.

REPRODUCIBILITY STATEMENT

The appendices provide complete proofs, arithmetic and endpoint conventions, experimental protocols, model configurations, timing definitions and full outcome counts. The accompanying artifacts retain verifier implementations, model and input identifiers, result records, counterexample replays and figure-generation code. Constructed comparisons, encoder studies and full retrieval populations have separate protocols and denominators. All 623 queries in the offline quality-bound study are retained, including unresolved conditions and matched inexpensive controls. Dataset acquisition and preprocessing instructions accompany the records. The reproduction instructions distinguish reaggregation of saved results from rerunning training and native verification: restoring and checking records establishes their integrity and the reported aggregates, but does not independently reprove saved invariant judgments. Full reruns additionally require the documented datasets, model dependencies and native solver environments.

AI USE STATEMENT

Generative AI was used to review the paper draft and code implementation; identify literature; and optimize Figure 1 in the paper. AI-assisted checks are included in the supplementary artifacts. These checks do not substitute for human author verification. Responsibility for the manuscript, proofs, code, data provenance, and disclosure rests with the authors.

ETHICS STATEMENT

The retrieval experiments use public NFCorpus, SciFact and LETOR data; the controlled experiments use generated inputs and graphs. Dataset attribution and licensing requirements apply. NFCorpus contains medical information, but the experiments evaluate ranking stability and do not establish clinical validity. Certificates are conditional on the stated numerical uncertainty sets and fixed candidate pools, and do not guarantee factual correctness, fairness or relevance outside those conditions.

REFERENCES

Arhaan Ahmad, Tanay V. Tayal, Ashutosh Gupta, and S. Akshay. Sensitivity verification for additive decision tree ensembles. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=h0vC0fm1q7.

Michael A. Bender and Martín Farach-Colton. The LCA problem revisited. In *LATIN 2000: Theoretical Informatics*, volume 1776 of *Lecture Notes in Computer Science*, pp. 88–94. Springer, 2000. doi: 10.1007/10719839_9.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval*, volume 9626 of *Lecture Notes in Computer Science*, pp. 716–722. Springer, 2016. doi: 10.1007/978-3-319-30671-1_58.

Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11 (23-581):81, 2010.

Abraham Charnes and William W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3–4):181–186, 1962. doi: 10.1002/nav.3800090303.

Hongge Chen, Huan Zhang, Si Si, Yang Li, Duane Boning, and Cho-Jui Hsieh. Robustness verification of tree-based models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Francisco Criado, Michael Joswig, and Francisco Santos. Tropical bisectors and Voronoi diagrams. *Foundations of Computational Mathematics*, 22(6):1923–1960, 2022. doi: 10.1007/s10208-021-09538-4. URL https://link.springer.com/article/10.1007/s10208-021-09538-4.

María Jesús de la Puente. On tropical Kleene star matrices and alcoved polytopes. *Kybernetika*, 49 (6):897–910, 2013. URL https://www.kybernetika.cz/content/2013/6/897.

Leonardo de Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems*, volume 4963 of *Lecture Notes in Computer Science*, pp. 337–340. Springer, 2008. doi: 10.1007/978-3-540-78800-3_24.

Mike Develin and Bernd Sturmfels. Tropical convexity. *Documenta Mathematica*, 9:1–27, 2004. URL https://arxiv.org/abs/math/0308254.

Laurens Devos, Wannes Meert, and Jesse Davis. Versatile verification of tree ensembles. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2654–2664. PMLR, 2021. URL https://proceedings.mlr.press/v139/devos21a.html.

Laurens Devos, Lorenzo Cascioli, and Jesse Davis. Robustness verification of multi-class tree ensembles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21019–21028, 2024. doi: 10.1609/aaai.v38i19.30093.

John Hershberger. Finding the upper envelope of n line segments in $O(n \log n)$ time. *Information Processing Letters*, 33(4):169–174, 1989. doi: 10.1016/0020-0190(89)90136-1.

Thomas S. Hune, Judi Romijn, Mariëlle Stoelinga, and Frits W. Vaandrager. Linear parametric model checking of timed automata. Technical Report RS-01-5, BRICS, Department of Computer Science, University of Aarhus, 2001. URL https://tidsskrift.dk/brics/article/view/20459.

Marianne Johnson and Mark Kambites. Convexity of tropical polytopes. *Linear Algebra and its Applications*, 485:531–544, 2015. doi: 10.1016/j.laa.2015.08.006.

Michael Joswig and Katja Kulas. Tropical and ordinary convexity combined. *Advances in Geometry*, 10(2):333–352, 2010. doi: 10.1515/advgeom.2010.012.

Alex Kantchelian, J. D. Tygar, and Anthony Joseph. Evasion and hardening of tree ensemble classifiers. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2387–2396. PMLR, 2016. URL https://proceedings.mlr.press/v48/kantchelian16.html.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Nikolai Krivulin. A constrained tropical optimization problem: Complete solution and application example. In *Tropical and Idempotent Mathematics and Applications*, volume 616 of *Contemporary Mathematics*, pp. 163–177. American Mathematical Society, 2014. doi: 10.1090/conm/616/12308.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. doi: 10.1561/1500000016.

George S. Lueker. A data structure for dynamic range queries. Technical Report 123, Department of Information and Computer Science, University of California, Irvine, September 1978. URL https://escholarship.org/uc/item/9k40c7jc.

Antoine Miné. A new numerical abstract domain based on difference-bound matrices. In *Programs as Data Objects*, volume 2053 of *Lecture Notes in Computer Science*, pp. 155–172. Springer, 2001. doi: 10.1007/3-540-44978-7_10.

Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013. URL https://arxiv.org/abs/1306.2597.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010.

Francesco Ranzato and Marco Zanella. Abstract interpretation of decision tree ensemble classifiers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4):5478–5486, 2020. doi: 10.1609/aaai.v34i04.5998.

Siegfried M. Rump. Error estimation of floating-point summation and dot product. *BIT Numerical Mathematics*, 52(1):201–220, 2012. doi: 10.1007/s10543-011-0342-4.

Richard P. Stanley. Two poset polytopes. *Discrete & Computational Geometry*, 1(1):9–23, 1986. doi: 10.1007/BF02187680.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

John Törnblom and Simin Nadjm-Tehrani. Formal verification of input-output mappings of tree ensembles. *Science of Computer Programming*, 194:102450, 2020. doi: 10.1016/j.scico.2020.102450.

Namrita Varshney, Ashutosh Gupta, Arhaan Ahmad, Tanay V. Tayal, and S. Akshay. Data-aware and scalable sensitivity analysis for decision tree ensembles. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.07453.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609.

Dennis Wei, Haoze Wu, Min Wu, Pin-Yu Chen, Clark Barrett, and Eitan Farchi. Convex bounds on the softmax function with applications to robustness verification. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6853–6878. PMLR, 2023. URL https://proceedings.mlr.press/v206/wei23c.html.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Wei Chen, Yixing Fan, Maarten de Rijke, and Xueqi Cheng. Certified robustness to word substitution ranking attack for neural ranking models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2128–2137. Association for Computing Machinery, 2022a. doi: 10.1145/3511808.3557256.

Yihan Wu, Hongyang Zhang, and Heng Huang. RetrievalGuard: Provably robust 1-nearest neighbor image retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24266–24279. PMLR, 2022b. URL https://proceedings.mlr.press/v162/wu22o.html.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

SUPPLEMENT GUIDE

The supplement contains the attained marginal geometry, complete normalized verification (including the signed pair-image construction), the generated computational comparison, and the fixed LETOR and NFCorpus DEV and TEST studies. Appendix J collects the SciFact task and subsequent retrieval, external-search and offline quality-bound evidence. Appendix K reports numerical conformance and persistent-process verification measurements. Appendix L reports full-population certificate retention for the specified binary32 pipeline. Exact endpoint ownership, arithmetic and implementation status, full witness recovery and all completed LETOR conditions are retained.

A SHARP COORDINATE BOUNDS AND LIMITS OF MARGINAL VERIFICATION

Difference-bound matrix closure and saturation are classical (Miné, 2001, Section 3). The following normalized-coordinate result uses their attained anchored extremals (de la Puente, 2013; Johnson & Kambites, 2015) to characterize the attainable normalized-coordinate endpoints.

A.1 ATTAINABLE ENDPOINTS FROM A CLOSED DIFFERENCE MATRIX

Let $U \in \mathbb{R}^{m\times m}$ be finite, with $U_{ii} = 0$ and $U_{ij} \le U_{ik} + U_{kj}$ for all i, j, k. Define

$$\mathcal{P}(U) = \{r \in \mathbb{R}^m : r_i - r_j \le U_{ij} \text{ for all } i, j\}, \qquad \mathcal{N}_i(r) = \frac{r_i - \min_j r_j}{\max_j r_j - \min_j r_j}, \tag{7}$$

with $\mathcal{N}(r) = 0$ for a constant vector. For a vector d containing zero, write

$$\psi(d) = \begin{cases} \dfrac{\max_j d_j}{\max_j d_j - \min_j d_j}, & \max_j d_j > \min_j d_j, \\ 0, & d = 0. \end{cases} \tag{8}$$

Proposition 5 (Sharp coordinate hull). *For each i, the smallest interval containing $\{\mathcal{N}_i(r) : r \in \mathcal{P}(U)\}$ is*

$$[\ell_i, u_i] = \left[\psi\big((-U_{ji})_{j=1}^m\big), \psi\big((U_{ij})_{j=1}^m\big)\right]. \tag{9}$$

Both endpoints are attained. After exact closure, all m intervals require $O(m^2)$ arithmetic operations.

Proof. The function ψ is coordinatewise nondecreasing on vectors containing zero. Indeed, for $d \le e$, set $A_d = \max_j d_j \ge 0$ and $B_d = -\min_j d_j \ge 0$. Then $A_d \le A_e$ and $B_d \ge B_e$; cross multiplication gives $A_d/(A_d + B_d) \le A_e/(A_e + B_e)$ when both denominators are positive. If $d = 0$, its value is zero; if $e = 0$, then $d \le 0$ and both values are zero. Thus the convention in Equation (8) covers the degenerate cases.

For any feasible r, fix i and set $d_j = r_i - r_j$. These contrasts contain zero and satisfy $-U_{ji} \le d_j \le U_{ij}$, while $\mathcal{N}_i(r) = \psi(d)$. Monotonicity proves containment. For upper attainment choose $r_j^+ = -U_{ij}$. Closure gives $r_j^+ - r_k^+ = U_{ik} - U_{ij} \le U_{jk}$, so $r^+ \in \mathcal{P}(U)$; its contrasts about i are the entire upper row. For lower attainment choose $r_j^- = U_{ji}$. Then $r_j^- - r_k^- = U_{ji} - U_{ki} \le U_{jk}$, and its contrasts are the negative column. These feasible vectors attain the stated endpoints. Scanning each row and column establishes the operation count. □

Sharpness and scope. Equation (9) is the exact *coordinate interval hull* over $\mathcal{P}(U)$. Different coordinates can have different attaining vectors. Their Cartesian product need not equal the joint normalized feasible set. Even one coordinate's attainable set can be disconnected: for two candidates with $|r_1 - r_2| \le 1$, it is $\{0, 1\}$, with hull $[0, 1]$. Exact hulls therefore need not decide every learned comparison. The constant-vector case must remain included whenever feasible.

Separate numerator/range propagation can be looser: $1 \le r_2 - r_1 \le 2$ gives $\mathcal{N}_2(r) = 1$, whereas separate ratio propagation gives $[1/2, 1]$. A scalar tree threshold at $3/4$ distinguishes these intervals.

Exact closure and the machine reference. For supplied finite lower/upper difference bands L, H, first form $U_{ij} = \min\{H_{ij}, -L_{ji}\}$, impose a zero diagonal and compute exact shortest-path closure, rejecting inconsistent bands. Rational Floyd–Warshall takes $O(m^3)$ arithmetic operations; rational bit costs are additional. For the canonical input $\mathrm{RN}_{64}(\mathcal{N}_i(r))$, monotonicity of round-to-nearest, ties-to-even gives the enclosing interval $[\mathrm{RN}_{64}(\ell_i), \mathrm{RN}_{64}(u_i)]$. Its endpoints remain attained over the abstract difference constraints. Normalization is exact before the single binary64 conversion.

A.2 Positive difference bounds make independent profiles complete

Fix the closed zone $\mathcal{P}(U)$, candidate-wise scalar machine scores, and total tie priority. Uniform certification concerns all of $\mathcal{P}(U)$, under exact normalization and once-rounded binary64 conversion.

Proposition 6 (Independent profiles are complete for positive bounds)**.** *Suppose $m \geq 4$ and $U_{ij} > 0$ for every $i \neq j$. Every pair of distinct normalized coordinates attains all of $[0,1]^2$. Consequently, for fixed candidate-wise scalar machine scores, a set S of size k is uniformly top-k if and only if every selected score minimum exceeds every unselected score maximum, or equals it with priority favoring the selected candidate. Extrema are exact whole-profile extrema on binary64 $[0,1]$.*

Proof. Assign any two coordinates their desired values and two others zero and one; put the remaining coordinates in $[0,1]$. For $0 < T \leq \min_{i \neq j} U_{ij}$, $r = Tz$ is feasible and $\mathcal{N}(r) = z$. In particular, every binary64 pair is attained unchanged by rounding. Thus every selected minimum and unselected maximum are jointly attainable; the stated priority-aware separation is both necessary and sufficient. □

The conclusion concerns the supplied difference zone and need not hold with nonpositive off-diagonal bounds.

B COMPLETE FINITE-PROFILE VERIFICATION: PROOF AND IMPLEMENTATION SCOPE

We prove Theorem 3. Throughout, $\mathcal{P}(B) = \{r : r_u - r_v \le B_{uv}\}$, the matrix is finite and exactly closed, and $B_{uu} = 0$. Candidate i supplies an ordered, disjoint cover of real $[0,1]$ by P_i nonempty rational-endpoint intervals with exact endpoint ownership and fixed exactly comparable scores. Singletons and repeated scores are permitted. Priority resolves score equality. Profiles are inputs to the theorem; model extraction, quantizer preimages and operand bit costs are separate charges. Binary64 profiles specialize this contract through the exact nearest-even preimages described below; the generated checks cover that specialization, not a new generic-quantizer implementation.

B.1 WHEN CELLWISE MARGINALS ARE COMPLETE

Corollary 7 (Cellwise criterion for universal marginal completeness)**.** *Fix a nonempty normalized image $\mathcal{Z} \subseteq [0,1]^m$, including the zero vector whenever constant raw inputs are feasible. Fix a proposed set S with $0 < |S| < m$, total tie priority, and each candidate's finite disjoint cell partition of $[0,1]$ with exact endpoint ownership. Let A_i be the cells of candidate i reached by some $z \in \mathcal{Z}$, and $J_{ij} \subseteq A_i \times A_j$ the cell pairs reached jointly. The priority-aware exact marginal score test is equivalent to uniform top-$|S|$ membership of S for every assignment of exactly comparable real scores to the fixed cells if and only if*

$$J_{ij} = A_i \times A_j \qquad (i \in S,\ j \notin S).$$

Proof of Corollary 7. Write I_{iu} for cell u of candidate i. Precisely,

$$A_i = \{u : \exists z \in \mathcal{Z},\ z_i \in I_{iu}\}, \qquad J_{ij} = \{(u,v) : \exists z \in \mathcal{Z},\ z_i \in I_{iu},\ z_j \in I_{jv}\}.$$

For assigned cell scores s_{iu}, put $\alpha_i = \min_{u \in A_i} s_{iu}$ and $\beta_i = \max_{u \in A_i} s_{iu}$. These extrema exist and are attained: the domain is nonempty and each partition is finite and covers $[0,1]$. Write $i \prec j$ when the fixed priority favors i. The marginal test requires, for every cross pair, $\alpha_i > \beta_j$, or $\alpha_i = \beta_j$ and $i \prec j$. This test always implies uniform membership because each selected candidate then outranks each outsider at every feasible vector.

Suppose all cross-pair relations are products. If the marginal test fails on i, j, choose cells attaining α_i, β_j. Their pair is jointly reachable, so some feasible vector makes j outrank i, either strictly in score or at equality by priority. Its top-$|S|$ set cannot be S. This proves completeness for every score assignment.

Conversely, suppose some cross pair misses $(u,v) \in A_i \times A_j$. Give i score 0 on cell u and 1 on every other cell; give j score $1/2$ on cell v and -1 on every other cell. Give every other selected candidate constant score 2, and every other outsider constant score -2. The two special cells never co-occur, so i strictly outranks j at every feasible vector. The other constant scores ensure that the top-$|S|$ set is always S, independently of priority. Yet individual reachability gives $\alpha_i = 0 < 1/2 = \beta_j$, so the exact marginal test fails. All assigned scores are rational and exactly comparable. This proves necessity. Constant inputs and singleton cells or excluded endpoints need no separate exception: their exact reachability was included in A_i, J_{ij}. □

B.2 RELATION TO CLASSICAL ZONE GEOMETRY

The raw zone is the min-plus column space of the Kleene star B, equivalently the max-plus column space of $-B^\top$. Its anchored coordinatewise least and greatest vectors are classical dominators and their duals (de la Puente, 2013; Johnson & Kambites, 2015). Specifically, in the gauge $r_i = 0$ they are $(-B_{ij})_j$ and $(B_{ji})_j$. Applying the monotone anchored-contrast functional in Proposition 1 gives its attained coordinate hull as an explicit corollary of this structure. Thus the coordinate hull follows from established difference-bound geometry.

Likewise, the reciprocal-range substitution below is classical linear-fractional programming (Charnes & Cooper, 1962), and general constrained comparisons of ensemble outputs predate this work (Devos et al., 2021). The specialized result is complete processing of the coupled normalized profiles with the stated work bound and full adverse-potential recovery. Generic fixed-dimensional ensemble verification is already polynomial (Chen et al., 2019); here the raw uncertainty dimension grows with

m, while a ranking failure reduces to a pair after conditioning extrema. The bound applies to this supplied-profile decision problem. It does not imply optimality or measured acceleration over other verification algorithms.

Polynomial decidability alone is not specific to difference zones. For a general rational polyhedral raw domain $Ar \le b$, fix extrema roles and an adverse pair of profile cells. Writing $r = Tz + c\mathbf{1}$, with $\lambda = 1/T > 0$ and $\beta = c/T$, gives the linear constraints

$$Az + \beta A\mathbf{1} \le \lambda b, \qquad 0 \le z \le 1, \quad z_a = 0, \quad z_h = 1,$$

together with the two supplied real cells. Retain the strict constraint $\lambda > 0$; replacing it by nonnegativity can admit recession directions that have no finite-range preimage. Strict linear feasibility can be tested by maximizing a common slack bounded above by one and accepting only a positive optimum. The constant branch separately tests whether some $r = c\mathbf{1}$ is feasible. Enumerating roles and adverse cell pairs therefore gives a polynomial number of rational linear-feasibility problems, each in the original dimension plus two variables. The special benefit of the closed difference structure is eliminating that growing-dimensional feasibility problem *before* processing cells, then using ordered polygon queries, factoring common scale constraints to reuse roles, or constructing the entire pair image with linear boundary complexity (Section B.11). Classical box feasibility (Krivulin, 2014), parametric strict difference constraints (Hune et al., 2001), and static range aggregation (Lueker, 1978) supply the underlying tools. No exponential-to-polynomial separation from the generic procedure is claimed.

B.3 Constant-time access to conditioned constraints

Let

$$c_v = \min_u B_{uv}, \qquad b_u = \min_v B_{uv}, \qquad g = \min_{u,v} B_{uv}.$$

These arrays take $O(m^2)$ work once. Requiring a to be a minimum adds $r_a - r_v \le 0$ for every v. This is consistent exactly when $c_a \ge 0$: necessity follows from $0 \le r_v - r_a \le B_{va}$, and sufficiency from the feasible column potential $r_v = B_{va}$. Its exact closure is

$$M^a_{uv} = \min\{B_{uv}, B_{ua} + c_v\}. \tag{10}$$

Indeed a path using a new zero edge can be reduced to one passing through a; further visits remove nonnegative cycles. Minimizing the remaining endpoint gives c_v. This is the ordinary shortest-path update for a closed difference matrix (Miné, 2001).

The row minima of this matrix are available without constructing it:

$$d^a_u = \min_v M^a_{uv} = \min\{b_u, B_{ua} + g\}.$$

Requiring h also to be a maximum is consistent exactly when $d^a_h \ge 0$. Necessity follows from $0 \le r_h - r_v \le M^a_{hv}$; the feasible negative row potential $r_v = -M^a_{hv}$ proves sufficiency. The final closure is

$$N^{a,h}_{uv} = \min\{M^a_{uv}, d^a_u + M^a_{hv}\}. \tag{11}$$

The same path argument proves this identity. Thus any requested entry costs constant work after the initial row/column/global minima. Discard $a = h$, inconsistent roles, and $N^{a,h}_{ha} \le 0$. The latter cannot support a positive range; if it is positive, a column potential attains that range. No dense matrix is constructed per role in the stated bound.

B.4 Exact pair image and extension

Fix valid roles a, h and candidates i, j. Suppress the superscript on N, and assign the four labeled normalized values $q_a = 0, q_h = 1, q_i = x, q_j = y$. Their exact pair image is

$$Q_{a,h,i,j} = \left\{(x,y) \in [0,1]^2 : \exists \lambda \ge 0, \quad q_u - q_v \le \lambda N_{uv} \quad (u,v \in \{a,h,i,j\})\right\}. \tag{12}$$

Repeated labels remain consistent: the opposite zero-diagonal inequalities force the corresponding values equal. For example, $i = a$ forces $x = 0$. The h, a inequality gives $1 \le \lambda N_{ha}$, hence $\lambda \ge 1/N_{ha} > 0$; allowing nonnegative scale introduces no spurious zero-scale solution. Splitting positive, negative and zero scale coefficients and comparing their lower and upper bounds eliminates one variable from at most 16 inequalities. The result is a closed polygon with a constant number of

rational inequalities, including empty, segment and point cases. Its size does not depend on m or the profile sizes.

For exactness, any full potential with these roles supplies a point of (12) by $\lambda = 1/(r_h - r_a)$. Conversely, given a feasible (x, y, λ), prescribe $r_u = q_u/\lambda$ on the distinct retained indices $J = \{a, h, i, j\}$. Set

$$r_v = \min_{u \in J}(N_{vu} + r_u). \tag{13}$$

The retained pair constraints ensure this preserves their values. Triangle closure gives $r_v - r_w \le N_{vw}$ for every pair. The conditioned constraints make a, h actual extrema, so normalization recovers x, y. This proves both exact projection and full witness extension.

Constants belong to $\mathcal{P}(B)$ exactly when every entry of B is nonnegative. Their normalized vector is zero and is checked separately. They are not represented by setting $\lambda = 0$ in the nonconstant system.

B.5 Rounding cells and exact projected endpoints

The geometric arguments accept the supplied rational real cells directly. For the binary64 specialization, construct those cells as follows. For a machine cell $[\ell, u]$, its nearest-even real preimage in $[0, 1]$ has lower endpoint zero if $\ell = 0$, otherwise $(\mathrm{pred}(\ell) + \ell)/2$; its upper endpoint is one if $u = 1$, otherwise $(u + \mathrm{succ}(u))/2$. Interior midpoints belong to the cell exactly when the endpoint float has even significand. All midpoints are computed as exact rationals, including subnormals and powers of two. Zero is canonicalized. These preimages form an ordered disjoint partition of the real unit interval. No change of the model's branch or accumulation semantics is permitted (Törnblom & Nadjm-Tehrani, 2020).

Intersect $Q_{a,h,i,j}$ with a slab $x \in I$ given by one real cell. Eliminating x projects it to an interval J_I in y. Strict lower/upper constraints produce a strict comparison whenever either participating constraint is strict. Zero-coefficient inequalities retain their strictness. Thus an open zero-width result is empty, and a singleton survives only when both endpoint constraints own it. Simply closing the slab and projecting polygon vertices would be unsound on excluded boundary segments. Since the polygon has constant size, exact projection takes constant arithmetic work.

The competing profile cells intersecting a nonempty J_I form one contiguous index range. A segment tree with binary endpoint search therefore already gives $O(P_i \log(P_j + 1))$ processing, or the symmetric orientation. The following batching argument obtains the linear balanced-profile bound.

B.6 Why four monotone scans suffice

For a nonempty polygon with nontrivial x projection D, define

$$f(x) = \min\{y : (x, y) \in Q\}, \qquad h(x) = \max\{y : (x, y) \in Q\}.$$

Compact convex polygon geometry makes f convex and h concave, continuous and piecewise affine on D. The projected lower and upper values for a slab are respectively $\inf f$ and $\sup h$ on its allowed intersection with D. An endpoint is included exactly when an allowed point attains the extremum.

Order lower cuts by increasing value, with closed before open at equality. The sequence of projected lower cuts over successive nonempty slabs is nonincreasing and then nondecreasing. The minimizers of f form an interval; f strictly decreases to its left and increases to its right. Slabs meeting the interior of a minimizer interval attain the common minimum. At either boundary an adjacent slab can approach the value without owning it, placing an open cut outside the closed minimum cuts in the stated order. A singleton minimizer belongs to exactly one profile cell and satisfies the same rule. Thus endpoint ownership does not introduce extra turns.

Order upper cuts by increasing value, with open before closed at equality. Applying the preceding argument to $-h$ makes their sequence nondecreasing and then nonincreasing. Clipping preserves the order of nonempty slabs. A singleton D belongs to only one slab; segments and points are therefore covered as well.

First and last intersecting competing-cell indices are monotone functions of these cuts. Split each cut sequence at its turn and merge each of its two monotone runs with the competing boundaries. At most four complete boundary scans suffice, costing $O(P_i + P_j)$ exact comparisons. Equality

requires ownership by both intervals. Classical static range-minimum/maximum structures use linear preprocessing and constant-time queries on a word RAM (Bender & Farach-Colton, 2000). A query returns both the extreme score and an attaining cell, with a fixed index tie convention.

For each scanned selected-candidate cell, query the maximum competing score. It is adverse if it is larger, or equal with better candidate priority. Scanning the outsider instead uses a range minimum. If only a strict score certificate is requested, every equality is adverse regardless of priority.

B.7 Completeness, recovery and total work

An invariant decision passes every cross-pair test. Conversely, a different top-k set replaces some $i \in S$ by some $j \notin S$ that outranks it. The responsible potential supplies actual minimum/maximum roles and profile cells; their feasible adverse pair is therefore checked. Along with the constant branch, this proves the equivalence in Theorem 3.

A proposed set need not be supplied by a nominal input. To decide whether *any* uniform set exists, take a feasible closed-column potential $r_u = B_{uv}$ for any fixed v, normalize it, and obtain its top-k set from the supplied real profiles (which already encode rounding in the binary64 specialization). If a feasible counterexample refutes that set, no other set can be uniform, since a uniform set must also be selected at this feasible potential. This initialization fits within the shared preprocessing bound.

For recovery, choose rational y in the nonempty intersection of the attaining competing cell and projected interval, respecting strict endpoints. Intersect the slice of the original polygon with the scanned cell to recover a rational x. The remaining scale system has only non-strict inequalities and a positive attained maximum lower bound. Choose that bound, even when the upper scale limit is infinite. Equation (13) then gives the full rational adverse potential in $O(m)$ arithmetic work. Checking all original inequalities costs $O(m^2)$, followed by normalization and profile evaluation; the binary64 specialization also checks rounding and replays the original scorer. Such a witness is feasible for the abstract zone; additional application constraints may restrict which such vectors are realizable.

Shared role preparation costs $O(m^2)$ after closure, and all range structures cost $O(\sum_i P_i)$ once. There are at most $m(m-1)$ roles and $k(m-k)$ candidate cross pairs, giving the polygon branch of (4). Streamed geometry uses $O(m^2 + \sum_i P_i)$ storage apart from the model, proof outputs and optional replay buffers. The word size must address the profile arrays. Exact rational coefficients and midpoint comparisons have additional bit costs. A bounded number of arithmetic operations per geometric step does not make arbitrary-precision arithmetic constant-time.

For highly unequal profile sizes, retaining the binary-search orientation allows the smaller bound $\min\{P_i + P_j,\ P_i L(P_j),\ P_j L(P_i)\}$, where $L(P) = 1 + \lceil \log_2(P+1) \rceil$, per pair and role. This is C_{ij} in Theorem 3; the checked linear prototype uses the batched scans. These bounds compare algorithmic procedures; their practical costs are evaluated separately.

Charging tree-model profile construction. For numerical threshold trees with constant leaf values, complete profiles can be constructed in polynomial work as well. Let T be the total number of model nodes, K the number of trees, d their maximum depth, F the input width, and H the cost of the fixed output postprocessing per point. Assume that the variable scalar affects the output only through the selected leaves; all other feature values are fixed. For candidate i, a traversal with reachable scalar intervals visits each node at most once and collects at most T cut records. Sort and deduplicate these into $C_i \le T$ distinct starts, evaluate one machine representative of each induced cell using the original evaluator, and merge adjacent equal-score cells. Thus $P_i \le C_i + 1$ and one conservative deterministic bound is

$$O\left(\sum_{i=1}^{m} [T \log(T+1) + (C_i+1)\{F + K(d+1) + H\}]\right). \tag{14}$$

The F term charges forming/checking point inputs, as in the retained implementation; model loading, compilation and runtime setup are also charged when incurred. Full ordered machine evaluations avoid assuming that an incrementally updated real leaf sum preserves binary64 accumulation. This bound covers the stated threshold-tree scorer, not arbitrary direct scalar dependence in postprocessing, and gives no linear-in-model-size claim.

Uniform parameters and finite-format limits. The operation bound is uniform over supplied rational real-cell partitions. It therefore permits unbounded profile counts; it does not require unboundedly many distinct score values. Growing specified input formats or finite monotone quantizers with rational inverse cells provide another such family, with inverse-cell construction separately charged. For fixed binary64, however, each nonempty machine cell contains a representable input, so P_i is capped by the fixed finite number of binary64 values in $[0, 1]$. The regime $P_i = \Theta(m)$ is then meaningful only within that finite cap, not as $m \to \infty$ at fixed format. An unbounded comparison uses the uniform real partitions or increasing input precision. Rational operand sizes, score comparison costs and format-dependent model evaluation remain additional bit costs. Equation (14) applies when the profiles are extracted under its stated threshold-tree assumptions; supplied profiles do not imply that a new format has been implemented or checked. Forming cuts and their inverse cells, and evaluating any growing format, must also be charged; arbitrary supplied partitions do not imply a T-node extraction procedure.

Signedness does not by itself collapse the extrema-role count. Fix distinct p, q and take

$$\mathcal{P} = \{r : \max r - \min r \le 2,\ r_p - r_q \le -1\}.$$

Its exact closure has $B_{uu} = 0$, $B_{pq} = -1$, and $B_{pu} = B_{uq} = 1$ for $u \notin \{p, q\}$; every other off-diagonal entry equals two. The unit entries follow through the negative edge, and every two-edge path is at least its displayed direct bound, proving closure. Only q has a negative column minimum and only p a negative row minimum. Thus the eligible pairs are $a \neq q$, $h \neq p$, $a \neq h$, numbering

$$N = (m-1)^2 - (m-2) = m^2 - 3m + 3.$$

Every pair occurs as unique minimum and maximum: set $r_a = 0$, $r_h = 2$, set $r_p = 1/2$ unless $a = p$ and $r_q = 3/2$ unless $h = q$, and set every remaining coordinate to one. The range is two and $r_q - r_p \ge 1$. This proves quadratic actual roles, not quadratic intrinsic difficulty: role keys may coincide, and for $m \ge 4$ the normalized (p, q) image is just $y - x \ge 1/2$ in the unit square. Necessity follows from the gap and range bounds; two free anchors at zero and two prove sufficiency. The example supplies no verifier lower bound.

With eligible-list size N and $\ell_N = 1 + \log_2(N + 1)$, polygons cost $O(NC_{ij})$ and a generic role index costs $O(N\ell_N^3 + P_iP_j\ell_N^4)$. The Cartesian construction below instead costs $O((m^2 + P_iP_j)\ell^2)$, including its dense-product preparation, where $\ell = 1 + \log_2(m + 1)$. Section B.10 further reduces this to $O(m^2 + P_iP_j)$ by selecting an active scale constraint and batching the queries. For $N = \Theta(m^2)$ and balanced $P_i = P_j = P$, the latter expression is smaller than the polygon bound when $1 \ll P \ll m^2$. Small role sets, very small profiles or sufficiently large profiles can favor other branches. In the uniform family $N = \Theta(m^2)$, $P_i = \Theta(m)$ and fixed k, the displayed verification bounds are $O(m^4)$ and $O(m^3)$, with every per-pair index build included. Under $T = O(m)$ and $F + K(d + 1) + H = O(m)$, ordinary closure and the stated profile construction each cost $O(m^3)$ operations. These compare specified procedure upper bounds; deduplication or other structure can help individual inputs, and no measured acceleration follows.

B.8 Implementation evidence and its limits

The separately retained sweep and linear prototypes use exact rational geometry and explicit input/work caps. Invalid inputs, arithmetic errors or caps produce failure with no invariant decision. Their first sealed generated checks each agree with the older independent range-variable enumerator on all 246 fixed whole decisions and independently validate 81 full adverse witnesses. The linear check additionally covers 113,672 static range queries and 128 batching cases, with direct extrema and interval intersection oracles. Shared low-level geometry tests are labeled as reused; they are not a second independently implemented geometry engine.

The signed-role-index prototype agrees with the older raw-range affine oracle on 182 whole decisions over the 32 signed fixtures from that same 42-fixture collection, with 57 independently replayed full adverse witnesses. The 10 nonnegative fixtures are explicitly outside this prototype's scope. Its first sealed check also tests 3,321 cell queries, reconstructs 939 full cell witnesses (112 below an excluded scale supremum), and compares 40,176 weighted-index queries over 36 point populations against direct scans. Twelve invalid-input/cap rejections and two valid-but-wrong proposed sets are retained. The cell oracle enumerates raw-range affine constraints and roles; it does not use the new common-scale reduction or range index. Input closure, profile and rounding helpers are reused and

disclosed. The prototype does not implement automatic choice between index and sweep; that hybrid is an algorithmic bound, not a new measured execution.

The Cartesian replacement retains this signed geometry and witness recovery. Its separate primitive check covers 13,056 queries over 36 populations, including all 16 strict/weak boundary masks, against direct weighted-product enumeration. Its integration check repeats the 182 whole decisions and 3,321 cell queries above, with the same 57 whole and 939 cell witnesses; these are reused cases, not an additional independent population. A separate saved-result audit replays all 998 witnesses against original constraints without importing the implementation or reference oracle. The Cartesian prototype rejects nonnegative zones. Its saved checks do not test the stronger batching algorithm; that implementation has a separate fixed check below.

The signed six-table implementation now covers the active-constraint algorithm in Section B.10, with exact owned-cell kernels and the same once-rounded binary64 whole-profile contract. Its fixed run repeats the 182 earlier signed decisions in both storage orders (364 calls), retaining 114 full adverse witnesses, and checks two wrong proposed sets with two additional witnesses. The 3,078 distinct-candidate cell queries retain the old population except the 243 same-candidate diagnostics excluded by the new cell API. The principal reference remains the older raw-range oracle; its answers and role caches are not passed to the implementation.

The same run checks 192 suffix-table builds and 62,208 queries against direct Cartesian enumeration, plus 1,452 suffix-position checks. Twenty-three analytically specified owned cells are checked in both storage orders. Another 128 trace configurations emit all 2,592 cell-feasibility labels using the actual moving cuts, including nonadverse cells. Direct rank checks cover 6,440 denominator-valid and 1,336 invalid type-cell visits; both scan orientations agree. These tests exercise strict active selection, weak-zero exceptions, original-coefficient recovery and movement through cells assigned to other types or with empty common scale. Finally, 44 fixed checks cover 19 invalid/capped rejections, six empty cells and 19 callback interruptions; interrupted tables release their storage. All five stages complete under external CPU, memory and output limits. Resource records charge setup, imports, reference checks and terminal hashing; they are cap ledgers rather than comparative timings. The nonnegative order sweep and automatic hybrid remain analytical and unimplemented. The endpoint-input interval specialization has the separate implementation and fixed check described in Appendix B.13.

The reproducibility package includes the protocols, implementation snapshots, reference oracles, inputs and complete check results. The signed whole-image implementation has the separate checks in Appendix B.11. These finite software tests supplement the mathematical proofs; they are separate from application outcomes and comparative runtime measurements.

B.9 A COMMON-SCALE INDEX FOR ALL EXTREMA ROLES

This subsection supplies the signed-zone reduction underlying Theorem 3. Assume $g = \min B < 0$; constants are then infeasible. The same row and column minima b, c are shared across candidate pairs. The reduction uses classical box feasibility and weighted range searching; its specific step is separating seven common scale constraints from five monotone role constraints.

Classical box feasibility, including open endpoints. For a closed finite difference matrix D, intervals with endpoints L_u, U_u are jointly feasible exactly when

$$L_u - U_v \le D_{uv} \qquad \text{for every ordered pair } u, v, \tag{15}$$

strictly when the participating lower or upper endpoint is excluded. For closed endpoints this is a specialization of tropical linear feasibility (Krivulin, 2014, Lemma 3): use the max-plus matrix $-D^\top$, lower vector L and upper vector U. A direct construction is

$$q_u = \min_v (D_{uv} + U_v). \tag{16}$$

The box inequalities give $q \ge L$; diagonal entries give $q \le U$; triangle closure gives $q_u - q_w \le D_{uw}$. Necessity is immediate. For open endpoints, each affected inequality has positive rational slack. Choose a common rational $\epsilon > 0$ smaller than every such slack divided by its number of moved endpoints. Raise excluded lower endpoints and lower excluded upper endpoints by ϵ. The resulting closed box is feasible and lies inside the original box. The diagonal inequalities reject every open zero-width interval. This is an elementary finite-slack extension of the classical criterion. Exact strict constraints and affine parameter bounds also precede this work (Hune et al., 2001).

Exactly seven common constraints. Let I_i, I_j be nonempty rational real cells, with the endpoint values and exclusion flags of Section 3. At a fixed $\lambda > 0$, initially put all coordinates in $[0, 1]$, replacing only the two candidate intervals. Positive scaling preserves closure, so (15) applies to $D = \lambda B$. Its full set of constraints is equivalent to (5), with strictness respectively

$$0, \quad s_{li},\ s_{ui},\ s_{lj},\ s_{uj}, \quad s_{li} \vee s_{uj},\ s_{lj} \vee s_{ui}.$$

For necessity, all lower endpoints are at least zero and upper endpoints at most one. The full constraints therefore imply the global, row and column inequalities even when a minimizing index is i or j; an excluded candidate endpoint preserves strictness. For sufficiency, classify an ordered pair by whether neither, one or both of its endpoints are candidate bounds. These cases use the global inequality, a row/column inequality, or one of the two cross inequalities. A repeated candidate uses the already checked nonemptiness of its interval. No index is removed from b, c or g.

Intersect these seven half-lines with the strict constraint $\lambda > 0$, checking zero coefficients directly and retaining endpoint ownership. If the interval Λ is nonempty, then

$$0 < L = \sup \Lambda \leq -1/g < \infty.$$

Let η mean that $L \in \Lambda$. The intersection takes constant exact work. This is the point at which the signed assumption is used; nonnegative zones need not have a finite upper scale and admit the order-domain specialization in Section B.12.

Role constraints and aliases. If a is a minimum, then $0 \leq r_u - r_a \leq B_{ua}$ for every u, so $c_a = 0$. Conversely the closed column $(B_{ua})_u$ attains a minimum at a. Similarly, h can be a maximum exactly when $b_h = 0$, as witnessed by $(-B_{hu})_u$. These are individual eligibility conditions, not a claim that every eligible pair is jointly attainable.

Now impose $q_a = 0, q_h = 1$. Beyond the common box constraints, the new inequalities are exactly

$$\begin{aligned} l_i &\leq \lambda B_{ia}, & l_j &\leq \lambda B_{ja}, \\ 1 - u_i &\leq \lambda B_{hi}, & 1 - u_j &\leq \lambda B_{hj}, \\ 1 &\leq \lambda B_{ha}, \end{aligned} \tag{17}$$

with flags $(s_{li}, s_{lj}, s_{ui}, s_{uj}, 0)$. All remaining new comparisons are implied by the common constraints or by $c_a = b_h = 0$. The five coefficients are nonnegative. The last inequality excludes $a = h$ and every $B_{ha} = 0$. If $a = i$, the first coefficient is zero, forcing $l_i = 0$ with its endpoint included; if $h = i$, the third forces $u_i = 1$ included. The same applies to j. Thus candidate/role aliases enforce exactly the intersection with an anchor singleton. Equivalently one can retain repeated box labels: the zero diagonal forces their coordinates to agree. No alias is dropped.

Lemma 8 (Testing roles at the common scale supremum)**.** *For a nonempty common interval Λ, set*

$$D = (B_{ia}, B_{ja}, B_{hi}, B_{hj}, B_{ha}), \quad v = (l_i, l_j, 1 - u_i, 1 - u_j, 1), \quad s = (s_{li}, s_{lj}, s_{ui}, s_{uj}, 0).$$

A role and cell pair are jointly feasible exactly when

$$D_t \geq v_t/L \quad (t = 1, \ldots, 5), \qquad \textit{strict iff } s_t \vee (\neg\eta \wedge v_t > 0). \tag{18}$$

Write $s'_t = s_t \vee (\neg\eta \wedge v_t > 0)$ for these strictness flags. An excluded supremum leaves a weak zero threshold weak. This preserves feasible aliases and zero coefficients.

Proof. Every v_t, D_t is nonnegative, making the five role inequalities monotone in positive scale. If η holds, their feasibility is equivalent to satisfying them at L. If η fails, any positive numerator must satisfy $LD_t > v_t$ to be attained below L. For a zero numerator, an originally strict inequality requires $D_t > 0$, whereas a weak one allows $D_t = 0$ at every positive scale. These observations prove necessity and the zero-threshold exception.

For sufficiency when η fails, let H be the maximum of zero, the lower endpoint of Λ, and every v_t/D_t for which $D_t > 0$. Nonemptiness of Λ and the strict positive-threshold conditions ensure $H < L$. A zero coefficient has $v_t = 0, s_t = 0$. Choose $\lambda = (H + L)/2$. It belongs to Λ, including when its lower endpoint is open, and satisfies all five comparisons with their original ownership. When η holds, use $\lambda = L$. Construct the box with its compatible anchor singletons and apply (16), with inward shifts for open bounds. The resulting rational q has minimum zero and maximum one, and $r = q/\lambda$ belongs to $\mathcal{P}(B)$ with exact normalization q. □

Four-dimensional aggregation with attaining roles. For fixed i, j, store each eligible role with key $(B_{ia}, B_{ja}, B_{hi}, B_{hj})$, weight B_{ha} and role pointer (a, h). There are $N \le m(m-1)$ keys; an empty set is handled directly. Query the upper orthant with thresholds $(v_1, v_2, v_3, v_4)/L$ and flags (s'_1, s'_2, s'_3, s'_4). Among matching keys return the maximum weight, with a fixed role tie order. A role is feasible exactly when this weight is at least $1/L$, strictly if $\neg\eta$. Thus a four-dimensional maximum query replaces enumeration over all roles for that cell pair.

This is the classical static range-tree aggregation bound described by Lueker (1978, Sections 1–2), crediting the earlier static and decomposable-search constructions. Maximum weight with an attaining pointer is associative and commutative; no reporting/output-size term is needed. Presort each coordinate once and partition those sorted lists in linear work at recursive splits. A one-dimensional suffix maximum is the base case. The four-dimensional construction takes $O(N \log^3(N+1))$ operations and index words; an orthant query takes $O(\log^4(N+1))$. Duplicate coordinates use stable IDs for construction and inclusive/strict boundary cuts before or after the entire equality group. No perturbation or floating comparison is used. Keys and weights refer to exact rationals; their bit costs are additional. The structure is classical, and no best-possible orthogonal query bound is asserted.

Cartesian aggregation. The first two key coordinates depend only on a and the last two only on h. Put $A_0 = \{a : c_a = 0\}$, $H_0 = \{h : b_h = 0\}$ and $r = |A_0|$, $s = |H_0|$. Retain the full product $A_0 \times H_0$ with weight B_{ha}. These weights are nonnegative. Zero weights, including $a = h$, never pass the positive threshold $1/L$, so this padding preserves the query. Empty role sides immediately give no hit.

Build a balanced tree on A_0 sorted by B_{ia} and another on H_0 sorted by B_{hi}. Each minimum-tree node U also stores its members in B_{ja} order, and each maximum-tree node V stores its members in B_{hj} order. Stable IDs resolve ties without changing threshold equality. For these two lists $(a_1, \dots, a_u)$ and $(h_1, \dots, h_v)$, store

$$F_{U,V}[p, q] = \max_{p \le x \le u,\ q \le y \le v} B_{h_y a_x},$$

with an attaining role under a fixed role tie order. Empty suffixes have a sentinel below every weight. The recurrence taking the maximum of $B_{h_q a_p}$, $F[p+1, q]$ and $F[p, q+1]$ builds one table in $O(uv)$ work and words; its boundary row and column have the same bound.

Write $\ell_r = 1 + \lceil \log_2(r+1) \rceil$ and define ℓ_s similarly. Each minimum leaf belongs to $O(\ell_r)$ nodes and each maximum leaf to $O(\ell_s)$ nodes. Therefore

$$\sum_{U,V} |U||V| = \left(\sum_U |U| \right) \left(\sum_V |V| \right) = O(rs\ell_r\ell_s).$$

Sorting the first coordinates and merging the second-coordinate lists costs $O(r\ell_r + s\ell_s)$, absorbed by this bound when $r, s \ge 1$. All repeated role weights and node-pair tables are charged.

A first-coordinate threshold gives a suffix decomposed into $O(\ell_r)$ disjoint minimum nodes and $O(\ell_s)$ disjoint maximum nodes. In each selected node, binary-search its second-coordinate threshold once, cutting before or after the full equality group according to strictness. Reuse these positions over every selected node pair and aggregate their table entries. The resulting disjoint products cover exactly the filtered role product, so its maximum and attaining role are exact. The searches and lookups cost

$$O(\ell_r^2 + \ell_s^2 + \ell_r\ell_s) = O(\ell^2).$$

This ragged-size bound includes binary searches; it does not assume fractional cascading. The construction uses classical balanced decompositions and suffix maxima, not a new general range-search data structure.

Completeness, charged construction and hybrid bound. Every adverse pair of profile scores is tested on its exact real cells. The preceding equivalence accepts precisely those pairs with some compatible extrema and full potential. Thus it can replace the polygon search for a signed zone in the earlier completeness proof, including score equality resolved by priority. Computing the witness's finite inward slack, extending all coordinates, and checking every original constraint costs $O(m^2)$ exact operations once, followed by normalization and profile evaluation; the binary64 specialization additionally checks once-only rounding and full machine replay. A returned potential concerns the abstract zone, not another graph.

For each compared pair, Cartesian construction and all cell-pair queries cost

$$O\big(rs\ell_r\ell_s + P_iP_j(\ell_r^2 + \ell_s^2 + \ell_r\ell_s)\big) \subseteq O((m^2 + P_iP_j)\ell^2).$$

Keys depend on i, j, so every build is paid. Streaming pairs reuses $O(m^2\ell^2)$ index words. The complementary polygon bound is $O(NC_{ij})$, with a shared eligible list. A sparse generic index remains an alternative with its separately stated work and storage; it need not beat the dense Cartesian method on every input. The checked Cartesian prototype scans the two eligible sides and builds their dense weight product per candidate pair, without implementing automatic branch choice. The next subsection sharpens the signed bound used in (4). No optimality or measured speedup follows from these procedure bounds.

B.10 An active scale constraint gives six two-dimensional tables

The four key coordinates in the preceding construction are not independent for a valid cell query: at least one common upper-scale constraint attains L. Exploiting this dependence gives the stronger signed branch in Theorem 3. Put $\tau = 1/L$ and retain $\mathcal{A}_0 = \{a : c_a = 0\}$, $\mathcal{H}_0 = \{h : b_h = 0\}$, endpoint flags and η from Section B.9.

Choose an active constraint with the correct ownership. Every finite positive upper bound on Λ comes from a negative coefficient in (5). If L is owned, choose any attaining upper constraint. If it is unowned, choose a *strict* attaining upper constraint, with a fixed deterministic tie order. Such a constraint exists: otherwise all upper constraints hold at L, and nonemptiness supplies all lower constraints there, contradicting nonownership. Choosing an arbitrary weak active constraint in this case would not justify the reductions below.

The global inequality is weak. When selected it has $L = -1/g$ and is owned. Choose a fixed pair (p, q) attaining $B_{pq} = g$. Any feasible common box vector satisfies $z_p - z_q \le -1$ while $0 \le z \le 1$, so $z_p = 0, z_q = 1$. The common box already has both extrema; its feasibility and rational extension suffice without a role query. Otherwise one of the six candidate row, column or cross constraints is selected.

Six fixed tables. For each applicable row below, $d > 0$, the two keys and the stored weight depend only on B, i, j. Only τ and the query thresholds depend on cells. Rows whose stated coefficient is nonnegative are omitted.

Active constraint	d, τ	Keys $(X(a), Y(h))$	Weight $W(h, a)$
R_i	$-b_i,\ d/(1-l_i)$	(B_{ja}, B_{hj})	$\min\{B_{ha}, B_{ia} + d\}$
C_i	$-c_i,\ d/u_i$	(B_{ja}, B_{hj})	$\min\{B_{ha}, B_{hi} + d\}$
R_j	$-b_j,\ d/(1-l_j)$	(B_{ia}, B_{hi})	$\min\{B_{ha}, B_{ja} + d\}$
C_j	$-c_j,\ d/u_j$	(B_{ia}, B_{hi})	$\min\{B_{ha}, B_{hj} + d\}$
X_{ij}	$-B_{ij},\ d/(u_j - l_i)$	(B_{ia}, B_{hj})	B_{ha}
X_{ji}	$-B_{ji},\ d/(u_i - l_j)$	(B_{ja}, B_{hi})	B_{ha}

Here R_i, C_i, X_{ij} denote the common row-i, column-i and cross-ij inequalities, respectively. For a minimum-side key belonging to candidate c, query $X(a) \ge l_c\tau$ with its adjusted lower flag from (18). For a maximum-side key belonging to e, query $Y(h) \ge (1 - u_e)\tau$ with its adjusted upper flag. Test the maximum stored weight against τ, strictly exactly when $\neg\eta$. A selected constraint's denominator is positive; otherwise it could not give the finite positive supremum of a nonempty Λ.

Row and column reductions, including strict boundaries. For R_i, choose t with $B_{it} = b_i = -d$. Closure and maximum eligibility give $0 \le B_{ht} \le B_{hi} - d$, hence $B_{hi} \ge d = (1 - l_i)\tau \ge (1 - u_i)\tau$. The maximum-side i condition is redundant, with strict slack if $u_i > l_i$. If $u_i = l_i$, the nonempty singleton owns both endpoints. A selected unowned R_i would require $s_{li} = 1$, so this singleton case necessarily has $\eta = 1$ and only a weak redundant comparison. The minimum-side threshold is $l_i\tau = \tau - d$. Its adjusted flag equals $\neg\eta$: an owned active row must have $s_{li} = 0$, while the unowned selection rule requires $s_{li} = 1$. Thus its conjunction with the original weight comparison is exactly $\min\{B_{ha}, B_{ia} + d\} \ge \tau$, with the common strictness flag. This remains exact when $l_i = 0$; the original strict-zero exception is retained rather than silently replacing $>$ by $\ge$.

For C_i, choose t with $B_{ti} = c_i = -d$. Minimum eligibility and closure give $0 \le B_{ta} \le B_{ia} - d$, so $B_{ia} \ge d = u_i\tau \ge l_i\tau$. The minimum-side i comparison is redundant, with the same positive-width or owned-singleton reasoning. The maximum-side threshold becomes $(1 - u_i)\tau = \tau - d$; its flag equals $\neg\eta$ by the active column's ownership. Combining it with the weight yields $\min\{B_{ha}, B_{hi} + d\} \ge \tau$ with that flag. Interchanging i, j proves R_j, C_j.

Cross reductions. For X_{ij}, $B_{ij} = -d$ and $d = (u_j - l_i)\tau$. Closure gives $B_{ja} \ge B_{ia} + d$ and $B_{hi} \ge B_{hj} + d$. The surviving minimum-i comparison therefore implies $B_{ja} \ge u_j\tau \ge l_j\tau$, and the surviving maximum-j comparison implies $B_{hi} \ge (1 - l_i)\tau \ge (1 - u_i)\tau$. Strictness is preserved. If $l_j < u_j$, the first implication has strict slack. If $l_j = u_j$, J is a closed singleton; an unowned selected X_{ij} must then have $s_{li} = 1$, making the surviving minimum-i comparison strict. The owned case requires only a weak discarded comparison. The second implication is symmetric: when I is a singleton and L is unowned, selection requires $s_{uj} = 1$. This treats zero thresholds and candidate/role aliases as well. Interchanging i, j proves X_{ji}. Thus each table query is equivalent to all five original role conditions.

Construction and witness recovery. Sort the eligible minima by X and maxima by Y, using stable identifiers within equal keys. For these orders $a_1, \ldots, a_r$ and $h_1, \ldots, h_s$, store a suffix maximum with attaining pointer:

$$F[p, q] = \max_{x \ge p,\, y \ge q} W(h_y, a_x).$$

The recurrence combining the current weight, $F[p + 1, q]$ and $F[p, q + 1]$ fills the table in $O(rs)$ work and words. Empty sides return no hit before allocation; for nonempty sides boundary rows have the same bound. Sorting costs $O(m \log m) \subseteq O(m^2)$, and the constant number of tables costs $O(m^2)$. Transformed weights take constant arithmetic work. Two binary searches plus one lookup give an online $O(\log m)$ query; the batching below removes these searches from full profile processing. Zero-weight and diagonal pairs remain harmless because $\tau > 0$. An attaining role passes the original exact test, so the preceding rational scale selection and box extension recover its full witness unchanged.

Batch all cells in separate table passes. Let p, q index the ordered cells of candidates i, j. Their lower and upper endpoints are nondecreasing, including singleton cells. Each of a table's two numerical query thresholds is monotone along either index on its positive-denominator domain. The direction pairs below correspond to increasing (p, q); arrows permit equality.

Table	Minimum-key threshold	Maximum-key threshold
R_i	$(\uparrow, \uparrow)$	$(\uparrow, \downarrow)$
C_i	$(\downarrow, \uparrow)$	$(\downarrow, \downarrow)$
R_j	$(\uparrow, \uparrow)$	$(\downarrow, \uparrow)$
C_j	$(\uparrow, \downarrow)$	$(\downarrow, \downarrow)$
X_{ij}	$(\uparrow, \downarrow)$	$(\uparrow, \downarrow)$
X_{ji}	$(\downarrow, \uparrow)$	$(\downarrow, \uparrow)$

For example, the X_{ij} thresholds are $dl_i/(u_j - l_i)$ and $d(1 - u_j)/(u_j - l_i)$: both increase with l_i and decrease with u_j where $l_i < u_j$. The other signs follow from the table formulas. A zero numerator is constant. Along each grid line, valid denominators form an interval, so invalid portions are skipped before division.

Process each table type in its own pass, using the shorter profile as the outer loop. Along every inner line, move two pairs of rank pointers through the sorted keys, retaining both the first weak and first strict cut positions. Each numerical threshold moves in its fixed direction. Ownership flags and the assigned active constraint may change arbitrarily without reversing these scans: ranks advance on the whole valid line, even at cells whose scores are not adverse or whose selected constraint belongs to another pass. Recompute the constant-size common interval and deterministic assignment at each visited cell; only a cell assigned to this pass may return its hit. A separate constant-work-per-cell pass handles the owned global constraint. Each inner line costs $O(m + \max\{P_i, P_j\})$. Including all constant-many passes and construction, the work is

$$O\big(m^2 + m \min\{P_i, P_j\} + P_iP_j\big) = O(m^2 + P_iP_j).$$

If the shorter profile has at most m cells, the middle term is at most m^2; otherwise it is at most P_iP_j. Only the current table and a constant number of pointers need remain live; no P_iP_j query array is stored. Streaming candidate pairs gives $O(m^2 + \sum_i P_i)$ working storage. Combining this signed branch with the complementary polygon bound, the whole-image sweep and the nonnegative branch below proves (3)–(4). Every per-pair build, comparison sort and grid pass is charged. This is an arithmetic/word bound; rational bit costs, profile construction and closure remain separate. It establishes neither an algorithmic lower bound nor a measured advantage of the implemented signed tables.

B.11 Exact signed pair images and profile sweeps

The active-scale reduction also permits constructing the entire pair image before inspecting profile cells. Its structure gives an additional procedure whose cost is nearly linear in the supplied profile sizes. The ingredients of closed-zone box extension, reciprocal-range substitution and static range extrema are classical (Miné, 2001; Krivulin, 2014; Charnes & Cooper, 1962; Bender & Farach-Colton, 2000). Difference domains also have classical tropical/Kleene-star geometry (Develin & Sturmfels, 2004; Joswig & Kulas, 2010; Johnson & Kambites, 2015). Our object is the two-coordinate image after division by a common, data-dependent range, rather than a new representation of the raw domain. The point established here is that the factored active-scale role caps give boundary complexity linear in the candidate count and a suitable direction for an owned-cell sweep. The implementation and its finite correctness checks are described below.

Fix distinct candidates i, j and a finite, consistent, closed rational difference matrix B with $g = \min B < 0$. Retain $b_u = \min_v B_{uv}$, $c_v = \min_u B_{uv}$ and put $R = \max B > 0$. Let $\mathcal{E}$ contain the distinct pairs (a, h) with $c_a = b_h = 0$ and $B_{ha} > 0$, and write $N = |\mathcal{E}| \le m(m-1)$. One shared $O(m^2)$ scan computes these summaries and enumerates $\mathcal{E}$ after closure. Every candidate pair thereafter reads this stored role list. Define the exact pair image

$$\mathcal{Q}_{ij} = \{(\mathcal{N}(r)_i, \mathcal{N}(r)_j) : r \in \mathcal{P}(B)\}.$$

Signedness excludes constants, so normalization is defined by a positive range throughout this section.

Proposition 9 (Linear-complexity signed pair image)**.** *After the shared preparation above, $\mathcal{Q}_{ij}$ has an exact rational straight-line representation with $O(m)$ vertices and edges, including all lower-dimensional parts. It can be constructed in $O(m^2)$ arithmetic/comparison operations and $O(N+m)$ working words. For verification it suffices to retain at most seven pieces with this total complexity. Each piece has interval fibers in at least one coordinate direction. This directional assertion is per piece; a common direction for the whole image is not required. The whole image need not be convex.*

Closed point patches. First fix exact point coordinates $q_i = x, q_j = y$. Substitute $l_i = u_i = x$ and $l_j = u_j = y$ into (5); every comparison is now weak. Each negative coefficient supplies an affine upper bound on the reciprocal range:

$$\begin{aligned} L_G &= -1/g, \\ L_{R_i} &= (1-x)/(-b_i), \qquad L_{C_i} = x/(-c_i), \\ L_{R_j} &= (1-y)/(-b_j), \qquad L_{C_j} = y/(-c_j), \\ L_{X_{ij}} &= (y-x)/(-B_{ij}), \quad L_{X_{ji}} = (x-y)/(-B_{ji}). \end{aligned} \tag{19}$$

A nonglobal type is present only when its denominator is positive. Zero coefficients in the common inequalities are checked directly. Every feasible raw range is at most R, so its reciprocal satisfies $\lambda \ge 1/R$. This bound also follows from any complete role: $1 \le \lambda B_{ha} \le \lambda R$. For each present type Q, define

$$\mathcal{C}_Q = \left\{(x,y) \in [0,1]^2 : L_Q \ge 1/R, \quad \text{the point common inequalities hold at } \lambda = L_Q\right\}. \tag{20}$$

Substitution makes these constant-many constraints linear. Thus $\mathcal{C}_Q$ is a compact convex rational polygon, possibly empty, a segment or a point. Its inequalities imply $L_Q \le L_{Q'}$ for every other present upper bound. We allow overlapping active boundaries and require no tie partition.

For any attainable point the common feasible scale interval has a positive attained supremum $L \le -1/g$. An actual extrema role remains feasible at L: all five coefficients and numerators in (17) are nonnegative. Some type in (19) attains L, placing the point in its patch. Conversely, a complete

role at L_Q together with the common constraints gives a full potential by closed-box extension. Hence constructing the role union inside every patch is exact. Profile endpoints have not been closed by this argument: they are imposed later on this point image.

The global patch requires no role union. Fix (p, q) with $B_{pq} = g$. At $\lambda = -1/g$, every common-box solution has $q_p - q_q \le -1$, hence $q_p = 0, q_q = 1$. Therefore $\mathcal{C}_G \subseteq \mathcal{Q}_{ij}$.

Three projective role forms. Write $(r, s) = (i, j)$ or (j, i) and $(z, w) = (q_r, q_s)$. For each nonglobal type the following transformation reduces a role to

$$D_{ha} = \{(u, v) \in \mathbb{R}^2_{\ge 0} : u \le A_a,\ v \le C_h,\ u + v \le V_{ha}\}. \tag{21}$$

All coefficients are rational and take constant work per stored role.

For R_r, put $d = -b_r > 0$ and $L = (1 - z)/d$. Closure and maximum eligibility give $B_{hr} \ge d$, so the maximum-side r constraint is redundant. The minimum-side constraint combines with the range constraint as $\min\{B_{ha}, B_{ra} + d\} \ge 1/L$. Consequently use

$$\begin{aligned} (u, v) &= \left(\frac{w}{1 - z}, \frac{1 - w}{1 - z}\right), \\ (A_a, C_h, V_{ha}) &= \frac{1}{d}\left(B_{sa}, B_{hs}, \min\{B_{ha}, B_{ra} + d\}\right). \end{aligned} \tag{22}$$

The inverse is $z = 1 - 1/(u + v)$, $w = u/(u + v)$. For C_r, put $d = -c_r > 0$, $L = z/d$. Closure and minimum eligibility give $B_{ra} \ge d$, making the minimum-side r constraint redundant. The corresponding formulas are

$$\begin{aligned} (u, v) &= \left(\frac{w}{z}, \frac{1 - w}{z}\right), \\ (A_a, C_h, V_{ha}) &= \frac{1}{d}\left(B_{sa}, B_{hs}, \min\{B_{ha}, B_{hr} + d\}\right), \end{aligned} \tag{23}$$

with inverse $z = 1/(u + v)$, $w = u/(u + v)$.

For X_{rs}, put $d = -B_{rs} > 0$, $L = (w - z)/d$. Closure gives $B_{sa} \ge B_{ra} + d$ and $B_{hr} \ge B_{hs} + d$, so the surviving minimum-r and maximum-s conditions imply the other two. Set

$$\begin{aligned} (u, v) &= \left(\frac{z}{w - z}, \frac{1 - w}{w - z}\right), \\ (A_a, C_h, V_{ha}) &= \left(\frac{B_{ra}}{d}, \frac{B_{hs}}{d}, \frac{B_{ha}}{d} - 1\right). \end{aligned} \tag{24}$$

Here $u + v + 1 = 1/(w - z)$, proving the last inequality of (21). Its inverse is $z = u/(1 + u + v)$, $w = (1 + u)/(1 + u + v)$. A negative V_{ha} gives an empty role, whereas $V_{ha} = 0$ may contribute the origin and is retained. Candidate/anchor aliases are also retained: their zero keys enforce the original zero or one coordinate.

On $\mathcal{C}_Q$, the respective denominators satisfy $1 - z \ge d/R$, $z \ge d/R$, or $w - z \ge d/R$. Thus the projective maps are regular on each compact patch, preserve line segments and convexity there, and send the patch to a constant-complexity convex polygon $\mathcal{C}'_Q$. For row and column types this polygon has $u + v \ge 1$, excluding the inverse map's pole. The exact active piece is the inverse image of

$$\mathcal{C}'_Q \cap D_Q, \qquad D_Q = \bigcup_{(a,h)\in\mathcal{E}} D_{ha}.$$

Factored caps give a candidate-linear frontier. Every nonempty D_{ha} is downward closed and contains the origin. Its upper frontier is

$$f_{ha}(u) = \min\{C_h, V_{ha} - u\}, \qquad 0 \le u \le E_{ha} := \min\{A_a, V_{ha}\}. \tag{25}$$

If no nonempty role remains, return $D_Q = \varnothing$ before taking any maximum; this differs from a nonempty union with zero-width support. Therefore D_Q lies between $v = 0$ and the nonincreasing upper frontier $H(u) = \max f_{ha}(u)$ on $[0, U]$, where $U = \max_{ha} E_{ha}$. Its affine pieces have slope

zero or -1. Although there are N roles, (22)–(24) give only m possible A_a caps and m possible C_h heights.

Every positive-length horizontal frontier piece has height C_h. Because H is nonincreasing, a fixed height can occur on only one connected plateau, after merging adjacent collinear pieces. There are thus at most m horizontal pieces. A positive downward jump can occur only at an A_a: if a disappearing role instead ends at $u = V_{ha}$, its endpoint height is zero and cannot lower a positive maximum. This also bounds a positive terminal vertical edge. Hence there are at most m positive jump locations. Outside the horizontal pieces and jumps, consecutive slope--1 pieces must coincide if they meet continuously; otherwise they do not meet. There are consequently at most $2m + 1$ maximal diagonal pieces. Including jump fibers, the ground segment and left axis gives $O(m)$ vertices and edges.

This argument uses the actual factored caps, without assuming arbitrary role polygons can be realized by a closed zone. Negative V_{ha} roles are empty. Zero caps, $V_{ha} = 0$ and anchor aliases remain valid: a role with $A_a = 0$ may contribute only to the initial closed fiber. If $U = 0$, the whole union is a vertical segment or point; zero-height unions lie on the ground axis. These parts obey the same bound.

Grouped construction in quadratic arithmetic work. Bucket nonempty roles by minimum identity a while reading the stored role list, then process groups in decreasing distinct A_a order. Sort the maximum identities by C_h once. Maintain

$$M_h = \max\{V_{ha} : \text{role } (a, h) \text{ has been added}\},$$

with an absent value for keys without an added role. Add the entire current A group before emitting its strip down to the next smaller A value. There are $O(N)$ updates overall. In the open strip the active roles are fixed, and its frontier is exactly

$$\max_h \min\{C_h, M_h - u\}, \qquad u \le \max_h M_h,$$

where absent keys are omitted. The last inequality is essential for nonnegative-height support. If all keys are absent, emit no strip.

One $O(m)$ scan constructs this envelope. For tied C retain the largest M; in decreasing C order discard any key whose M is no larger than a preceding M. It is dominated in both caps. The survivors satisfy $C_1 > \cdots > C_s$ and $M_1 < \cdots < M_s$. Neighboring functions cross at $t_k = M_k - C_{k+1}$, with

$$M_k - C_k < t_k < M_{k+1} - C_{k+1}.$$

The crossings are strictly increasing, so the winners occur in survivor order, each supplying at most one plateau and one diagonal. Crop this envelope to the strip, $u \ge 0$ and $u \le \max_h M_h$.

Emit the cropped strips closed. At a strip's lower cut, roles having that smaller A value have not yet been added; the next strip adds them and supplies the full closed fiber there. Their union is exact, and no included role violates its A bound at a cut. End at zero, including an $[0, 0]$ fiber after adding an $A = 0$ group if present. Thus disappearing roles, support endpoints, axes and points are retained. Streaming in decreasing u, keep the actual vertical jump at a shared cut, remove seam coverage internal to the union, and merge consecutive collinear frontier segments. This takes constant work per emitted segment and leaves the $O(m)$ frontier just proved. Temporary pieces are discarded after each strip; they are not accumulated in quadratic storage.

There are at most m strips, each costing $O(m)$, and candidate-indexed buckets cost $O(N + m)$ space. Sorting the $O(m)$ cap keys costs $O(m \log m)$; sorting all N role triples is unnecessary. Hence the construction costs $O(N + m^2 + m \log m) = O(m^2)$ arithmetic/comparison work and $O(N + m)$ words. A grouped implementation and its finite checks are described below; measured costs do not prove these asymptotic bounds.

Clipping and the whole point image. Clipping to $\mathcal{C}'_Q$ preserves $O(m)$ complexity: each of its constant-many lines meets only $O(m)$ frontier/axis edges. Component bookkeeping, including collinear overlaps and isolated contacts, fits $O(m \log(m + 1))$ work. The regular inverse projective map preserves straight lines, incidences and lower-dimensional parts. Distinct affine upper bounds in (19) tie only on a straight line. Their active patches have disjoint two-dimensional interiors and meet

on a constant set of tie lines, which cut only $O(m)$ edges in total. The whole union therefore retains $O(m)$ complexity. Alternatively, retain the constant-many pieces and duplicate boundary coverage during verification. All these additional operations fit the $O(m^2)$ construction bound.

A valid sweep direction for every piece. For a row type, fix w. Increasing z increases both coordinates in (22); downward closure of D_Q makes the feasible z values a prefix. For a column type, increasing z decreases both coordinates in (23), giving a suffix. Intersecting either fiber with the convex patch preserves an interval. For a cross type, increasing w at fixed z, or decreasing z at fixed w, decreases both coordinates in (24). Both coordinate fibers are therefore intervals after clipping. The global piece is convex. This proves the directional assertion of Proposition 9, completing its construction. In particular, R_i, C_i are swept in y, with intervals in x; R_j, C_j use the reverse orientation. The proof uses this orientation; downward closure alone would not justify the other orientation.

Proposition 10 (Tight worst-case boundary complexity)**.** *For every $m \geq 4$, a full-dimensional finite, consistent, closed rational signed difference zone has a normalized pair image with $\Omega(m)$ maximal boundary edges. Hence Proposition 9's $O(m)$ boundary bound is tight in the worst case. This is an output-size statement, not a running-time lower bound.*

Proof. First let $m = 2n + 2$. Use anchors p, q and candidates a_t, h_t, $t = 1, \ldots, n$. Choose rational $d > 0$ and

$$0 < A_1 < \cdots < A_n < d, \qquad d > C_1 > \cdots > C_n > 0;$$

for example $A_t = dt/(n+1)$ and $C_t = d(n+1-t)/(n+1)$. In the translation gauge $r_q = 0$, impose the difference constraints

$$r_p = d, \qquad -A_t \leq r_{a_t} \leq d, \qquad 0 \leq r_{h_t} \leq d + C_t, \qquad r_{h_s} - r_{a_t} \leq d \quad (s \neq t).$$

Every displayed interval denotes differences from r_q, and the anchor equality includes both inequalities. Finite coordinate bounds give a finite difference matrix; its closure preserves the nonempty feasible set. For instance, all $r_a = 0, r_h = d$ are feasible. The closed matrix is signed because $B_{qp} = -d$.

Put $u = -\min_v r_v$ and $v = \max_v r_v - d$. The exact attainable set is

$$D = \bigcup_{t=1}^{n} [0, A_t] \times [0, C_t]. \tag{26}$$

Indeed, if $u, v > 0$, the minimum is some a_t and the maximum some h_s. The off-diagonal inequality would give $d + u + v \leq d$ unless $s = t$. Thus $u \leq A_t, v \leq C_t$. On either axis the coordinate attaining the nonzero amplitude supplies a containing rectangle; the origin is also included. Conversely, every point of rectangle t is attained by

$$r_{a_t} = -u, \quad r_{h_t} = d + v, \quad r_{a_s} = d,\ r_{h_s} = 0\ (s \neq t), \quad r_q = 0,\ r_p = d.$$

Off-diagonal $h - a$ differences are u, v, or $-d$, all at most d. The coordinate bounds and both attained extrema hold, including their endpoints. These full witnesses also satisfy the closure.

The incomparable rectangles have n horizontal frontier edges, n vertical jump/right edges, and two axis edges, all of positive length. The normalized anchor pair is the regular projective image

$$(z_p, z_q) = \left(\frac{d+u}{d+u+v}, \frac{u}{d+u+v} \right), \qquad u = \frac{dz_q}{z_p - z_q}, \quad v = \frac{d(1-z_p)}{z_p - z_q}.$$

Here $z_p - z_q > 1/3$, so the map and inverse are regular on the compact sets. They preserve boundary incidence and noncollinearity. The pair image therefore has exactly $2n + 2 = m$ maximal boundary edges.

To obtain full dimension, replace the anchor equality by $d \leq r_p - r_q \leq d + \delta$ for rational $\delta > 0$, leaving all other constraints unchanged. The fixed-gap slice remains included. Conversely, for any vector in the relaxed zone, gauge $r_q = 0$, put $t = r_p \geq d$ and multiply every coordinate by d/t. This contracts each coordinate interval into itself and preserves every off-diagonal bound $h - a \leq d$. It returns a vector in the fixed-gap slice with identical normalization. The pair image is therefore unchanged. The vector $r_p = d + \delta/2$, all $r_a = r_h = d/2$, strictly satisfies every nontrivial inequality,

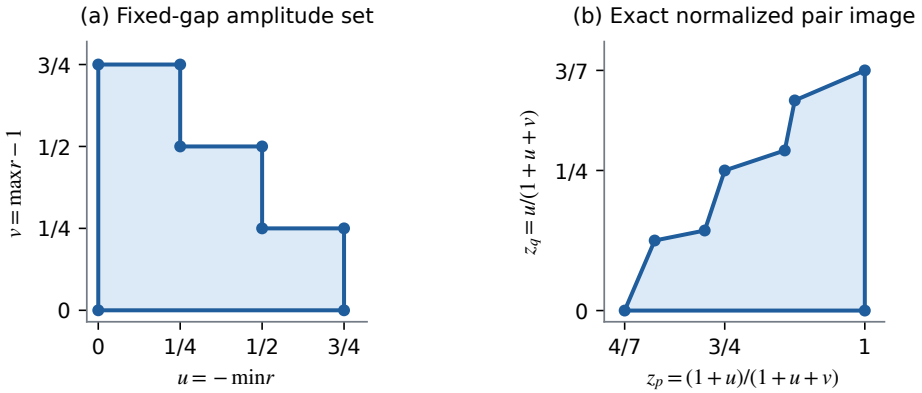


Figure 4: The tightness construction with $n = 3$, $m = 8$, $d = 1$, $A_t = t/4$ and $C_t = (4 - t)/4$. Left: the fixed-gap amplitude set is the union of three incomparable grounded rectangles. Right: its exact normalized anchor image has eight maximal boundary edges. Relaxing the anchor gap to $[d, d + \delta]$ gives a full-dimensional raw zone with this same normalized image, by the contraction argument in Proposition 10. The drawing illustrates the analytical construction; it is not sampled data.

proving full dimension with the free translation included. For odd m, add one candidate e satisfying $r_q \le r_e \le r_p$. It changes neither extremum, and $r_e = r_p/2$ leaves both new inequalities strict. Its pair image has $m - 1$ edges. This proves the assertion without requiring any fixed offset between distinct candidates. □

Proposition 11 (Owned-profile sweep over the pair image)**.** *With the shared zone preparation and linear static score-range preparation of the supplied profiles, all adverse cells for a fixed compared pair can be excluded, or one feasible adverse cell pair returned, in*

$$O\big(m^2 + (m + P_i + P_j + 1)\log(2 + m + P_i + P_j)\big) \tag{27}$$

arithmetic/comparison and word-RAM operations, including the pair-image construction. Working storage is $O(N + m + P_i + P_j)$ beyond the supplied matrix and shared prepared role list. A hit supplies a full rational counterexample after a once-only $O(N + m^2)$ role search, extension and original-constraint check, followed by full normalization and profile/scorer replay.

Proof. For one active piece, denote its justified scan coordinate by t and the other coordinate by s. Its boundary has $K = O(m)$ edges and every t fiber is an interval. Sort all distinct vertex t coordinates. At each critical coordinate retain the actual *closed* fiber, a segment, point or empty set. Between consecutive critical coordinates use an *open* t interval. If nonempty, its two fiber endpoints follow fixed affine boundary edges, giving one constant-complexity trapezoid. The two affine functions may coincide, yielding a line.

An event sweep builds these $O(K)$ strata in $O(K \log(K + 1))$ work. Insert and remove incident boundary edges at their vertex coordinates; edges crossing an open slab cannot cross there, since such a crossing would already be a boundary vertex. At an event combine continuing edges, incident endpoints and segments on the scan line before proceeding to the next open slab. This keeps closed bridges and isolated contacts. Every edge causes only constantly many updates. Lower-dimensional components are included from the outset: an isolated segment on a scan line is a critical fiber, while a sloping segment gives degenerate trapezoids.

The strata's scan-coordinate intervals are ordered and disjoint, with singleton critical strata owning the boundary coordinates. Merge these intervals with the ordered, exactly owned scan-profile partition. There are only $O(K + P_{\text{scan}} + 1)$ nonempty intersections. Coincident cuts and singleton cells require only constant work per cut; no profile cell is compared separately with every role or stratum.

For each intersection, project the constant-complexity stratum restricted to that scan cell onto s by exact linear elimination. Convexity gives one interval, with exact endpoint ownership. A participating

strict inequality makes the corresponding eliminated comparison strict; zero coefficients are checked directly. Thus an open zero-width projection is empty, an excluded limiting endpoint remains excluded, and a constant edge can still attain its value at an allowed interior point. Closing the slab before projecting its vertices would not preserve these distinctions.

The other profile's intersecting cells form a contiguous index range. Binary endpoint searches cost $O(1+\log(P_{\text{other}}+1))$, and a static range-extremum query returns both the extreme score and an attaining cell in constant time after linear preprocessing (Bender & Farach-Colton, 2000). Scanning a selected-candidate cell uses the maximum outsider score; scanning an outsider cell uses the minimum selected score. Equality is adverse exactly when the outsider has better fixed priority. Repeated scores and singleton cells require no special relaxation.

Each piece therefore costs $O(K\log(K+1)+(K+P_{\text{scan}}+1)(1+\log(P_{\text{other}}+1)))$. Summing the constant-many orientations and adding the $O(m^2)$ grouped image construction proves (27). The merge projects individual convex strata, so it never replaces a disconnected projection by its hull. Streaming pieces gives the stated storage bound.

At a hit, retain the stratum and both attaining owned profile cells. Their intersection is a nonempty constant-complexity rational polyhedron; elimination and an interior rational choice where needed recover (x,y). Set $\lambda = L_Q(x,y)$. For a nonglobal piece, scan the original N roles once and test all five inequalities (17) at these point coordinates. Some role passes by the exact construction. The global piece instead uses the fixed pair attaining g. This recovery needs no role pointer to survive clipping or the boundary sweep.

Form the closed coordinate box with these extrema anchors, the candidate singletons and $[0,1]$ elsewhere. With upper bounds U_v, the classical extension

$$q_u = \min_v\{\lambda B_{uv} + U_v\}$$

preserves the prescribed coordinates and satisfies the scaled difference constraints. Its anchors give minimum zero and maximum one, so $r = q/\lambda \in \mathcal{P}(B)$ has exactly the returned normalized pair. Direct extension and checking every original inequality cost $O(m^2)$ once. Normalizing and evaluating the full vector verifies changed top-k membership. The binary64 specialization retains exact nearest-even conversion and replay of the original scorer as additional charged work. □

Corollary 12 (Reuse on an identical indexed zone)**.** *Fix the indexed matrix and normalization contract of Theorem 2. For tasks $t = 1,\dots,T$, let $P_i^{(t)}$ be valid supplied-profile sizes, S_t a target with $0 < |S_t| < m$, and the tie priority a fixed total order within each task. Put $E_t = S_t \times S_t^c$, $E = \bigcup_t E_t$ and $D_{ij}^{(t)} = m + P_i^{(t)} + P_j^{(t)}$. Retaining each first-used pair image permits total work*

$$O\left(m^2 + |E|m^2 + \sum_{t=1}^{T}\left[\sum_i P_i^{(t)} + \sum_{(i,j)\in E_t} D_{ij}^{(t)}\log(2+D_{ij}^{(t)}) + I_t m^2\right]\right), \tag{28}$$

where $I_t = 1$ when task t returns a full counterexample and is zero otherwise. Retained geometric output and working storage can use $O(m^2 + |E|m + \max_t \sum_i P_i^{(t)})$ words. These are exact arithmetic/comparison and word-RAM bounds after closure and profile construction; rational bit costs, input/cache validation and original scorer replay are additional.

Proof. Each $\mathcal{Q}_{ij}$ depends only on B, i, j, so its $O(m)$ pieces and owned strata can be retained after their $O(m^2)$ construction. Apply Proposition 11 to each task's profiles and adverse score order. Linear score-range preparation is charged anew. At a hit, the once-returned role scan, extension and original-constraint check cost $O(m^2)$; no role pointer must survive in the cached boundary. Summing these costs proves the claim. Transient construction uses $O(m^2)$ words; the retained boundary/stratum outputs use $O(|E|m)$. □

The construction term replaces repeated pair-image builds by at most one build per requested ordered pair; it is not a lower bound or a measured speedup. Changed profiles or priorities require fresh score processing, and changed targets can require previously unused pairs. A different candidate order, uncertainty matrix or normalization contract invalidates reuse. Validating each newly supplied dense matrix adds its own $O(m^2)$ input work; any shared sufficient checks are also charged separately. All

comparators may reuse their own valid domain-dependent preparation. An implementation retaining additional role tables must report its actual memory rather than the compact-output bound above.

Consequences and computational scope. The grouped construction gives the main theorem's additional branch $m^2 + (m + P_i + P_j)L(m + P_i + P_j)$, with $L(P) = 1 + \lceil \log_2(P + 1) \rceil$. The once-returned full witness fits its shared dense-zone budget. Existing branches remain useful: for $P_i, P_j = \Theta(m)$ the batched tables cost $O(m^2)$, while for $P_i, P_j = \Theta(m^2)$ the image sweep costs $O(m^2 \log m)$ instead of the tables' $O(m^4)$. These procedure comparisons imply no running-time lower bound or algorithmic optimality claim. Constants and the nonnegative order-domain branch are unchanged. Closure, rational bit costs, model-to-profile construction and scorer replay remain separate costs. The operation bounds do not imply a measured computational advantage on every workload.

Original and clipped role-event implementations. The original and clipped implementations use the earlier $O((N + 1) \log(N + 2))$ role-event construction. It sorts all role knees and endpoints, including zero, and maintains heaps for the maximum horizontal height H and diagonal intercept D. Between consecutive events the frontier is $\max\{H, D - u\}$, omitting absent terms, with at most one crossing. Ending roles remain present when evaluating a closed event fiber, before the following open interval is processed. This preserves exact ties and lower-dimensional contacts. The construction retains $O(N + 1)$ event subdivisions and does not coalesce them into the candidate-linear representation proved above. The general segment-envelope construction rate is classical (Hershberger, 1989); the factored-cap argument establishes the smaller geometric output size. Implementation timings and counters for these variants include their retained subdivisions and clipping work.

Grouped implementation. The grouped implementation processes candidate-indexed cap groups and coalesces maximal affine frontier spans within each transformed patch, retaining exact clipping, closed contacts and directional strata. It reuses the owned-profile comparison and full-vector recovery interfaces. Its fixed qualification contains 3,690 checks: the original 3,574, 34 clipping checks and 82 additional grouped cases, all completed without failed or unvisited entries. These are finite software checks, separate from the proof and from measured workloads. The resulting internal spans and strata are not an explicit maximal-boundary representation of the full seven-piece union; their counters must not be identified with the theorem's maximal boundary-edge count.

Finite implementation checks. The signed-zone prototype implements the seven active patches, grounded frontiers, exact convex clipping and directional strata. It merges one owned profile with each stratum sequence and uses an attaining-index range query on the other, without enumerating their cell product. Rational input bounds and owned cell endpoints remain exact; profile scores are finite binary64, with signed score zeros canonicalized. Profile construction, exact operand costs and one full raw-vector recovery remain charged. Nonnegative zones are explicitly unsupported by this prototype; the automatic hybrid is not implemented. The check evaluates supplied profiles, without running a fitted model.

A prospectively fixed 3,574-case bank contains 20 geometric primitive cases, 2,910 owned-cell feasibility queries, 60 direct owned-profile comparisons, 540 full binary64-profile decisions, eight structural configurations and 36 invalid/capped rejections. Its 30 signed zones comprise 12 named targets, 12 fixed-seed rational constructions and six identity-permuted copies. The independent oracle enumerates original extrema roles in raw-range affine feasibility; it does not use the new pair-image construction. All checks pass. The 540 whole decisions comprise 321 invariants and 219 counterexamples; these are dependent software cases, not a robustness rate or scientific cohort. The 275 feasible cell queries and 57 adverse owned-profile comparisons recover full rational vectors. Primitive inputs separately exercise closed jumps, zero-dimensional contributions and degenerate clipping, without claiming that every synthetic role shape is a realizable closed-zone input.

The eight structural configurations use two fixed zones with profile-size pairs $(1, 257)$, $(257, 1)$, $(65, 97)$ and $(129, 129)$. They check a linear bound on merge visits and identical geometry counts across profile lengths. The diagnostic cell-query path is disabled during whole-profile checks. These source/counter checks test the implemented mechanism; they do not prove an asymptotic bound from finite measurements or demonstrate speedup. A separate replay checks the saved geometry and raw witnesses; invariant labels are checked against the reference oracle rather than independently

re-solved. The reproducibility package contains all inputs and results. These finite checks supplement the mathematical proofs.

Incremental-clipping regression. The clipped variant replaces the nonglobal slab/fiber clipping routine with incremental exact closed-halfplane intersection, retaining every public interface and the remaining geometry/profile/witness code. It passes the identical 3,574-case bank and 34 additional exact-contact, degenerate-set, invalid-input and resource-cap checks. The supplement covers narrow rational strips, non-dyadic cuts, lines, points, contradictory rows and piece unions. All 540 whole decisions again agree with the reference (321 invariant and 219 adverse). A saved-result audit replays 1,102 implementation/reference witnesses against 14,188 original bounds and independently checks the saved supplementary sets. These are repeated finite software checks, not new application samples. Different rational intermediates and operation counts can reach an unchanged cap at different points; exhaustion remains a failure, never an invariant verdict. The reproducibility package contains the implementation and complete results. The matched timing comparison appears in Appendix E.5.

B.12 Nonnegative zones: exact order regions

Assume $g = 0$, so every entry of B is nonnegative. Define $u \preceq v$ when $B_{uv} = 0$, and quotient mutual-zero classes to obtain a finite poset Q. Closure supplies transitivity. Before fixing the attained extrema, the normalized constraints form the classical order polytope (Stanley, 1986). With the constant convention, the exact image is

$$\{0\} \cup \{z \in [0,1]^Q : z \text{ is isotone}, \ \min z = 0, \ \max z = 1\}.$$

Necessity follows from zero bounds. For sufficiency, let ϵ be the smallest positive entry of B. Then $r = \epsilon z$, lifted to all candidates, satisfies every positive bound since $z_u - z_v \le 1$, and every zero bound by isotonicity. If no positive entry exists, Q has one class and the image is only zero. Positive magnitudes therefore do not affect this image.

Let M, H be the minimal and maximal classes. Prescribe values x, y for the compared classes I, J. They must respect their mutual order, including equality when $I = J$. Remove prescribed classes with value > 0 from M and those with value < 1 from H, giving sets M_0, H_1. A nonconstant extension exists exactly when

$$M_0 \neq \varnothing, \quad H_1 \neq \varnothing, \quad \text{and } M_0, H_1 \text{ are not the same singleton.}$$

Indeed any attained zero propagates to a minimal class and any attained one to a distinct maximal class, proving necessity. Conversely choose distinct $a \in M_0, h \in H_1$ and prescribe $z_a = 0, z_h = 1$. A prescribed class can lie below minimal a only by equaling it; the symmetric argument holds above maximal h. Also $h \preceq a$ would force $h = a$. Thus all prescriptions are isotone. Setting each remaining value to the maximum of zero and its prescribed predecessors extends them, with attained extrema. The scaling above recovers a full rational raw witness. The point $(0, 0)$ is always included separately via a constant potential.

This criterion is constant on eleven disjoint regions: four corners, four open edges, the interior diagonal, and the two strict interior triangles. Order, class IDs, extrema membership and cardinalities take $O(m^2)$ preparation from closed B. Only two prescribed classes can be deleted from either extremal set; cardinalities and its first three class IDs give remaining emptiness and singleton identity in $O(1)$. Each pair therefore has a constant-size exact region mask, without extrema-role enumeration.

For supplied real profiles, corners use point scores and edges use one interior score scan. Clip cells to $(0, 1)$ for the remaining regions. The diagonal is an ownership-aware merge of the two interval partitions: advance the earlier upper cut, ordering equal-coordinate open cuts before closed cuts. Compare scores whenever their exact intersection is nonempty. For the triangle $x < y$, nonempty cells I, J intersect it precisely when $\inf I < \sup J$. Scan I in reverse order and grow the suffix of J cells satisfying this strict inequality, maintaining its maximum score and attaining cell. For $x > y$, a forward scan grows the prefix satisfying $\inf J < \sup I$. Scores need not be monotone. Each pointer advances at most its profile length, so all eleven regions cost $O(P_i + P_j)$ with constant auxiliary scan storage. Candidate priority resolves score equality.

On a triangle hit, choose rational $x \in I \cap (-\infty, \sup J)$ and then $y \in J \cap (x, \infty)$; both intervals are nonempty. A singleton supplies its owned endpoint, and otherwise an interior midpoint suffices.

Reverse the construction for the other triangle. Edges and diagonal use their exact owned intersections. The isotone extension then gives the full witness. Complete top-k verification follows by testing every selected– outsider pair, with total work $O(m^2 + \sum_i P_i + \sum_{i \in S, j \notin S}(P_i + P_j))$ after closure and profile construction, and $O(m^2 + \sum_i P_i)$ storage. Bit costs and full witness replay retain the theorem's scope. This is a specialization of classical order geometry, not a new general polytope theorem or an implemented performance result.

B.13 Interval-derived zones: eight roles and linear profile processing

Independent raw intervals have extra structure even though shared normalization couples their scores. The following specialization supplies a stronger complete comparator than general role indexing for that input class. It follows from role dominance and the polygon sweep. Its endpoint-input binary64-profile implementation is evaluated by the generated checks described below; its operation bound is separate from measured runtime.

Proposition 13 (Complete verification from interval endpoints)**.** *Let the raw uncertainty be a nonempty rational box $\mathcal{B} = \prod_{v=1}^{m}[L_v, U_v]$ with finite endpoints. Give candidate v a complete supplied real-cell score profile of size P_v, fixing every other model input and supplying total tie priority as ranks. For a proposed set S, $0 < |S| < m$, uniform top-k membership under shared min–max normalization, with the zero-vector convention for constants, is decidable in*

$$O\left(m + \sum_v P_v + \sum_{i \in S,\, j \notin S} (P_i + P_j)\right)$$

exact arithmetic/comparison and word operations, using $O(m + \sum_v P_v)$ working storage. Noninvariance yields a full rational witness in $\mathcal{B}$. Model-to-profile construction, rational bit costs and recognition of this structure from a dense matrix are additional costs.

Proof. Define $B_{vv} = 0$ and $B_{vw} = U_v - L_w$ for $v \neq w$. This matrix is closed: for distinct intermediate identities, a two-edge path adds the nonnegative width $U_q - L_q$ to the direct bound; repeated-index cases use the diagonal and interval nonemptiness. Every box vector belongs to $\mathcal{P}(B)$. Conversely, $r \in \mathcal{P}(B)$ can be translated into $\mathcal{B}$ precisely when

$$\max_v (L_v - r_v) \leq \min_w (U_w - r_w).$$

These are exactly the difference inequalities, with the repeated-index cases following from $L_v \leq U_v$. Translation preserves normalization, so the two domains have the same normalized image, including constants.

Put $L_* = \max_v L_v$ and $U_* = \min_v U_v$. The diagonal exception gives

$$b_v = \min\{0, U_v - L_*\}, \qquad c_v = \min\{0, U_* - L_v\}, \qquad g = \min\{0, U_* - L_*\}.$$

Thus eligible minimum and maximum identities are $A = \{a : L_a \leq U_*\}$ and $H = \{h : U_h \geq L_*\}$. Fix compared candidates $i \neq j$. Coordinatewise larger coefficients in $D(a, h) = (B_{ia}, B_{ja}, B_{hi}, B_{hj}, B_{ha})$ preserve all five role inequalities at every positive scale, including their endpoint flags. The common inequalities do not depend on the role.

Partition each role side into identities i, j and the remaining eligible identities $A_0 = A \setminus \{i, j\}$, $H_0 = H \setminus \{i, j\}$. Retain at most two distinct fixed/fixed pairs. For a generic minimum and a fixed maximum, take a minimum-L member of A_0; for a fixed minimum and generic maximum, take a maximum-U member of H_0. These two cases contribute at most two pairs each. They dominate every role in their category without a diagonal alias.

For generic/generic roles, let a_* minimize L on A_0 and let h_* maximize U on H_0, breaking ties by identity. If either side is empty, there is no role. If $a_* \neq h_*$, this pair dominates the category. Otherwise write their common identity as t. Retain (a_2, t), where a_2 minimizes L on $A_0 \setminus \{t\}$, and (t, h_2), where h_2 maximizes U on $H_0 \setminus \{t\}$, whenever defined. A distinct role with $a \neq t$ is dominated by the former; a role with $a = t$ is dominated by the latter. Equal-identity roles cannot satisfy $\lambda B_{ha} \geq 1$. Hence at most $2 + 2 + 2 + 2 = 8$ representative roles suffice for every cell pair. This counts representatives, not possibly tied nondominated labels.

This dominance argument does not use $g < 0$ or a finite upper scale. Discard any representative with $B_{ha} = 0$. For each remaining role, substitute point coordinates x, y into the seven common inequalities (5) and five role inequalities (17), and impose $0 \le x, y \le 1$. This is a constant-size closed polyhedron in (λ, x, y). Because $B_{ha} > 0$, its last role inequality implies $\lambda \ge 1/B_{ha} > 0$. Eliminating λ therefore gives the exact constant-complexity closed pair polygon, also when λ is unbounded above. Polyhedral projection does not add infinite-scale witnesses. Empty polygons, segments and points are allowed. Their union is exactly the nonconstant pair image. Constants contribute the zero vector separately exactly when $L_* \le U_*$.

The owned-cell polygon sweep above costs $O(P_i + P_j)$ per representative, hence per pair. Its strict slab projection and four monotone scans retain singleton cells, excluded endpoints and priority equality. Prepare the endpoint extrema, eligibility flags, and four smallest eligible L and four largest eligible U values, with identities, once in $O(m)$ work. Four entries suffice after excluding i, j and at most one overlapping role identity. Representative extraction, all needed B entries and each polygon then cost constant work; no dense matrix or closure is formed. Shared profile range extrema cost $O(\sum_v P_v)$ preparation and space.

For a hit, use the original cell endpoint flags to select a finite rational scale from its nonempty one-dimensional interval. An unbounded interval admits a finite choice above its lower endpoint. Put $t = 1/\lambda$ and find the raw minimum s in the intersection

$$[L_* - t, U_*] \cap [L_a, U_a] \cap [L_h - t, U_h - t] \cap \bigcap_{v \in \{i,j\}} [L_v - tu_v, U_v - tl_v].$$

For each compared cell, the lower s endpoint is excluded exactly when u_v is excluded, and the upper endpoint exactly when l_v is excluded. Check that any compared minimum/maximum alias owns $0/1$. These intervals are precisely the compatibility conditions for the raw box and the fixed range. Choose an owned rational s, set $r_a = s, r_h = s + t$, and choose every other r_v in $[L_v, U_v] \cap [s, s + t]$, also intersecting the translated owned cell for i, j. This constructs and checks the full raw-box witness in $O(m)$ work. Scan the supplied profiles and use linear-time selection under the fixed priorities to recover changed membership in $O(m + \sum_v P_v)$ work; a sorted full ranking is not required. For the constant branch, any common raw value supplies its witness.

Every changed set has an adverse selected/unselected pair. Checking constants and these complete pair images proves the claim. Streaming roles and pairs uses the stated storage, and the once-returned witness cost is absorbed by the displayed total bound. □

Thus shared normalization by itself does not require a dense extrema-role index. Any benefit from general indexing must be assessed against this simpler interval specialization when it applies. In particular, for $\beta \ge 0$, the enclosure $B_{vw} = x_v - x_w + \beta$ for $v \ne w$ is interval-derived via $[L_v, U_v] = [x_v - \beta/2, x_v + \beta/2]$ and admits this simpler complete procedure. It remains an outer uncertainty set when other application constraints are omitted; no application-specific realizability follows.

Implemented endpoint-input specialization. The separate implementation constructs endpoint extrema and representative roles directly, then uses exact owned-cell polygon projection and the linear profile scan. It forms no dense matrix or closure. Its fixed check completes 13 oracle unit cases, 1,508 pair-cell queries, 718 whole-ranking decisions and 47 invalid, empty or resource/interruption cases (2,286 total). The reference oracle operates on raw-box variables rather than implementation role polygons. All 788 feasible cell witnesses and 268 whole-ranking counterexamples are replayed against original bounds, normalization and owned cells; whole-ranking witnesses additionally check scores, priority ties and changed membership. These finite checks under external resource caps supplement the proof and are distinct from the application experiments. The reproducibility package includes the implementation, reference oracle, inputs and complete results.

B.14 A STRICT SEPARATION FROM COORDINATEWISE ROLE ENVELOPES

The motivating example in Section 3 is a later analytical construction, not a replacement or additional query in the fixed study below. In the order (p, q, t_1, t_2), its exact closed difference matrix is

$$B = \begin{pmatrix} 0 & -1 & 1 & 1 \\ 5/4 & 0 & 9/4 & 9/4 \\ 2 & 1 & 0 & 1/4 \\ 2 & 1 & 1/4 & 0 \end{pmatrix}.$$

Fixing translation by $r_p = 0$ gives exactly $1 \le r_q \le 5/4$, $-1 \le r_{t_1}, r_{t_2} \le 2$ and $|r_{t_1} - r_{t_2}| \le 1/4$. Every displayed off-diagonal entry is an attained difference supremum on this set; hence the matrix is closed and consistent. The zone has nonempty interior after fixing translation, and no two coordinates have a fixed difference. In every feasible vector $r_q - r_p \ge 1$. If both p, q normalized strictly inside $(0, 1)$, the other two coordinates would have to attain the minimum and maximum; their separation is at most $1/4$, a contradiction. The two adverse score cells are interior even after taking exact binary64 preimages, so their scores cannot occur together. Candidate p therefore strictly outranks q, t_1, t_2 on the whole zone. Both nonconstant profiles attain all their scores: setting $r_q = 1, r_{t_1} = r_{t_2} = -1/3$ gives normalized $p = 1/4$, whereas setting $r_q = 1, r_{t_1} = r_{t_2} = 4/3$ gives normalized $q = 3/4$.

In fact the exact real pair image in Figure 2 is

$$\big(\{0\} \times [1/2, 1]\big) \cup \big([0, 1/2] \times \{1\}\big).$$

Necessarily p is a minimum or q is a maximum. In the first case $z_p = 0$ and $z_q = r_q / \max r \ge 1/2$, since $r_q \ge 1$ and $\max r \le 2$. In the second, write $\min r = -a$ with $0 \le a \le 1$; then $z_q = 1$ and $z_p = a/(r_q + a) \le 1/2$. Conversely, every vertical point $(0, y)$ is realized by $(0, 1, 1/y, 1/y)$ for $1/2 \le y \le 1$, and every horizontal point $(x, 1)$ by $(0, 1, -x/(1-x), -x/(1-x))$ for $0 \le x \le 1/2$. Both constructions obey all raw constraints, including endpoints and tied extrema. Their coordinate hulls give the depicted marginal rectangle.

For explicit comparison with the declared filters, put $\delta = 2^{-57}$. The exact nearest-even preimages of the adverse machine cells are

$$I = [1/8 - \delta,\ 3/8 - 4\delta), \qquad J = [5/8 - 8\delta,\ 7/8 - 8\delta).$$

Their lower midpoint ties belong to the first included machine value, and their upper midpoint ties belong to the excluded next cell. The separate score extrema do not certify p over q; both profiles are nonmonotone, so the compatible-order filter does not apply. The minimum possible raw range is one and the maximum is $9/4$. The declared globally scaled and closed difference bounds for (p, q) are $C_{pq} = -4/9, C_{qp} = 1$. The point $(1/4, 3/4) \in I \times J$ obeys both strictly, so that filter leaves the adverse rectangle unresolved.

The actual extrema roles are $(p, q), (p, t_1), (p, t_2), (t_1, q), (t_2, q)$. Thus possible minimum and maximum sets each include candidates outside $\{p, q\}$, defeating the declared endpoint-role exclusion. The five roles have three distinct keys for this compared pair. Their coordinatewise envelope, including the weight, is

$$E = (1,\ 9/4,\ 2,\ 1,\ 9/4).$$

The seven common scale inequalities for I, J all hold with strict slack at $\lambda = 1/2$ whenever strictness is required. At the same scale every envelope inequality holds with strict slack: $l_p < E_1/2$, $l_q < E_2/2$, $1 - u_p < E_3/2$, $1 - u_q < E_4/2$, and $1 < E_5/2$. Consequently the envelope also passes at the common supremum, with its ownership rules. But every actual role has either $a = p$ or $h = q$, forcing a zero first or fourth coefficient, incompatible with the interior cells. The coordinatewise maxima combine different roles and supply no joint witness. This proves strict incompleteness of these particular filters, not of all inexpensive verifiers. Direct enumeration of five roles already resolves this example; it establishes neither index necessity, runtime benefit nor a new application population.

B.15 FIXED COMPLETE-VERIFIER FEASIBILITY STUDY

Before running a comparison, we fixed 52 fixtures and 206 decision queries: 36 primary fixtures spanning $m \in \{4, 8, 16\}$, requested profile sizes 2 and 8, and two fixed seeds; six simple signed

controls; and ten reused correctness fixtures. Each uses $k = 1$ and $k = \min(3, m-1)$, with duplicate k removed, under identity and reversed storage orders carrying the same semantic priority and reference potential. Requested profile counts can shrink after adjacent-score merging and sharp hull clipping; all three counts are retained. Orders and k values are dependent queries, not independent samples or confidence-interval replicates.

The primary families are dense signed zones, aligned correlated profiles intended to separate joint from marginal verification, and shifted profiles with feasible adverse choices. Seed-zero and four-candidate correlated fixtures are predeclared structural controls. The remaining candidate stratum has 80 queries, distinct from 88 structural and 38 reused-control queries. The reproducibility package includes the exact construction formulas, all inputs and every per-query result.

An older independent raw-range affine enumerator supplies each label. Signed queries use the Cartesian implementation; nonnegative controls use the retained polygon sweep. The latter do not count as executions of the new order-domain procedure. Each adverse raw potential is separately replayed against the original constraints, raw profiles, nearest-even rounding and semantic priority. Sufficient filters are applied cumulatively: sharp marginal score ranges, compatible profile/order comparisons, global normalized-difference bounds, endpoint-role exclusion, and coordinatewise role envelopes at the same common scale. The envelope retains the maximum of each role coefficient separately; it can exclude a cell pair without building the full index, but does not establish joint role feasibility.

Table 4: Diagnostic workload of 206 queries. The final two columns count invariants certified by marginal checks and by all sufficient filters.

Query stratum	Queries	Invariant	Marginal	All filters
Candidate	80	34	26	34
Structural control	88	52	36	52
Reused control	38	28	24	26
Total	206	114	86	112

Table 4 reports the diagnostic outcomes. All 206 complete-verifier labels agree with the oracle; 92 adverse witnesses pass independent replay, and all 103 storage-order pairs agree semantically. At each primary size there are invariants that fail the marginal test and checked noninvariants. However, every invariant candidate query is resolved by the role envelope. The two residual reused queries are one reused three-candidate obstruction under the two orders; they cannot supply a new candidate regime.

The prespecified feasibility criterion required at least one invariant candidate surviving every filter, with two nonconstant active profiles, multiple deduplicated role keys and more than two rigid-offset classes. No candidate met this criterion, so no comparative timing phase followed. No cases were retuned or retried. The run stayed within its fixed 600 CPU-second, 900 wall-second, 1 GiB worker-memory and 100 MiB output caps. These resource observations enforce the protocol, not a speed comparison. The constructed bank is neither an application cohort nor a random sample of all difference zones. Its negative finding leaves practical advantage of the full index unestablished; it does not invalidate the completeness proof or establish that the filters suffice on every input.

B.16 Soundness of marginal score certificates

The general verification theorem accepts a supplied difference zone. The following proposition states how a sound raw-feature enclosure and sound score bounds certify the ranking of a reference input.

Proposition 14 (Soundness under the declared machine contract)**.** *Fix an ordered candidate pool and its reference raw-feature vector $r^\star$. Suppose a sound enclosure gives a finite, consistent, exactly closed matrix B with $r^\star \in \mathcal{P}(B)$. Let $[L_i, U_i]$ be the coordinate hulls from Proposition 1 and set*

$$I_i = [\mathrm{RN}_{64}(L_i), \mathrm{RN}_{64}(U_i)] \cap \mathbb{F}_{64}, \qquad z_i^\star = \mathrm{RN}_{64}\big(\mathcal{N}_i(r^\star)\big),$$

where $\mathbb{F}_{64}$ denotes finite binary64 values and $\mathcal{N}$ is Eq. (1), computed exactly before one nearest-even binary64 conversion. Fix every other input byte, the finite machine scorer F_i, including its branch and accumulation order, and a total candidate priority π (smaller is preferred). Suppose $[\alpha_i, \beta_i]$

encloses $\{F_i(z) : z \in I_i\}$. *A size-*k *set* S, $1 \le k < m$, *is the reference top-*k *set whenever, for every* $i \in S$, $j \notin S$,

$$\alpha_i > \beta_j \quad \textit{or} \quad (\alpha_i = \beta_j \textit{ and } \pi(i) < \pi(j)). \tag{29}$$

The reference orders $F_i(z_i^\star)$ *by decreasing score and then increasing priority. Strict separation in every pair gives a strict certificate. The same selected set is valid for every* $r \in \mathcal{P}(B)$ *under this fixed decision map.*

Proof. The enclosure assumption places $r^\star$ in $\mathcal{P}(B)$. The coordinate hulls contain $\mathcal{N}_i(r)$ for every feasible r; monotonicity of nearest-even rounding places its machine input in I_i. Thus each score lies in $[\alpha_i, \beta_i]$. The displayed condition makes every selected candidate outrank every unselected candidate, including equality resolved by priority. Hence their top-k set is S, in particular at $r^\star$. □

For the declared boosted-tree scorer, whole-ensemble scalar profiles give exact extrema over each I_i (Eq. (14)); any sound enclosing score ranges also suffice. Exact coordinate hulls need not form the exact joint normalized set, and a failed sufficient test is inconclusive. Invalid enclosures, inconsistent constraints or arithmetic rejection yield no certificate. The cases $k = 0$ and $k = m$ have trivial set identity. The guarantee fixes one variable scalar feature per candidate, all other model inputs, the scoring arithmetic and the tie priority.

Table 5: Ten-method development comparison. Dense methods share 60 supported signed cases; native intervals have 14 supported endpoint cases. Unsupported cases remain recorded outside these supported-domain denominators. The adapter CPU limit is five seconds per call.

Method	Cases completed in all three repeats	Successful calls
Grouped pair image	58/60	174/180
Clipped pair image	54/60	162/180
Original pair image	48/60	144/180
Direct zone QF_LRA	51/60	153/180
Cartesian index	46/60	138/180
Linear RMQ	32/60	96/180
Binary sweep	32/60	96/180
Active-scale tables	30/60	90/180
Cell enumeration	27/60	81/180
Native intervals (different domain)	10/14	30/42

C GROUPED VERIFICATION ON CONSTRUCTED WORKLOADS

We evaluate ten methods on a 62-case development workload and five methods on 52 separately specified inputs. Timings from these experiments are reported separately from the nine-method comparison in Appendix D. The 52-case experiment tests computational coverage and cost within stated constructed families; it does not establish a distribution of learned-model inputs.

C.1 TEN-METHOD DEVELOPMENT COMPARISON

All ten methods are executed on the same 62-case development workload. The execution contains 150 qualification entries and 1,860 measurement entries, including 198 unsupported and 498 censored measurement calls. There are no failed, invalid or unvisited entries. The audit replays 845 retained counterexamples, including late returns without crediting them as completed timing samples.

On the 54 cases completed by both grouped and clipped methods, grouped has lower median adapter CPU on 49; the median within-case clipped/grouped ratio is 1.267. Against Cartesian, grouped completes 12 additional cases, but is faster on only 19 of their 46 jointly completed cases; the median Cartesian/grouped ratio is 0.960. This is a completion–time tradeoff on reused development inputs. Other processes were active on the host, so these measurements do not represent isolated-hardware timings.

C.2 FIXED 52-CASE INPUT DESIGN

This study fixes five methods: grouped image, clipped image, Cartesian index, active-scale tables and direct zone QF_LRA.

Graph-metric cases without conditioned robustness labels. The 16 cases cross $m \in \{8, 16, 32, 64\}$, profile length $P \in \{32, 128\}$, and two generating graphs: a ring and two rings joined by one bridge. Edge weights are exact multiples of $1/256$ in $[1, 2]$, selected by domain-separated SHA-256 keys with seed 2026092101. Let d be exact graph distance and $D = \max_{ij} d_{ij}$. Set $s_0 = 0$, $s_1 = D + 1$, choose the remaining potentials by the fixed hash rule, and supply

$$B_{ij} = d_{ij} + s_i - s_j.$$

The matrix is closed, $r = s$ is feasible with positive slack off the diagonal, and $B_{01} \leq -1$ guarantees positive raw range. These are complete dense matrices, despite the sparse graphs used to generate them. The two values of P share the same matrix for a fixed graph and m.

Each profile has starts j/P and scores $L_{ij} + j/(4P)$, where $L_{ij} \in \{-8, \ldots, 8\}$ is hash-selected. All values are exactly representable in binary64 and adjacent scores differ. The proposed top-2 set is computed at $r = s$ using exact normalization, one nearest-even rounding and fixed priority ties. No protected scores or known counterexamples are inserted. Robustness labels are initially unknown, and no case is filtered by its label, difficulty or residual-pair count.

Controlled-scaling cases within the proved construction. The 36 cases extend the controlled frontier and positive-width offset construction in Appendix D, using seeds 2026092102 and 2026092103. Each specification has an analytically invariant and an adverse variant with identical profiles. Three axes are fixed:

- Core size: $c \in \{6, 10, 18\}$, $k = 2$, $P = 128$ (12 cases).
- Profile size: $c = 10$, $k = 2$, $P \in \{16, 64, 256, 1024\}$ (16 cases).
- Group size: $c = 10$, $k \in \{1, 3\}$, $P = 128$ (8 cases).

There are $m = c+2k-2$ candidates, $K = k^2$ intended residual comparisons and $2kP+c-2$ supplied cells. All nonconstant cells are scalar-attained; each intended pair has P comparison-effective classes and survives the shared marginal and adverse-hull screens. Exact checks retain the proof range $6 \le c \le 18$, $P \le 1024$, and the offset/rounding margins. Adverse variants carry full feasible raw witnesses. The group axis changes both m and K; it is not an isolated intervention on the number of pairs. Related seeds, profile settings and paired labels are not independent statistical samples.

C.3 Execution, timing and validity

All methods receive identical decision inputs; expected labels, known witnesses and construction metadata are withheld. They share exact input validation, closure, machine-profile preparation, marginal screening and coarse-hull screening. Every returned counterexample is checked against the original raw constraints, exact normalization, nearest-even rounding, original profiles, priorities and complete selected-set membership.

There are 780 measurement calls: 52 cases, five methods and three fresh-process repetitions. Within each cohort, method order rotates by $(i+r) \bmod 5$, with zero-based case index i and repeat r. Before measurement, all 75 entries of a reused 15-case qualification set are classified: 55 supported calls pass and 20 are declared unsupported. These inputs exercise two residual invariants, six residual counterexamples, one hull-screened invariant and two marginal-screened invariants per method; the four-pair invariant must visit all four pairs. Separate 37-fixture adapter regressions test malformed inputs, endpoint certificates, priority ties, rounding ownership, explicit caps and interruption.

The primary limit is five adapter CPU seconds, including adapter loading, deferred imports, preparation, screening, solving, recovery and replay. The post-import cold view starts after common-adapter and deferred engine imports; whole-worker CPU additionally includes startup and input authentication. No input or query structure is reused between calls. Exclusive phase timers separate construction from comparison, including lazily built active tables. Unlike the nine-method study, this adapter timer also includes common-adapter loading; timings from the two protocols are therefore reported separately.

Worker limits are ten CPU seconds, fifteen wall seconds, one GiB monitored RSS, one MiB output and 256 KiB logs. Arithmetic limits remain 64 candidates, 20,000 supplied cells, 2,048-bit rationals, 8,192 elimination rows, one million projections, five million counted work units, one million table entries and 100,000 geometry elements where applicable. Z3 5.1.0 uses a fixed QF_LRA encoding, seed zero and one thread; its native intermediates are not claimed to share Python operation counters. The host is an Apple M1 Pro with 16 GiB memory and Python 3.11.7. Calls are serialized and host activity is sampled; ordinary operating-system and source-preparation activity is not an idle-host guarantee.

C.4 Results and their scope

All 855 scheduled entries are retained. Measurement yields 591 valid returns and 189 adapter-CPU censors, with no unsupported, failed, invalid or unvisited measurement entry. Every case/method status is stable across its three repeats. The saved audit replays 477 counterexamples, including 30 from the gate, and finds no conflicting valid verdicts.

All 16 graph-metric inputs turn out to admit counterexamples. Grouped, clipped and Cartesian complete all of them. Grouped is faster than Cartesian on eight and slower on eight; their median within-case Cartesian/grouped adapter ratio is 1.038. Thus the graph-metric cohort extends coun-

Table 6: Fixed 52-case study: cases with valid within-budget answers in all three repetitions. A counterexample requires original-input replay. Invariant completion and independent computational corroboration are separate: direct QF_LRA corroborates 8 of the 18 grouped controlled-scaling invariants within budget; the remaining invariants retain their analytical construction authority.

Method	Graph	Controlled invariant	Controlled adverse	Total
Grouped image	16/16	18/18	18/18	52/52
Clipped image	16/16	14/18	18/18	48/52
Cartesian index	16/16	6/18	16/18	38/52
Active-scale tables	15/16	2/18	4/18	21/52
Direct zone QF_LRA	12/16	8/18	18/18	38/52

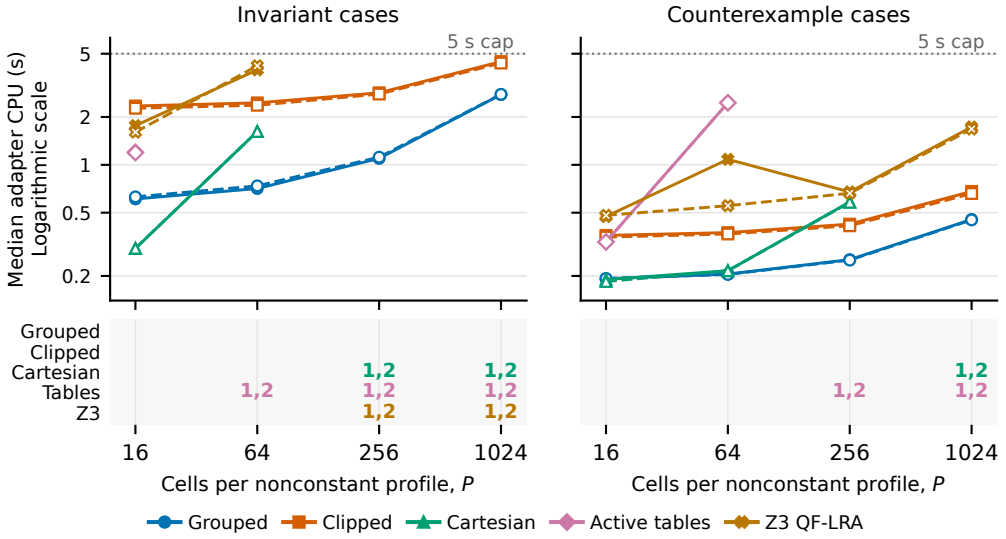


Figure 5: Controlled-scaling profile-axis adapter CPU medians for $c = 10$, $k = 2$, retaining both fixed seeds. The lower strips identify capped seeds; these groups have no numeric runtime point. Invariant and counterexample panels distinguish complete pair traversal from early stopping.

terexample coverage to the unconditioned graph families but supplies no new hard invariant cases and no completion advantage over these two competitors.

The main completion gain occurs on the 18 controlled-scaling invariants. Grouped completes all 18, versus 14 for clipped and six for Cartesian. On the 14 grouped/clipped jointly completed invariants, grouped is faster in every case and the median clipped/grouped ratio is 2.540. On the six grouped/Cartesian matches, grouped is faster on four, with median ratio 2.242. The grouped/QF_LRA comparison has eight matches and median ratio 4.576, or 4.621 when deferred imports are excluded. Ratios use within-case medians of three successful calls for both methods; censored times are never imputed.

Grouped visits all $K \in \{1, 4, 9\}$ comparisons on controlled-scaling invariants, but stops after its first comparison on every graph-metric and controlled-scaling counterexample. The latter results therefore do not demonstrate the cost of processing all residual pairs. On the controlled-scaling profile axis, grouped invariant medians increase from 0.610–0.629 seconds at $P = 16$ to 2.768–2.775 seconds at $P = 1024$; adverse medians increase from 0.192–0.193 to 0.451–0.453 seconds. Both fixed seeds remain visible in Figure 5.

The exclusive timings help locate costs. Summing grouped CPU over each cohort, zone preparation and the shared hull screen consume approximately 35.5% and 30.6% of adapter CPU on graph-metric cases. On controlled-scaling invariants, image construction and comparison consume approximately 35.8% and 45.4%. These are aggregate cost shares for the retained calls, not per-instance guarantees. The released tables preserve all stages, all capped cases, and implementation counters. Internal stratum or subdivision counts are not measurements of maximal boundary complexity and do not experimentally establish the tight asymptotic bound.

Computational costs and limitations. Input construction and exact checking consume 8.409 CPU seconds. The two 185-entry regression executions consume 127.507 CPU seconds in total; the second uses strengthened validation. Record consistency checks add 3.090 seconds. Measurement, including qualification and supervision, consumes 2,173.844 CPU seconds and 1,890.035 wall seconds. Saved-result checking and analysis add 20.318 CPU seconds. These stages total 2,333.168 CPU seconds, excluding implementation development and interactive analysis.

The controlled-scaling frontier construction deliberately preserves residual invariants and shared-core structure that grouped compression can exploit. Its gains cannot be interpreted as 36 independent hard geometries or as general application utility. The ten controlled-scaling invariants without independent within-budget corroboration remain analytically constructed cases, not newly proved arbitrary instances. The fixed graph cohort, separate strata, negative comparisons and retained censors limit the claim to improved computational completion in the measured regimes. The reproducibility package provides the source and inputs, execution schedules, individual results, witness replay and derived figure data for both comparisons.

D NINE-METHOD COMPARISON ON A FIXED CONSTRUCTED WORKLOAD

This comparison uses fixed seed 2026091803, nine methods and one execution schedule. Its input families and profile sizes were specified before construction or measurement. The seven-method comparison and matched clipping ablation appear separately in Appendix E; their timings are not pooled. The workload and implementations were developed by the same team, so this is a constructed benchmark rather than an external sample of learned models.

D.1 COMPLETE WORKLOAD AND ANALYTICAL LABELS

The 62 measurement inputs comprise 25 invariant and 37 adverse cases. The controlled family has 36 inputs: core size $c \in \{8, 14\}$, selected/competing group size $k \in \{1, 2, 3\}$, profile length $P \in \{16, 128, 1024\}$, and both labels. It uses the shifted-metric core from Appendix E, with fresh domain-separated hash coefficients. For $n = c - 2$, the core has $c^2 - 3c + 3$ eligible roles: 43 or 157. The full candidate count is $m = c + 2k - 2$.

Each additional selected coordinate has raw value $r_p + e_s$, where $e_s \in [s\delta/k, (s + 1/2)\delta/k]$; each competing coordinate has $r_q - f_t$ with the analogous interval, $1 \le s, t < k$ and $\delta = 2^{-30}$. These positive-width independent offsets remain strictly between the anchors, so they cannot attain a raw extremum. If $a(u)$ is a coordinate's core anchor and $[L_u, U_u]$ its signed offset interval, the exact closed extension is $B'_{uv} = B_{a(u),a(v)} + U_u - L_v$ for distinct indices and zero on the diagonal. All candidates stay in the normalization pool.

The profile boundaries follow the proven core frontier, with fixed margins for offsets and nearest-even rounding. A prescribed crossing-cut owner sequence cycles through the k^2 selected/competing identities. Consequently, no one candidate supplies the worst pointwise profile throughout the core domain. All P cells are scalar-attained, adjacent scores differ, and all k^2 pairs survive exact marginals and the coarse adverse-hull screen. The invariant construction excludes every adverse cell pair; changing one prescribed bridge gives the paired adverse input and a full feasible witness. This removes a single-candidate dominance shortcut, but does not rule out more general shared-core or grouped compression. The pair geometries are not independent hard instances.

The 12 seeded cases use $m \in \{10, 18\}$, $k \in \{2, 4\}$ and the same three profile sizes. Closed shifted-metric zones retain positive-width uncertainty and fail the rectangle equality of interval-derived matrices. Hash-selected scores vary across all candidates and cells. Protected nominal cells fix the proposed set; a protected cell reached by one explicit feasible raw perturbation forces a counterexample. At least k adverse pairs survive the common screens, but a method may stop at its first full witness. These are counterexamples by construction, not unbiased random labels or independent model profiles.

There are also 12 signed endpoint controls with $m \in \{8, 16\}$, $k = 2$, the three profile sizes and both labels, plus two nonnegative endpoint controls at $m = 8, k = 2, P = 16$. Authoritative endpoint bounds or an exact interval certificate allow the native endpoint procedure. Arbitrary coupled zones do not: replacing them with boxes would change the decision problem. The eight dense adapters support the 60 signed cases and explicitly exclude the two nonnegative ones. The native adapter supports the 14 endpoint cases and excludes the other 48. Unsupported calls remain in the fixed denominator and receive no timing credit.

The reproducibility package specifies all algebraic constructions, owner sequences, rational margins and hash rules. Input validation checks all 77 measurement and qualification inputs, including 171,256 triangle inequalities, 77 nominal vectors, 45 adverse vectors and 178 adverse-hull pair witnesses. Construction costs 20.568837 CPU seconds; the input audit costs 1.799668 CPU seconds. These offline costs are distinct from supplied-input verification and do not stand in for learned-model profile extraction.

D.2 METHODS, QUALIFICATION AND RESOURCE ACCOUNTING

The methods are original role-event image, clipped role-event image, active-scale tables, linear-RMQ polygons, binary polygon sweep, Cartesian indexing, adverse-cell enumeration, Z3 QF_LRA and native interval verification. The first eight receive identical decision fields, without expected labels, known witnesses or construction metadata. They share validation, exact closure, profile preparation

Table 7: Successful measurement calls over every planned family, including three repetitions. The remaining supported calls hit a preset limit; there are no other execution failures. U denotes the declared unsupported domain: six calls per dense method, 144 for native intervals. Native interval and dense endpoint timings compare different complete pipelines.

Method	Frontier	Seeded	Signed box	Nonneg. box
Original image	72/108	36/36	36/36	U
Clipped image	93/108	36/36	36/36	U
Active tables	36/108	24/36	30/36	U
Linear RMQ	24/108	36/36	36/36	U
Binary sweep	24/108	36/36	36/36	U
Cartesian index	69/108	36/36	36/36	U
Cell enumeration	21/108	24/36	36/36	U
Z3 (QF-LRA)	81/108	36/36	36/36	U
Native interval	U	U	24/36	6/6

and exact-marginal/coarse-hull screens. Dense endpoint conversion is charged, and returned quotient witnesses must be translated into the original endpoint box and replayed there.

Native intervals invoke the frozen public endpoint verifier exactly once; they construct no dense matrix for endpoint input and use no duplicated dense profile/screen preparation. An interval certificate accompanying a supplied matrix requires a charged exact equality check. Native-versus-dense costs thus compare complete procedures with different preprocessing. Z3 retains its rational QF_LRA encoding, version 5.1.0, seed zero and one thread. All engine, adapter, environment and replay bytes are authenticated. The host remains an Apple M1 Pro with 16 GiB memory, macOS 26.6.2 and Python 3.11.7. Fresh workers do not imply cold operating-system caches.

The completed 333-call adapter regression checks 136 semantic decisions, 34 expected unsupported cases, 153 expected rejection cases and ten declared bit-limit outcomes. All 55 adverse witnesses are replayed. It consumes 93.558405 CPU seconds, excluding an earlier incomplete regression that detected an incorrect wrapper status for nonnegative inputs. The corrected wrapper reports these inputs as unsupported without changing the engine or numerical inputs; the full regression is then repeated.

The 15-case qualification set runs all 135 scheduled entries before measurement: 97 supported calls pass and 38 are unsupported. Each dense method must actually execute two residual invariant and six residual adverse paths, one hull-screened invariant and two marginal-screened invariants. The multiple-pair invariant must process all four surviving pairs. Native qualification cases exercise four invariants and five counterexamples, including the constant branch.

For each of 62 cases and three repetitions, the nine-method order is rotated left by $(3i + r) \bmod 9$, with zero-based case i and repetition r. Only supported calls launch workers. Each method occupies each ordinal position 20 or 21 times overall; this is not complete within-case order balancing. Adapter CPU includes deferred imports, conversion, preparation, screening, solving, recovery and witness replay, with a five-second limit. The separate post-import input-to-answer view excludes only deferred imports. Whole-worker CPU additionally includes startup and bank authentication. Worker limits are ten CPU seconds, fifteen elapsed seconds, one GiB monitored RSS, one MiB output and 256 KiB logs. Structural limits are 64 candidates, 20,000 supplied cells, 2,048-bit rationals, 8,192 elimination rows, one million projections, five million counted work units, one million table entries and 100,000 geometry elements where applicable. Z3 uses external limits rather than a claimed equivalent Python-work counter. No cap is raised or failed call retried.

D.3 Completion and runtime results

The single measurement attempt completes all 1,809 classifications: 135 gate and 1,674 measurement records. Measurement contains 996 PASS, 486 capped and 192 declared unsupported entries, with no invalid, failed, interrupted or unvisited entry. Table 7 retains every family. Original and clipped image, tables, linear RMQ, binary sweep, Cartesian index, enumeration and Z3 complete respectively 144, 165, 90, 96, 96, 141, 81 and 153 of 180 supported signed calls. Native intervals complete 30 of 42 supported calls and cap on 12 signed endpoint calls. Every case/method status is stable

across its three repetitions; all 486 censors are adapter-CPU caps. All five clipped-capped cases are controlled invariants. The higher completion count does not imply uniformly lower successful-case cost: clipped adapter medians are lower than Z3 on 46 of 51 jointly completed cases, versus 25 of 51 for post-import cold CPU. They are lower than Cartesian indexing on 15 of 46 jointly completed cases. Imports and completion-conditioned populations matter for interpreting these descriptive directions.

The successful controlled image invariant calls process 1, 4 or 9 residual pairs for $k = 1, 2, 3$, respectively. Their adverse image counterparts stop after one pair despite having the same residual-pair census; they do not measure complete multi-pair traversal. At $P = 1024$, clipped image completes three of the six controlled invariant cases, versus one for Z3, two for original image and zero for the other dense methods. These small constructed strata are descriptive, not estimates of population success probabilities. On signed endpoint invariant controls, all dense methods exit at the common marginal screen; native intervals have no such screen. The native caps therefore do not contradict the endpoint processing bound or isolate a slower residual kernel.

There are 332 complete three-repetition case/method groups out of 558 planned. Only these receive medians. All 2,232 unordered case/method comparisons remain in the saved table, with ratios only for the 977 having three successes from both methods. A ratio of medians and a median of matched-repeat ratios are separate quantities. Zero denominators are blank; no stopped call is replaced by an exact five-second runtime. Successful-only comparisons are conditional on their displayed completion populations.

The complete attempt consumes 4,904.312009 aggregate CPU seconds and 4,365.137964 elapsed seconds, including gate, supervision and authentication. Its separate saved-result audit costs 51.666403 CPU seconds and 50.590026 elapsed seconds. The audit authenticates all 8,542 retained measurement files, all scheduled identities, source freezes and prerequisite inventories. It replays 734 full counterexamples, checking 9,226 coordinates, 107,220 original difference constraints and 180 certified endpoint checks. These counts include gate and repeated witnesses. Replay shares a reviewed pure helper and has author overlap; it is not external replication or independent invariant solving.

The reproducibility package includes all 1,809 calls, 558 case/method groups, 2,232 pairwise comparisons, phase costs and geometry/work counters. The results show stronger coverage for the clipped role-event implementation on this declared population, with substantial remaining caps. It does not measure the grouped algorithm or establish average-case behavior on a distribution of learned models. The grouped comparison is reported in Appendix C.

E FIXED SUPPLIED-PROFILE COMPUTATIONAL COMPARISON

This study compares seven implementations, including the original image prototype, under one supplied-input contract. A matched clipping ablation is reported in Appendix E.5. The workloads were specified algebraically before execution, with seed 2026091802 and no pilot timing. Their designer also implemented the image algorithm. Thus the study tests a declared computational regime; it is not an external workload assessment or evidence of a trained model producing that regime.

E.1 INPUTS, CONTROLS AND CONSTRUCTION

There are 40 fixed cases, with 20 analytically specified invariants and 20 counterexamples. Dense coupled cases use $m \in \{6, 10, 18\}$ and $(P_p, P_q) \in \{(8,8), (64,64), (128,128), (8,128), (128,8)\}$, with both labels (30 cases). Four sparse cases use a six-coordinate core, $m \in \{10, 18\}$, and $(64, 64)$. Two cases each retain a compressible forbidden-rectangle control, signed interval input, and an explicitly unsupported nonnegative control. Every case supplies the full rational matrix, complete binary64 profiles, priority permutation and selected set. All other candidates have constant low scores, leaving one potentially adverse ranking pair (p, q). No candidate is removed from the normalization pool.

For the dense core put $n = m - 2$, $s_p = 0, s_q = 2, s_{t_\ell} = 1$, and $B_{uv} = d_{uv} + s_u - s_v$ for a symmetric metric d. Define

$$A_\ell = \frac{\ell}{4(n+1)}, \quad C_\ell = \frac{n+1-\ell}{4(n+1)}, \quad d_{pq} = 1, \quad d_{p,t_\ell} = 1 + A_\ell, \quad d_{q,t_\ell} = 1 + C_\ell.$$

The outsider distances are $d_{t_a,t_h} = 1 + V_{ha}$, with $V = 1/2$ on adjacent indices. Otherwise $V = (2048 + 11b)/16384$, where b is the first SHA-256 byte of the ASCII string `2026091802|n|a|h`, using increasing outsider indices. Diagonals are zero. Every positive distance lies in $[1, 3/2]$, so triangle closure holds directly. The only negative entry is $B_{pq} = -1$; eligible roles number $m^2 - 3m + 3$, namely 21, 73 and 273. These counts do not assert that all roles contribute distinct boundary pieces. A four-distinct-index rectangle violates the equality required by $B_{uv} = U_u - L_v$, so these dense matrices are not interval-derived.

An independent construction formula for the complete normalized core image is

$$0 \le x \le x_{\max} = \frac{n}{5n+4}, \qquad F(x) \le y \le 1,$$
$$F(x) = \min\{1, \min_{a \ne h} F_{ah}(x)\},$$
$$F_{ah}(x) = \max\left\{(1 + 1/A_a)x, \frac{1 + C_h x}{1 + C_h}, x + \frac{1}{1 + V_{ha}}\right\}.$$

Indeed the active scale is $\lambda = y - x$, and the outsider-role constraints become $0 \le u \le A_a$, $0 \le v \le C_h$, $u + v \le V_{ha}$ for $u = x/(y-x)$ and $v = (1-y)/(y-x)$. Adjacent roles also cover the endpoint aliases. The invariant frontier is continuous and strictly increasing, with $F(x) \ge x + 10/13$. The reproducibility package provides the full derivation and exact rounding margins; the constructor imports no compared algorithm.

The p cuts are $X_k = \mathrm{RN}_{64}(x_{\max} k/P_p)$; the q cuts are $Y_k = \mathrm{RN}_{64}(F(\mathrm{RN}_{64}(x_{\max} k/P_q) - \epsilon) - 2\epsilon)$, with $\epsilon = 2^{-40}$, cut indices $1 \le k < P$ on each profile, and first start zero. This margin constructs the inputs; the verifiers still use exact rounding preimages, not a tolerance. Put $M = \min(P_p, P_q)$ and $\eta_k = 1/32 + ((5k) \bmod P)/(8P)$ on a profile of length P. Scores are $-\lfloor kM/P_p \rfloor + \eta_k$ for p and $-\lfloor kM/P_q \rfloor - \eta_k$ for q. Thus q wins exactly when its class is smaller. Feasibility forces the opposite class order in the invariant zone. The paired adverse zone raises only the outsider bridge $(1, n)$ to $V = 1/2$, retaining the same profiles. Its fixed feasible raw witness has $r_p = 3/16, r_q = 19/16, r_{t_1} = 11/8, r_{t_n} = 0$, and other outsider coordinates $11/16$, giving normalized pair $(3/22, 19/22)$ with q winning.

All supplied cells meet their exact scalar marginals and adjacent scores differ. Balanced profiles have P distinct adverse-comparison classes; unbalanced profiles have only M despite their larger supplied size. Sparse cases append affine copies $r_e = r_p + 1/2$ and remain compressible. The comparison

Table 8: Fixed largest balanced generated cases: $m = 18$, $P_p = P_q = 128$, with both analytically specified labels. CPU medians including deferred imports are reported only when all three repetitions pass. C/F/U/S/M denote capped/failed/unfinished/unsupported/mixed states; parentheses give passing repetitions out of three. Completion covers all 114 signed calls per method; six nonnegative-control calls are unsupported. The 5 s adapter CPU budget includes imports; capped calls have no reported median.

Method	Invariant CPU (s)	Adverse CPU (s)	Signed PASS
Pair image	C (0/3)	C (0/3)	84/114
Active tables	C (0/3)	C (0/3)	96/114
Linear RMQ	C (0/3)	C (0/3)	96/114
Binary sweep	C (0/3)	C (0/3)	96/114
Cartesian index	1.6	0.241	114/114
Cell enumeration	C (0/3)	C (0/3)	84/114
Z3 (QF-LRA)	4.22	0.393	114/114

records supplied, merged, attained and comparison-class counts. It also records actual geometry sizes instead of inferring them from roles. The common adverse-cell-hull rectangle is feasible on the frontier invariants, although every truly adverse cell pair is infeasible. This prevents that particular coarse screen from settling them; it does not establish hardness against every possible specialized control.

E.2 Equally informed methods and charged work

The seven methods are the whole pair image, active-scale tables, linear-RMQ role polygons, binary role sweeps, Cartesian role indexing, complete adverse-cell enumeration and Z3 QF_LRA (de Moura & Bjørner, 2008). They use the same full profiles, exact sharp-marginal screen and coarse adverse-hull feasibility screen. Each fresh worker pays for input validation, exact Floyd closure, profile preparation, common screening, its native preparation and solving, and full raw-witness replay. Thus closure is included here even though the theorem states its processing bound after closure. Offline generated-profile construction is excluded and is not model-to-profile extraction. No prepared zone, profile, image or query structure is reused across calls.

The SMT encoding uses rational variables $\lambda > 0$, $q_u \in [0, 1]$, $q_u - q_v \le \lambda B_{uv}$, and disjunctions requiring some $q_u = 0$ and some $q_v = 1$. Exact owned-cell guards imply their dyadic scores; the final disjunction requests an adverse selected/outsider comparison including priority ties. SAT gives raw witness $r = q/\lambda$; UNSAT gives the invariant verdict. The baseline is Z3 5.1.0.0, with `SolverFor("QF_LRA")`, seed zero and one thread. Its original wheel, installed files and loaded library are pinned. An UNSAT answer is a solver result, not an independently checked proof object. Internal solver arithmetic is governed by external limits rather than claimed to share the Python implementations' operation counters.

The execution host is an Apple M1 Pro with 16 GiB memory, macOS 26.6.2 and Python 3.11.7. Three repetitions rotate the method order by case and repetition. A fresh process and empty bytecode cache do not imply cold operating-system caches. The primary CPU measure covers the full adapter including deferred imports; the supplementary input-to-answer view excludes those imports. Whole-process and replay costs remain separately recorded. Every adapter has five process CPU seconds including imports; each worker also has ten CPU seconds, fifteen wall seconds and a monitored one GiB RSS limit. Fixed structural limits and aggregate budgets remain in the prospective protocol. Capped or failed calls receive no invariant verdict and no exact runtime imputed at the cap. Numeric medians require three of three successful repetitions. All planned cases remain in the denominators.

The fixed largest-balanced table and all 18 balanced dense cases were chosen for display before completed outcomes. Full CSVs also retain sparse, unbalanced, band and interval controls. A displayed-bound selector among five older methods is fixed without runtime inputs; the fastest older method on a case is only a post-hoc oracle. Neither is a newly timed eighth implementation. This is a comparison of the seven supplied-B implementations, not every complete verification method. The endpoint-native interval sweep and VERITAS are not timed, and the generated profiles do not establish an ML application.

E.3 Implementation checks and execution

The fixed prior bank passes 1,344 calls: 192 inputs for each method, with 735 independently replayed full counterexamples. It contains no residual invariant after common screening. A separate twelve-case gate therefore precedes measurement, retaining two unsupported controls and exercising actual residual invariant and counterexample paths in every backend. Qualification cases are finite checks, some deliberately reused; they are not independent application samples. Source reviews and separate saved-result audits have distinct scopes.

An incomplete measurement run was interrupted by a logging error and is excluded from every timing comparison. It consumed 769.779 CPU seconds and 671.698 wall seconds. After correcting only the logging interface, the full schedule was repeated with unchanged algorithms, inputs, order and limits. No results from the incomplete run were pooled with the completed run.

The completed run contains all 924 scheduled classifications: 70 supported qualification calls pass, 14 qualification controls are unsupported, and the 840 measurement entries comprise 684 successes, 114 CPU-capped calls and 42 unsupported controls. Every signed case/method has the same status across its three repetitions. Image, tables, linear RMQ, binary sweep, Cartesian indexing, cell enumeration and Z3 respectively complete 84, 96, 96, 96, 114, 84 and 114 of their 114 signed calls. No capped call is an invariant certificate. The attempt consumes 1,687.793 aggregate CPU seconds and 1,475.036 wall seconds; including the preserved first attempt gives 2,457.572 CPU seconds and 2,146.734 wall seconds. Prior adapter qualification and separate audit costs are additional.

The saved-result audit authenticates all scheduled records and source/input bindings and replays 377 full counterexamples, checking 3,820 coordinates and 47,048 original difference bounds. This count includes qualification witnesses and repeated measurement witnesses. Invariant labels use the reviewed analytical construction and authenticated solver results; the auditor does not perform another feasibility solve. Audit authorship overlaps the geometry and workload implementation. The replay is therefore an internal reproducibility check rather than an external replication.

E.4 Complete balanced-grid results and interpretation

Figures 6 and 7 show every balanced dense case under both CPU scopes, including all capped cases. No main-table case was selected after timings. Image processing caps on all ten dense $m = 18$ cases, regardless of profile length. Cartesian indexing and Z3 complete all 38 signed cases; Z3 is faster than image processing in every jointly completed dense case. At $m = 6, P = 128$, the invariant image median is 0.601 seconds versus Cartesian indexing's 1.369 seconds, a local crossover; at $m = 10, P = 128$, the corresponding 1.404 versus 1.430 seconds are very close. These do not establish superiority over the strongest alternatives. The all-five-old post-hoc oracle is eligible on 22 of the 40 planned cases; image processing beats it on none in either CPU view. The pre-result heuristic selector is a different, weaker comparator. Full per-repetition observations, all 40-case strata, eligible ratio populations and phase/counter records remain in the authenticated CSVs. No global speedup, statistical generalization, optimality or ML benefit is inferred from this generated study.

Reproduction records. The reproducibility package provides workload formulas, adapter and SMT definitions, source snapshots, inputs, per-case results and witness replay. Dependency versions and platform information are recorded separately because restoring saved inputs does not reproduce the original runtime environment.

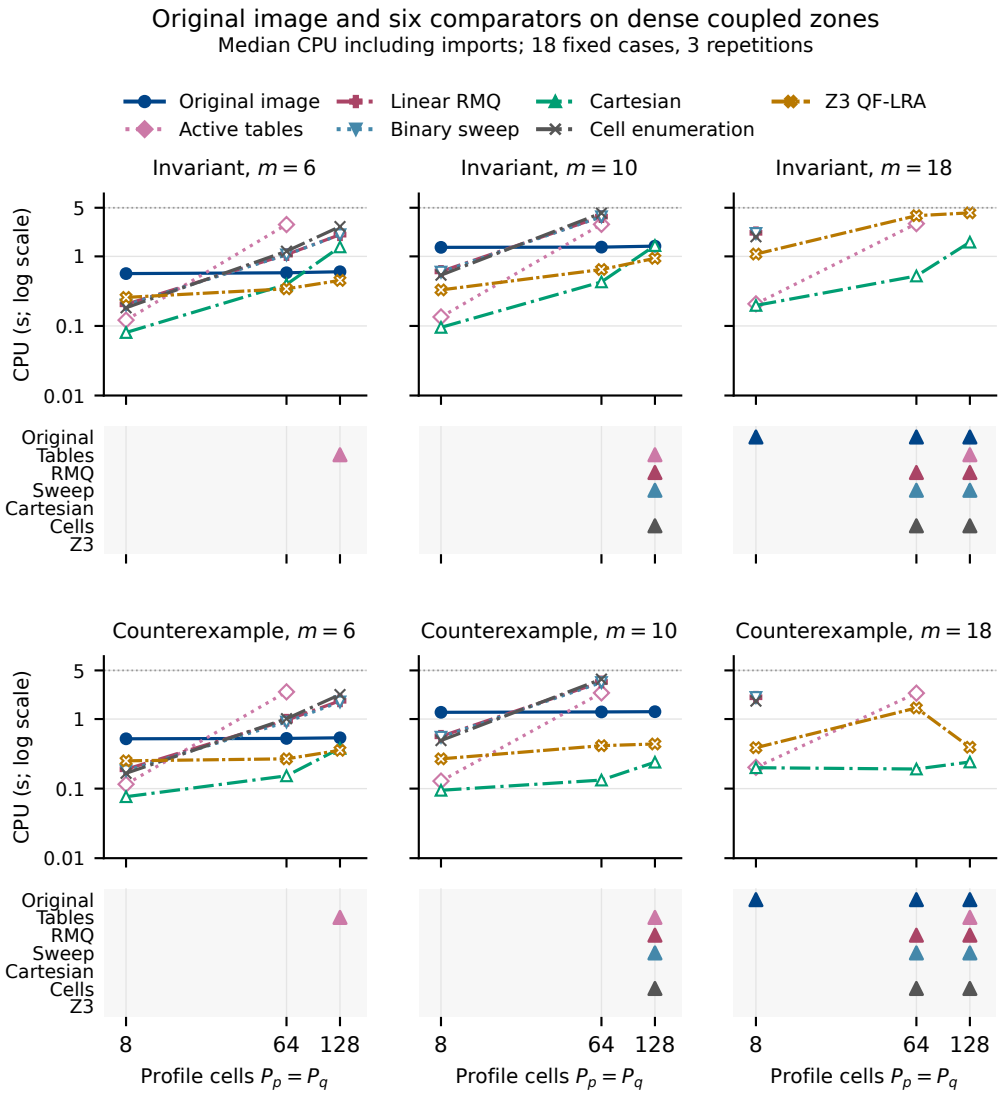


Figure 6: All 18 fixed balanced dense cases, with median adapter CPU including imports. Numeric points require three successful repetitions. The separate status strips show capped cases without assigning them the cap as an observed runtime. Cartesian indexing and Z3 complete every signed case; the original ungrouped image prototype caps for every dense $m = 18$ input. This prototype is distinct from the grouped construction evaluated in the main text.

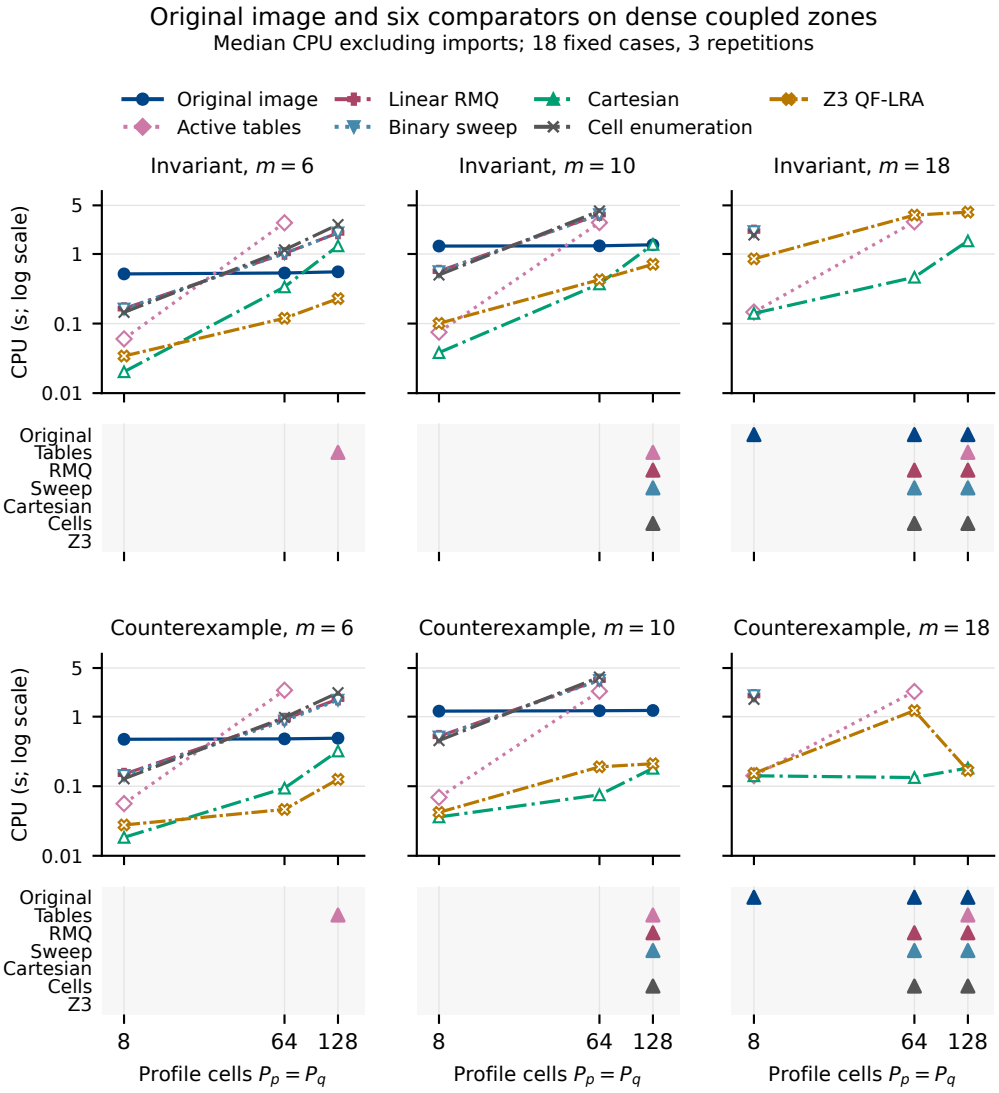


Figure 7: The same complete balanced grid with deferred imports excluded. Validation, closure, common profiles/screens, method preparation, solving and full witness replay remain charged. The adapter cap still includes imports; this secondary view does not rerun or uncap any call. Both figures use identical time limits, case populations and status rules.

Table 9: Matched development remeasurement at every dense dimension with $P_p = P_q = 128$. Medians include imports and require all three repetitions to pass. C denotes capped calls; no runtime is imputed. Completion across all signed cases is 84/114 for the original and 114/114 for clipping.

m	Label	Original CPU (s)	Clipped CPU (s)
6	Invariant	0.603	0.174
6	Adverse	0.540	0.146
10	Invariant	1.386	0.305
10	Adverse	1.280	0.267
18	Invariant	C (0/3)	1.035
18	Adverse	C (0/3)	0.946

E.5 Exact clipping and a matched development comparison

The original comparison motivated a localized implementation change: replace repeated pairwise boundary intersections inside each envelope slab/fiber by successive exact closed-halfplane intersections. A slab begins as its grounded trapezoid, triangle, segment or point. Each cut retains its inside vertices and exact boundary crossings, with explicit point/segment handling and collinear collapse. Equality remains inside; distinct pieces are never replaced by the convex hull of their union. The global active patch, role dispatch, envelope construction, owned strata, profile comparison and full witness recovery are unchanged. This is an implementation optimization, not a new geometry theorem. Its regression scope is stated in Appendix B.11.

The development study uses the exact saved 40-case bank and all 12 gate cases. Both original and clipped code are remeasured in fresh processes through the same adapter and authenticated source closure. Only their geometry package differs; common preparation, profiles, screens, solving interface and replay remain symmetric. The two-method source closure excludes unused alternative backends; both variants pay the same authentication and import policy. Historical seven-method timings are therefore not pooled with these fresh timings. Every internal cap and the five-second whole-adapter CPU budget remain fixed. The three repetitions alternate the two methods by case and repetition. Each method is first in 60 measurement blocks; the odd repeat count gives a 2:1 order split within each case rather than complete within-case balance.

The qualification set exercises, per variant, two residual invariants, five residual counterexamples, one hull-resolved invariant, two marginal-resolved invariants and two unsupported controls. Actual geometry, profile preparation, range queries and full counterexample replay are required. Its 20 supported calls pass.

All 264 planned records are retained: 24 gate entries and 240 measurements. The clipped version passes all 114 signed measurement calls; the original passes 84 and has 30 adapter-CPU caps, covering every repetition of all ten $m = 18$ dense cases. Each variant also retains six unsupported measurement calls. There are no failed, invalid, interrupted or unvisited entries. The saved-result audit authenticates every identity and receipt and replays 109 full counterexamples, including gate/repeated witnesses, against 12,652 original difference bounds. Invariant labels retain the reviewed construction and executed-check scope; the saved auditor does not solve their feasibility.

Among the 28 of 40 cases eligible for three-of-three paired comparisons, clipped adapter-CPU medians are lower in 26 and higher in two; input-to-answer medians excluding imports are lower in 27 and higher in one. These are per-case descriptive directions, not significance tests or a successful-only aggregate speedup. Early-screen differences do not isolate a clipping effect because those calls never reach that operation. The two unsupported cases and ten original-capped cases remain in the tables and have no paired ratio. A cap is not an exact five-second observation. For the largest balanced $m = 18, P_p = P_q = 128$ cases, clipped adapter-CPU medians are 1.035444 seconds for the invariant and 0.946091 for the adverse case; excluding imports gives 0.986879 and 0.898187 seconds, respectively. The original has no median for either label. Table 9 also retains both labels at the smaller dense dimensions.

The entire development attempt consumes 385.822224 aggregate CPU seconds and 337.216971 wall seconds; its saved audit consumes 1.487380 CPU seconds and 1.485600 wall seconds. Those totals include qualification, authentication and supervision and are not per-method speed ratios. The

reproducibility package includes all 264 calls, 80 case/variant summaries and 40 paired comparisons. The ratio of the two medians and the median of three matched-repeat ratios are separate columns. Complete cost views, counters, caps and source bindings stay available. No case is filtered according to timing or success.

This is outcome-informed development evidence: the original bank exposed the implementation cost and was then reused to evaluate the change. It does not supply independent performance confirmation, multiple-adverse-pair scaling, a comparison against newly timed Cartesian/SMT alternatives, or a general ML uncertainty contract. Workload and implementation/review authorship overlap.

Table 10: Clean Fold1 validation quality over all 157 queries; no significance or TEST generalization claim. Hyperparameters and preprocessing are fixed. The independent saved-result audit reproduces all nine macro values within 1.12×10^{-16}.

Fixed scoring rule	NDCG@1	NDCG@3	NDCG@10
Raw body BM25 (feature 21)	0.3503	0.4114	0.5127
Raw PageRank (feature 41)	0.1210	0.1716	0.3396
Fixed LambdaRank ensemble	0.3779	0.4265	0.5296

F LETOR INTERVAL-UNCERTAINTY EVALUATION

We evaluate clean ranking quality on the validation split and interval robustness on a fixed 32-query subset of TRAIN. All robustness conditions use one 100-tree model. The implementation checks below are reported separately from application results; no TEST model evaluation is performed.

F.1 DATA AND SCORER CONTRACT

The original MIN decimals and Fold1 memberships are joined by semantic $(\mathrm{qid}, \mathrm{docid})$, rejecting duplicate, missing or changed labels and cross-split query overlap. MIN preprocessing uses exact rational min–max normalization followed by one round-to-nearest, ties-to-even binary64 conversion; constant columns map to zero. Complete released candidate pools are retained. This declared preprocessing need not reproduce rounded values in the released query-normalized files. Semantic identity, split membership and labels are checked before fitting. There are 471 TRAIN, 157 validation and 156 TEST queries. The first 32 TRAIN query IDs ordered by the SHA-256 of ASCII `letor-box-v1:<qid>` form the development population, with integer qid as tie-break. The selection precedes model fitting and robustness outcomes. The source-fixed list and every semantic row identity are retained in the input freeze.

The model is LightGBM 4.6.0 CPU GBDT LambdaRank with 100 rounds, at most 31 leaves per tree, learning rate 0.1, configured `min_data_in_leaf=20`, seed 20260917, one thread, deterministic column-wise training, no row/feature subsampling, no extra regularization and no early stopping. The leaf setting is a configuration parameter rather than a hard lower bound on realized leaf counts. It uses all 46 features and all 9,630 rows from the 471 TRAIN queries. The frozen model contains 88 feature-41 splits across 57 trees; split usage establishes presence, not useful robustness or causal importance. The reproducibility package identifies the exact model. Clean NDCG uses gain $2^{\mathrm{label}} - 1$, discount $\log_2(\mathrm{position} + 2)$ with zero-based position, macro query averaging, fixed docid ties and NDCG zero for zero ideal gain. The independent saved-result audit reconstructs all validation identities, labels, baseline orders, normalization receipts and 1,413 per-query NDCGs.

Clean validation context. Table 10 reports descriptive macro NDCG on all 157 validation queries (2,707 rows), against raw body BM25 and PageRank. Every query is retained, including 37 with zero ideal gain, assigned NDCG zero. Scores tie-break by lexicographic document ID. The saved predictions, identities, labels and all 1,413 per-query NDCGs were independently audited; native/text nominal scores agree on every validation row. These are clean ranking measures, not robustness coverage or evidence that verification improves relevance.

Declared raw-feature uncertainty. Only PageRank, feature 41, is uncertain; the other 45 normalized model inputs remain fixed. For nominal raw values a with range $R_0 = \max a - \min a$, use

$$[L_i, U_i] = [\max(0, a_i - \rho R_0), a_i + \rho R_0], \qquad \rho \in \{0, 10^{-3}, 10^{-2}, 0.05, 0.1\}.$$

Recompute shared normalization over each raw vector in the box. This models corruption of the released feature values, without claiming that every vector is generated by a consistent perturbed web graph. Zero-range queries remain singleton boxes. Membership sizes are $k \in \{1, 3, 10\}$; $k \geq m$ is the trivial full set. Feature, radii, k and the deterministic 32-TRAIN-query development population were fixed before model fitting, giving 480 conditions. These queries were used to train the same model; this is development evidence.

Native machine profiles. The adapter accepts the declared finite numerical-threshold tree format and raw-score evaluation, with fixed nonuncertain inputs. It preserves native double-precision thresholds and leaves, left ownership at equality, signed zero score bits and ordered binary64 tree accumulation. In the dense float64 native interface, $|x| \leq \kappa$ is folded to zero, where κ is the binary64 promotion of the binary32 literal 10^{-35}; this occurs even when `zero_as_missing=false`. Profile construction accounts for this transformation in all input columns and follows only reachable branches. Nonfinite inputs or intermediate scores are rejected. These details are part of the specified scorer, not floating-point tolerances.

The finite native bank passed for 20 manually specified models and 25 profiles, including tiny thresholds, signed zero, fixed-feature branching and accumulation order. It checks model serialization, native leaf values, every atomic boundary and its neighbors, and explicit invalid-input rejection. Application profiles require a separate complete native parity check before use; the completed development check covers all 663 profiles, including their atomic boundaries and neighboring machine values against the native float64 raw scorer. The synthetic bank alone supplies no robustness certificate for LETOR rows.

F.2 Exact marginal controls and complete comparator

For each raw box, each candidate's exact attainable normalized scalar set is obtained by projection over at most eight representative extrema roles plus the constant branch. The owned union can be disconnected. Intersecting it with the complete machine profile yields attained score extrema; using only its coordinate hull could admit unattainable interior scores. Endpoint summaries, scalar unions and profile intersections cost $O(m + \sum_i P_i)$ and are prepared once per radius. Separate processing contexts share this read-only geometry, with fresh counters, fixed limits and isolated failures for each k/method call.

The sufficient sequence comprises exact marginal separation, compatible order, closed global-range bounds, endpoint-role availability, joint extremum-identity availability, common-scale feasibility and a coordinatewise role envelope. The identity filter checks whether one distinct minimum/maximum pair can satisfy the endpoint aliases. The envelope may combine bounds attained by different roles; neither test is a complete cell feasibility oracle. Adding all five role inequalities for one joint role at the same common scale is complete. Exhaustive adverse-cell enumeration using that conjunction is therefore a complete comparator and cannot be a source of incremental certificates for another complete method. Its per-pair processing is $O(P_i P_j)$, versus $O(P_i + P_j)$ for the interval sweep; these are operation bounds, not an observed timing advantage. A surviving sufficient relaxation is UNKNOWN, never a counterexample.

A separately frozen generated control bank passed 58 preparations, 155 candidate marginals, 4,340 scalar-cell queries, 399 profile-cell queries, 290 dense-bound comparisons, 394 independent whole-decision oracle calls, 2,758 sufficient-control calls, 394 complete-comparator calls and 30 safety cases. The independent oracle uses raw box variables. Checks cover disconnected attainability, exact RN64 ownership, constants, tie priority, distinct extrema, invalid/capped inputs and isolated shared-preparation forks. Every positive counterexample is replayed; finite checks supplement the proof without proving universal software correctness.

F.3 Evaluation protocol and computational costs

Each fixed query retains every radius and k combination. One complete profile is constructed per row and reused; each radius receives one marginal preparation. The controls, complete enumerator and interval sweep run under independent per-call budgets and a common global ceiling of 1,800 CPU seconds, 3,600 wall seconds, 2 GiB aggregate RSS and 256 MiB output. Per-call CPU/wall caps are 2/4 seconds for the sufficient sequence and 3/6 seconds for each complete method. The candidate cap is 128, above the complete TRAIN maximum of 121. No pool is shortened after a cap. Counters, profile sizes, exact raw bounds and all method labels are retained. A resource failure is FAILED or BLOCKED, not UNKNOWN or a counterexample. Every COUNTEREXAMPLE must change full top-k membership when its rational raw vector is normalized, rounded and evaluated by the fixed native scorer.

Table 11: Every fixed radius/k cell, each containing all 32 development queries. Each sufficient stage, starting with exact attainable marginals, has the same labels. All 21 interval-sweep failures are fixed per-call CPU/wall caps. Trivial full-set conditions remain included; they total 75 across this table.

		All sufficient stages		Complete enumeration		Interval sweep		
ρ	k	Inv.	Unknown	Inv.	Cex.	Inv.	Cex.	Failed
0	1	32	0	32	0	32	0	0
0	3	32	0	32	0	32	0	0
0	10	32	0	32	0	29	0	3
1/1000	1	32	0	32	0	32	0	0
1/1000	3	32	0	32	0	31	0	1
1/1000	10	32	0	32	0	28	0	4
1/100	1	32	0	32	0	32	0	0
1/100	3	30	2	30	2	29	2	1
1/100	10	29	3	29	3	26	2	4
1/20	1	30	2	30	2	30	2	0
1/20	3	24	8	24	8	24	7	1
1/20	10	22	10	22	10	21	9	2
1/10	1	30	2	30	2	30	2	0
1/10	3	20	12	20	12	20	11	1
1/10	10	21	11	21	11	20	8	4

Extraction, native parity, radius preparation, method processing and native witness replay are charged separately and in total. Internal rational witness recovery/replay is included in method processing. Fixed-order, capped observations constitute feasibility accounting rather than repeated comparative timing. The complete enumerator shares prepared marginals and repeats cheap pruning. The standalone interval verifier does not consume those verdicts: it rebuilds its box, profiles and range-query structures and directly sweeps the roles on every call. Consequently the observed method totals do not isolate quadratic versus linear profile processing under matched preparation or pruning. The independent saved-result audit authenticates every recorded artifact and reconciles 32 query rows, 160 radius rows, all 663 frozen candidate rows and 480 condition rows.

F.4 Complete fixed development outcome

The evaluation retains all 32 queries, 663 candidate rows, 160 radius preparations and 480 conditions. Table 11 reports every radius/k cell. Complete enumeration returns 430 invariants and 50 counterexamples. Of the invariants, 355 pass exact marginal checks and 75 select the entire candidate set, requiring no comparison. Every later sufficient stage has unchanged labels. The interval sweep returns 416 invariants, 43 counterexamples and 21 per-call `StageLimit` failures. Its 459 decisive results agree with complete enumeration. There are no contradictions, but 21 comparisons are not jointly decisive.

All 663 complete profiles pass native parity. The 6,157 atomic cells remain 6,157 merged cells, with at most 16 per profile and no constant profiles. All 32 nominal feature-41 ranges are positive; 75 of the 480 conditions are trivial full sets. Table 12 charges every phase, including profile extraction, parity and the 93 returned native witness replays. This is a fixed feasibility execution, not a matched runtime benchmark. The complete attempt costs 305.11 aggregate CPU seconds and 298.12 wall seconds, including construction, parity, shared preparation and all method calls. Every nontrivial invariant in this fixed population is marginal-certified. The joint interval method supplies no incremental certificate, and its capped execution does not establish a practical advantage. Because no invariant required joint reasoning beyond these checks, the study did not proceed to TEST evaluation. No TEST model scores, profiles or robustness evaluations were produced; its files were used only for identity and schema checks. A separate result check authenticates all 57 recorded files and independently replays all 93 saved raw-box witnesses. It recomputes method counts and recorded costs and checks 2,750 raw coordinates across those witnesses, including exact normalization, binary64 rounding, saved-profile scores and full membership under semantic document-ID priority. It imports neither verifier nor controls and performs no new native model evaluation. Invariant labels remain evidence from the authenticated executed pipeline; native parity and native witness scores are authenticated receipts, whereas the independent replay evaluates the saved profiles.

Table 12: Observed costs for the single fixed-order development execution. The 75 trivial full-set conditions skip method processing; the other 405 calls per method include all successes and caps. Shared preparation is charged once per query/radius for the controls and enumeration. The standalone sweep rebuilds its geometry and profiles per call and does not use their marginal verdicts. Method processing includes its own internal rational recovery and replay; 93 native witness replays are additionally charged (50 enumeration, 43 sweep). Whole-worker and supervised totals include costs outside the listed stage buckets, including source/data checks, output and final authentication. These nested totals must not be summed together. No repeated or randomized comparative timing claim is made.

Charged scope	Calls	CPU seconds	Wall seconds
Model load/capture	1	0.0358	0.0366
Query normalization	32	0.0935	0.0936
Nominal native prediction	32	0.0094	0.0094
Complete row profile export	663	0.7444	0.7447
Native row-profile parity	663	14.4387	14.4464
Shared radius preparation	160	13.7604	13.7611
Sufficient-control processing	405	0.1548	0.1550
Complete enumeration processing	405	0.4045	0.4047
Standalone interval verification	405	260.0707	260.1710
Native full-witness replay	93	0.0527	0.0528
Whole worker	1	296.8105	297.4279
Complete supervised attempt	1	305.1092	298.1151

Reproducibility. The reproducibility package contains the input definitions, fixed protocols, model and profile identifiers, complete condition-level results, native agreement checks and witness replay. The interval implementation and its raw-box reference oracle are provided separately.

G MiniLM verification on the NFCorpus development split

This study uses the BEIR distribution (Thakur et al., 2021) of NFCorpus (Boteva et al., 2016), a public medical-information-retrieval dataset available for academic use. It supplies inputs from a newly trained ranking pipeline. It is not a production deployment or an external replication of the verifier. Its timings are kept separate from the constructed study.

G.1 Data, one fixed fit and native profile agreement

The pinned archive contains 3,633 documents, 2,590 TRAIN queries and 324 DEV queries. Only TRAIN and DEV query texts and judgments are exported; no TEST judgments or TEST model evaluation are used. Each query receives its fixed BM25 top-32 pool, with no insertion based on relevance labels. The `multi-qa-MiniLM-L6-dot-v1` encoder is frozen at revision `4151c507ffb0f2fcd311cf431f54b5fc7d097851`; its 384-dimensional binary32 CLS vectors are used without length normalization. The tokenizer, 256-token limit, model bytes and CPU runtime are pinned before encoding.

The four features are the shared min–max-normalized exact neural dot product, normalized BM25, query-term overlap and log document length. Exactly one LightGBM 4.6.0 LambdaRank model (Ke et al., 2017) is fitted on all 2,590 TRAIN groups (82,880 rows), including 690 groups with all-zero relevance labels. The fixed configuration has 100 boosting rounds, 31 leaves, learning rate 0.1, one CPU thread and seed 2026091804. There is no DEV validation set, early stopping, parameter search or refit. Actual model bytes and effective parameters are frozen before DEV prediction or profile export.

Table 13 reports every DEV query using corpus-wide ideal DCG and the same candidate pools. Dense ordering reranks the BM25 pool; it is not full-corpus dense retrieval. These descriptive scores establish neither statistical significance nor a state-of-the-art retrieval result. Macro Recall@32 is 0.164499, and 94/324 pools contain no positively judged document. Those queries remain in all reported averages.

Table 13: Clean quality over all 324 DEV queries. Saved score-to-document alignment, ranking order, judgments and macro aggregation are independently rechecked from saved bytes. No new prediction is made by that audit.

Ordering within the fixed pool	NDCG@1	NDCG@3	NDCG@10
BM25	0.364198	0.318995	0.268302
Dense feature	0.360082	0.330831	0.285782
Fixed learned ranker	0.376543	0.349867	0.291093

Before model outcomes, a salted SHA-256 ordering fixes 16 DEV queries. All 512 query–candidate profiles are exported, with the other three features fixed to their saved bytes. Every profile is nonconstant: sizes range from 45 to 87 cells, with median 68 and 33,183 cells in total. The neural feature occurs in all 100 trees and accounts for 830 of the model's 3,000 splits. All 512 profiles pass the native atomic/boundary agreement checks; the saved receipts contain 337,311 native probe rows. These checks establish agreement with this frozen model, not a robustness decision.

G.2 Quantized queries and the supplied-zone contract

For each fixed query, the 8-bit and 16-bit unsaturated coordinate quantizers use powers-of-two steps fixed from TRAIN query vectors only. The transmitted code defines an owned reconstruction box $\mathcal{U}$ for a shared query vector q. With fixed document vectors d_i, the raw feature is $r_i = d_i^\top q$; exact dot products and normalization are followed by the specified binary64 nearest-even rounding before profile lookup. The proposed normalization maps a zero raw-score range to all zeros. The top-k membership is the decoded query's native nominal set, for $k = 1, 3$, with document-ID priority ties. Thus the study has 16 queries, 32 packets and 64 related conditions; these are not 64 independent samples.

Each packet supplies the closed difference-zone outer approximation

$$B_{ij} = \sup_{q \in \mathcal{U}} (d_i - d_j)^\top q, \qquad \mathcal{P}(B) = \{r : r_i - r_j \le B_{ij}\}.$$

The exporter constructs this matrix once per packet and checks its exact triangle inequalities and the feasibility of the saved original and decoded raw vectors. The same matrix is reused for top-1 and top-3. All 32 packets and 64 conditions qualify; none is replaced or removed.

An invariant over $\mathcal{P}(B)$ transfers to top-k membership throughout the owned query box, within the frozen 32-document pool, with the other three features fixed and the stated arithmetic. A raw counterexample in $\mathcal{P}(B)$ need not arise from one shared query in $\mathcal{U}$, including because the closed outer approximation does not retain all joint or endpoint-ownership constraints. The MiniLM DEV study therefore reports such counterexamples as abstract zone witnesses. It does not run a shared-query feasibility solver or claim a realized query perturbation from a zone witness.

For the saved original and decoded queries, 8-bit quantization changes top-1 membership on 1/16 queries and top-3 membership on 3/16; these four conditions belong to four distinct queries. At 16 bits, neither set changes on the saved queries. These observations do not certify the remaining uncertainty box. Two further 8-bit top-3 changes affect internal order only, outside the membership property. Six of the 16 fixed pools have no positive judgment, and three of the four queries with observed membership changes are among them. The selected cohort's macro Recall@32 is 0.080704. Consequently, observed membership changes alone do not establish a judged relevance loss or downstream benefit from verification.

G.3 Five-method verification comparison

The unchanged grouped-image, clipped-image, Cartesian-index, active-table and Z3 QF-LRA adapters receive the same supplied matrix, profiles, priorities and selected set. A 75-entry qualification gate precedes all $64 \times 5 \times 3 = 960$ measurement calls. Method order rotates by condition and repetition. Every call starts a fresh process and has a five-second adapter CPU cap, including imports, closure, profile preparation, both common screens, solving and full raw-witness recovery/replay. Worker CPU, wall time, memory, output and exact arithmetic also have fixed caps recorded in the protocol. Offline construction of the common supplied matrix is a separate cost. No historical timings are pooled and no capped time is imputed.

The schedule completes with 55 valid and 20 declared unsupported gate entries, then 879 valid and 81 adapter-CPU-capped measurement calls. There are no invalid results, worker failures, unavailable conditions or contradictory valid answers. Completion requires three valid within-budget returns, not a majority of repeats. The saved audit replays 369 counterexamples including the gate; 339 are measurement returns, with repeated conditions retained in that count.

Table 14: MiniLM DEV conditions completed in all three repetitions; each bits/k stratum contains the same 16 queries. All four settings complete for 15/16 queries with grouped or Cartesian, 14 with clipping, 13 with tables and two with Z3. No condition is omitted because a method caps.

Method	8-bit $k=1$	8-bit $k=3$	16-bit $k=1$	16-bit $k=3$	Total
Grouped image	16	15	16	16	63/64
Clipped image	16	14	16	16	62/64
Cartesian index	16	15	16	16	63/64
Active tables	15	14	16	16	61/64
Z3 (QF-LRA)	9	3	16	16	44/64

Across methods there are 36 invariant conditions, 27 abstract-zone counterexample conditions and one condition with no decision within budget: `PLAIN-1319`, 8-bit, $k = 3$, where all five methods cap in every repetition. All 32 16-bit conditions and four 8-bit top-1 conditions are certified invariant. Every invariant exits through the shared marginal screen; agreement among these calls is not independent full-solver corroboration. All successful complete-solver returns are counterexamples. The four observed codec membership changes are consistently noninvariant in the outer zone.

Paired costs and the role of common preparation. For conditions that both methods solve in all repetitions through the complete solver, the median within-condition adapter CPU ratios relative

to grouped are 1.137 for clipping (26 conditions), 1.035 for Cartesian (27), 1.798 for active tables (25), and 5.272 for Z3 (eight). All these conditions are counterexamples; these ratios exclude shared-screen returns and censored cases. Across all 63 jointly completed grouped/Cartesian conditions, the corresponding ratio is 1.028 including imports but 0.991 after imports. The timing view and selected completion subset therefore matter, and the data do not establish a large or universal grouped advantage.

For grouped's 27 complete-solver conditions, 26 stop after the first residual pair and one after two. The median condition-level share of adapter CPU spent in the common coarse-hull screen is 58.39%. The campaign consequently exercises real learned profiles and constructive witness recovery, but provides little support for a gain in exhaustive invariant certification.

The common packet exporter spends 80.841 CPU seconds over 32 packets (median 2.528 per packet, range 2.437–2.643), including exact zone-structure and saved-vector checks; top-1 and top-3 reuse that work. Its complete supervisor cost is 86.324 CPU seconds. The total runtime, feature, fitting and packet preparation cost is 1,564.286 CPU seconds, including an unsuccessful feature-preparation run. The single training call itself uses 0.758 CPU seconds; the native profile construction/parity phase uses 564.379. Measurement including the qualification gate uses 1,895.657 CPU seconds. Saved audit and analysis use 18.631 CPU seconds, plus 0.408 for record consistency checks. These stage costs exclude data acquisition, preliminary arithmetic checks and interactive analysis. They are not end-to-end retrieval latencies or an operational amortization claim.

Interpretation. The MiniLM DEV study demonstrates that the fixed signed-zone interface accepts inputs from a real public-data model with nonconstant neural-feature profiles and that valid raw witnesses can be recovered. It does not add an invariant certificate beyond marginals, a realizable shared-query witness from every abstract counterexample, or a measured retrieval benefit. It uses one model, 16 DEV queries and limited-recall pools, with no external replication. We retain it as application evidence alongside the stronger constructed invariant comparison, without treating the 64 correlated conditions as independent samples.

H QWEN3 ENCODER EXTENSION ON THE SAME NFCORPUS DEV COHORT

The Qwen3 encoder extension tests whether the supplied-zone interface also accepts a more recent embedding pipeline. It reuses the MiniLM DEV study's data splits, BM25 top-32 pools, three lexical features and 16 selected DEV queries (Appendix G). The protocol was fixed before Qwen dataset inference and fitting, after the MiniLM DEV results were known. This is a model extension on the same cohort, not an independent dataset replication.

Encoder, training and input checks. We freeze `Qwen3-Embedding-0.6B` (Zhang et al., 2025) at revision `97b0c614be4d77ee51c0cef4e5f07c00f9eb65b3`. Its 1,024-dimensional last-nonpadding-token vectors are L2-normalized before saving in binary32. Queries use the fixed web-search retrieval instruction; documents use title and text. Encoding uses a 256-token limit including the instruction, float32 eager attention, left padding, batch four and MPS with four CPU threads. There is no encoder fine-tuning. Exactly one new LambdaRank fit uses the MiniLM DEV study's 100-round, 31-leaf configuration and seed, with no DEV tuning or TEST evaluation. Exact saved-vector dot products, shared normalization and the binary64 evaluation rule are retained. All 512 exported profiles are nonconstant, with 66–91 cells (median 74); all pass native agreement over 326,737 probe rows. All 100 trees use the neural feature.

The TRAIN-only codec uses steps 2^{-9} and 2^{-17} for 8 and 16 bits. Decoded vectors are not renormalized. All 32 packets and 64 conditions qualify. An independent saved-input checker reconstructs the codec, owned cells, exact dot products, profile lookups and every supplied pair bound without importing the tested generator. As in the MiniLM DEV study, a zone invariant transfers to the specified query box, while an abstract zone counterexample need not have a realizable shared-query witness. These are certificates for saved-vector uncertainty and a frozen tree scorer, not for text perturbations through the encoder.

Quality on all 324 DEV queries. Table 15 uses the same pools and corpus-wide ideal DCG. The registered primary comparison, Qwen minus MiniLM learned NDCG@10, is $+0.012734$, with a

paired query-bootstrap percentile 95% interval $[+0.004462, +0.021045]$ (10,000 resamples, seed 2026092301). There are 102 improved, 161 unchanged and 61 worsened queries. This exploratory DEV interval is conditional on the two fixed fits; it does not cover training seed variation or establish independent generalization. Qwen's learned head also trails its dense ordering at NDCG@1. Recall@32 remains 0.164499, and 94/324 pools contain no positively judged document.

Table 15: Fixed-pool quality; the saved-score audit independently recomputes rankings and metrics from document identities and original judgments. The encoder/ranker pipeline causes the quality difference, not the verifier.

Ordering within the fixed pool	NDCG@1	NDCG@3	NDCG@10
BM25	0.364198	0.318995	0.268302
MiniLM dense	0.360082	0.330831	0.285782
MiniLM learned	0.376543	0.349867	0.291093
Qwen dense	0.403292	0.355707	0.300522
Qwen learned	0.395062	0.363341	0.303827

Matched verification. The same five adapters receive identical supplied zones, profiles and priorities under the MiniLM DEV study's five-second adapter CPU cap, including imports and common preparation. All 960 formal calls finish their scheduled attempts: 873 valid returns and 87 CPU caps (15 for tables, 72 for Z3). The separate 75-entry gate retains 55 valid and 20 declared unsupported returns. No condition is dropped and no valid answers conflict. Completion in Table 16 requires three valid within-budget returns. The saved audit replays 318 zone witnesses, including gate returns and repetitions.

Table 16: Qwen3 extension completion on four settings for the same 16 queries. All 39 invariant conditions exit through the common marginal screen; the remaining 25 conditions have replayed abstract zone counterexamples.

Method	8-bit $k = 1$	8-bit $k = 3$	16-bit $k = 1$	16-bit $k = 3$	Total
Grouped image	16	16	16	16	64/64
Clipped image	16	16	16	16	64/64
Cartesian index	16	16	16	16	64/64
Active tables	14	13	16	16	59/64
Z3 (QF-LRA)	8	1	15	16	40/64

No invariant in the Qwen3 extension requires the joint solver. On the 25 conditions jointly completed through the full solver, the median within-condition Cartesian/grouped adapter CPU ratio is 1.593. The corresponding clipping ratio is 1.103; tables give 3.260 on 20 jointly completed conditions. Z3 completes only one such condition, so its 6.601 ratio has that single-condition scope. Across all 64 grouped/Cartesian completions, the median ratio is only 1.029 including imports and 0.990 after imports. These completion-conditioned ratios exclude capped results and do not establish uniform speed superiority.

Observed changes and interpretation. Saved 8-bit encoding changes top-3 membership for `PLAIN-2007` and top-1 membership for `PLAIN-901`; no saved 16-bit membership changes occur. The first query's only positively judged pool document stays at rank two; the second pool has no positive judgment. Both swaps exchange unjudged documents, assigned zero by evaluation, and neither changes judged top-k quality. Thus observed instability alone supplies no measured relevance benefit from certification. For `PLAIN-901`, the 16-bit top-1 condition also has a zone counterexample despite unchanged saved-query membership; the diagnostic below establishes a shared-query witness for this condition.

Post-hoc shared-query recovery. After the Qwen3 extension outcomes were known, we fixed a separate diagnostic over all 25 counterexample conditions, using the first grouped repetition and no retraining. Each condition receives two numerical LP searches: one allows positive affine transformations of the saved raw-score vector, preserving its normalized target; the other targets the

saved adverse pair's original profile cells with fixed minimum/maximum identities. Each LP has a ten-second limit. Up to six candidates, in a fixed order, undergo exact owned-cell and identical-packet checks, exact dot products and shared normalization, binary64 rounding, and agreement of all 32 profile scores with the frozen native model. The first accepted candidate is canonical; judgments do not select candidates.

All 50 LP attempts finish. The 39 proposed queries yield 35 checked counterexamples and four unchanged memberships. Search finds witnesses in 13 conditions, one overlapping the two previously known original-query counterexamples: the union is **14/25**, including 12 newly established conditions. The other 11 remain unresolved. Numerical infeasibility is not an exact exclusion proof, and the recovered queries need not reproduce the saved abstract normalized vector. An independent saved-record audit checks all 89 candidate/control queries and their rankings.

The recovered 16-bit `PLAIN-901` witness changes top-1 but not evaluated quality. For 8-bit `PLAIN-237`, the canonical top-1 witness instead replaces the grade-one `MED-4184` with unjudged `MED-2662`, evaluated as zero: NDCG@1 falls from 1 to 0 and NDCG@10 by 0.081229. Across all 13 canonical search conditions, NDCG@10 decreases in one, increases in three and is unchanged in nine. These correlated, deliberately searched embedding-box witnesses establish possible relevance loss, not its frequency, an observed encoding loss, a text-realizable perturbation or a verifier-caused quality improvement. The number of additional joint-only invariant certificates remains zero. The reconstruction box does not enforce a unit norm, binary32 query coordinates or membership in the encoder's output set. This separate search/replay campaign costs 45.31 CPU and 46.15 wall seconds; its independent audit is accounted separately.

Computational costs. Encoding all 6,547 texts takes 540.60 wall seconds and 368.68 process CPU seconds on MPS; the latter excludes GPU execution cost. The complete fit and profile stage takes 64.64 CPU seconds, including 0.67 for the single training call and 37.92 for profile construction/native checks. Packet preparation takes 246.67 CPU seconds, including 244.55 for the 32 packet/zone exports. The measurement campaign including qualification takes 1,484.51 CPU seconds and 1,295.09 wall seconds. An unsuccessful launch adds 0.43 CPU seconds before any worker call; it changes no protocol, input or solver. Data acquisition, runtime preparation and independent result checking are separate costs. These costs are not end-to-end retrieval latencies.

This extension increases model coverage and improves fixed-pool NDCG@10. Encoder dimension, pooling, normalization, codec and fitted tree weights all change, so it is not a causal model-size ablation. Pretraining overlap with NFCorpus cannot be excluded; we make no clean zero-shot or state-of-the-art claim. The two pipelines reuse the same 16 verification queries, and their 64 conditions each are correlated. Their timing campaigns remain separate.

I NFCORPUS TEST APPLICATION: SHARED UNCERTAINTY AND COMPLETE VERIFICATION

This application complements the MiniLM and Qwen3 DEV computational studies. It evaluates the complete official TEST population, with a different encoder, ranker and quantizer. We report all verification outcomes and independently recompute their aggregates. A fresh encoding, training and profile reconstruction reproduces the original model, embedding arrays, features, quantized arrays, scores, condition records and full profile payload byte for byte. This reconstruction and the mechanism case below were specified after the original results were known; neither is an unexposed confirmation experiment. Their execution costs are reported separately. The input reconstruction stage itself did not repeat the full verification population. Subsequent full comparisons, SciFact results and query-witness searches appear in Appendix J.

I.1 FIXED MODEL AND FULL-POPULATION RETRIEVAL EVALUATION

We use the BEIR distribution (Thakur et al., 2021) of NFCorpus (Boteva et al., 2016): all 3,633 documents and the official 2,590/324/323 TRAIN/DEV/TEST queries. BM25 uses $k_1 = 1.2$, $b = 0.75$, lowercased `[a-z0-9]+` tokens, no stemming or stop-word removal, and document-ID tie priority. It fixes 32 candidates per query before neural scoring; no relevance-based insertion or query exclusion occurs. The frozen `all-MiniLM-L6-v2` encoder, revision `1110a243fdf4706b3f48f1d95db1a4f5529b4d41`, produces 384-dimensional L2-normalized binary32 vectors on CPU, with batch size 32 and a 256-token limit. The encoder is not fine-tuned.

One LightGBM 4.6.0 LambdaMART fit uses 100 rounds, at most 15 leaves per tree, learning rate 0.05, seed 20260922 and one training thread. It uses all 2,590 TRAIN groups, without early stopping, hyperparameter search or refitting. Its seven features are shared min–max-normalized neural similarity, BM25 score, reciprocal BM25 rank, title and abstract query-token coverage, and log-transformed document and title lengths. Only the first varies under query-vector uncertainty. The other six remain fixed and are not constructed from relevance labels. By comparison, the two DEV studies use four features, 31 leaves, learning rate 0.1, different encoders and a 16-query DEV verification cohort. Their absolute quality and certification rates are not pooled with this study.

Table 17: Reproduced and audited macro means over all 323 TEST queries. NDCG uses $2^{\text{rel}} - 1$ gains and ideal rankings from full supplied qrels. All methods retain the same candidate pools.

Ordering within the fixed pool	NDCG@1	NDCG@3	NDCG@10
BM25	0.414861	0.360472	0.308057
Neural similarity, full precision	0.399381	0.375138	0.321347
Learned ranker, full precision	0.440660	0.399262	0.330787
Learned ranker, 8-bit midpoint	0.437564	0.398439	0.329932
Learned ranker, 16-bit midpoint	0.440660	0.399262	0.330787

The full-precision ranker's NDCG@10 difference from BM25 is 0.022731. An exploratory paired-query bootstrap with 10,000 resamples and seed 20260922 gives a percentile 95% interval $[0.011674, 0.034111]$; it did not select the model. It excludes retraining variability, and query dependence limits an independent-sampling interpretation. This is a comparison with the fixed, untuned BM25 configuration, not with optimized retrieval systems. All methods have Recall@32 $= 0.196207$. Of 323 TEST pools, 84 have no positively judged document and 305 omit at least one relevant document; only 1,269 of 12,334 relevant query–document pairs enter the pools. Unjudged documents receive zero relevance in evaluation, without implying that they are known to be irrelevant.

Official query IDs are disjoint, but five TEST texts match TRAIN or DEV after lowercasing and whitespace normalization: `PLAIN-2440`, `PLAIN-2630`, `PLAIN-3053`, `PLAIN-3131` and `PLAIN-3251`. The official population remains primary. Excluding these five in a post-hoc sensitivity analysis gives BM25/ranker NDCG@10 of 0.308531/0.331129 over 318 queries, without refitting. Encoder pretraining overlap is not audited.

The semantic feature accounts for 327/1,400 tree splits and 61.87% of training split gain. This establishes use, not causal contribution. All 20,704 DEV/TEST scalar profiles are nonconstant, with 28–54 segments and median 45. Their 5,861,038 native boundary and adjacent-value probes pass. Independently audited reference and midpoint scores comprise 62,112 predictions; 3,235 query/method metric rows are recomputed with maximum absolute discrepancy 3.33×10^{-16}. These checks establish frozen-model agreement, separately from the uniform certificates below.

I.2 Three domains and attribution of certificates

Stored binary32 vectors define exact dyadic dot products, exact shared min–max normalization and one nearest-even rounding to binary64 before tree scoring. A zero raw range maps to all zeros. This numerical reference does not assert equivalence to arbitrary floating-point cosine or GEMM code. Uniform b-bit bins partition $[-1, 1]$, with width 2^{1-b} and midpoint decoding without renormalization. Actual cells retain half-open ownership, with the last bin including $+1$. Unlike the two DEV studies, the quantizer is not TRAIN-calibrated nearest-even quantization. No unit-sphere or text-realizability constraint is added to its query cell.

Writing $\bar{q}$ for the midpoint and $h = 2^{-b}$, the independent raw-score box I uses the exact individual supports $\bar{q}^\top v_i \pm h\|v_i\|_1$. The shared difference zone D uses $B_{ij} = \bar{q}^\top(v_i - v_j) + h\|v_i - v_j\|_1$. If Q is the raw-score image of the actual shared query cell and N is normalization, then $N(Q) \subseteq N(D) \subseteq N(I)$. The second inclusion uses a common raw-score translation; D itself need not be contained in I. Every method uses the same full profiles, document-ID priority and midpoint top-1/top-3 target.

Both relaxed domains first use marginal and global-range sufficient checks. Independent preparation retains attainable coordinate unions, whereas shared coordinate hulls can contain gaps, making their score bounds conservative. The shared pipeline additionally intersects remaining pairs with a joint adverse-hull feasibility screen. Independent QF_LRA SMT receives the independent residuals. Exact clipped pair images and a separate shared SMT encoding receive identical shared residuals and separate five-second wall budgets after process initialization, including construction, solving and witness replay. A backend answer obtained by repeating sufficient checks remains attributed to those checks.

Table 18: Invariance certificates from the original complete population of 1,292 TEST query/bit/k conditions, with 323 conditions per row. Conditions sharing a query are dependent. The shared complete column includes sufficient controls; it is not a count of independent full-backend proofs. Controls are the original marginal, global-range and joint-hull checks.

Condition	Independent pipeline	Shared controls	Shared complete
8-bit, $k = 1$	74	136	136
8-bit, $k = 3$	28	49	50
16-bit, $k = 1$	316	319	319
16-bit, $k = 3$	317	319	319
Total	735	823	824

There are 89 strict coupling gains: each pairs a shared invariant with a replayed counterexample in I, excluding differences caused only by timeouts. The same global-range control certifies 735 independent versus 822 shared conditions; the original shared joint-hull screen adds one, and complete verification adds one relative to that protocol. The latter condition is `PLAIN-741`, 8-bit, $k = 3$ (1/1,292 conditions). Both clipped pair images and shared SMT certify it, so this result is not unique to one implementation. A stronger role-envelope check also certifies it (Appendix K.2); it does not require complete solving beyond that check. After removing the five repeated texts, the 318-query sensitivity population retains 86 coupling gains and the same one increment over the original three-check baseline.

In the original pass, actual-query outcomes are 824 invariants, 26 counterexamples and 442 unresolved. Outer-domain invariants transfer by inclusion; original-query membership changes supply valid actual-cell counterexamples. At 8 bits the saved original and midpoint top-1/top-3 sets differ in 4/22 TEST conditions; at 16 bits neither set differs. These are pointwise observations, not certificates for other queries. All 442 direct latent-query SMT calls reach the five-second wall cap without an additional

certificate or counterexample. An abstract counterexample or a capped search is never classified as a real-query failure. Later searches over additional precisions recover further common-query witnesses (Appendix J.4); the original counts here are not the final cumulative witness totals.

Shared clipped verification leaves one outer-domain condition unresolved; shared SMT leaves 30, with 467/438 counterexamples respectively (including the 26 original-query counterexamples). Independent SMT has 735 invariants and 557 counterexamples, with none unresolved. Of 443 shared backend calls, clipping decides 442 and SMT decides 413; 30 are decided only by clipping and one only by SMT. This is a bounded operational comparison of these pipelines. Most 16-bit candidate profiles are locally constant: 9,777/10,336 (94.6%) have constant scores over their shared marginal hulls, compared with 327/10,336 (3.2%) at 8 bits.

Execution and costs. The original full verification stage uses two query workers in one fixed pass, after 16 conditions from four fixed DEV queries pass numerical and operational qualification. All caps, repairs and unsuccessful calls remain recorded. The supervisor completes all 323 queries and 1,292 conditions in 2,023.90 wall seconds, with no query-worker failures. Table 19 reports original backend costs. This single-pass protocol differs from the algorithm comparison's three-repetition, five-second adapter CPU budget including imports; their times are not pooled. Parallel calls, preparation and sufficient checks prevent the sum of backend times from equaling supervisor elapsed time. These data do not establish an isolated-hardware speed ranking.

Table 19: Original complete verification run. Wall times are sums over calls; process wall includes backend wall, imports, initialization, IPC and termination. CPU missing counts hard-terminated calls with no CPU measurement; their CPU cost is not treated as zero.

Backend	Calls	Backend wall (s)	Process wall (s)	CPU missing
Independent SMT	531	96.859	199.257	0
Shared clipped image	443	172.447	204.790	1
Shared SMT	443	716.895	803.202	30
Actual-query SMT	442	2,210.719	2,314.252	442

Across both bit depths, independent/shared preparation takes 88.18/132.01 wall seconds; independent/shared sufficient checks take 5.01/1.84 seconds, and the shared joint-hull screen takes 113.79 seconds. The result check authenticates 81 source files, all condition identities, exact witness replay and 47,616 native scores in 1,488 witness batches, with no score mismatch or conflicting decisions. A fresh replay reproduces these checks and all reported outcomes. It authenticates every numerical dependency; 44 declared external input files were unavailable in this replay, so the original full input environment was not restored. Independent reaggregation reproduces every status and gain count. These checks do not independently reprove all invariant certificates.

Separate input-reconstruction costs. In the fresh application reconstruction, encoding takes 43.47 wall/108.65 process CPU seconds; BM25 and exact feature preparation takes 23.66/23.57 seconds. The single fit takes 0.65/0.64 seconds, and combined profile construction, native checks and serialization takes 690.48/687.67 seconds. These stage measurements exclude downloads, installation, process launch and subsequent independent audits. The runtime uses Python 3.11.7 rather than the original 3.11.13, with matching scientific package versions; byte-identical numerical payloads do not assert runtime binary identity. The reproducibility package records both environments.

I.3 A POST-HOC CASE ILLUSTRATING THE COMPLETE VERIFICATION STEP

The previously reported `PLAIN-741`, 8-bit, $k = 3$ instance was selected before the local five-method comparison. We preserve its full 32-candidate zone, original profiles, priorities and midpoint target. After 75 synthetic qualification entries, all five unchanged comparison adapters run three rotated, fresh-process repetitions. They retain the algorithm comparison's marginal and coarse-hull screens; the population study's global-range residual list is not supplied to them.

Marginal checks leave comparisons $(2, 3)$ and $(2, 5)$. The coarse-hull screen removes $(2, 3)$ but leaves $(2, 5)$, comparing `MED-4232` with `MED-4234`. Every method then enters its complete solver and returns `INVARIANT` in all three repetitions (Table 20). Thus the unchanged grouped implementation

can establish an invariant on real learned profiles beyond the original comparison adapters' marginal and coarse-hull controls. By domain inclusion, this certificate also holds throughout the actual owned quantizer query cell, for the fixed-pool top-3 membership target. The example is not unique to grouped, and Cartesian indexing is faster in this case.

Figure 8 illustrates why that coarse screen is inconclusive. The coordinate hull of the adverse cells intersects the feasible pair image, but the actual adverse cell products do not. All remaining candidates still constrain the displayed pair through the original 32-candidate zone. The stronger role-envelope check also certifies this condition (Appendix K.2).

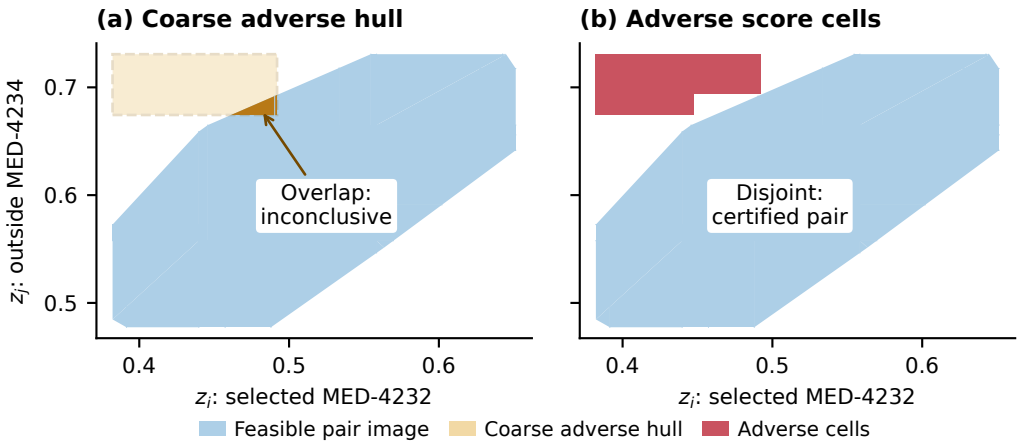


Figure 8: Post-hoc case `PLAIN-741`, 8-bit, $k = 3$, residual pair `MED-4232`/`MED-4234`. Blue shows the pair image of the full outer difference zone. Left: the coarse adverse coordinate hull overlaps it, so that sufficient screen cannot certify the pair; dark shading highlights the overlap. Right: the actual adverse score-cell products are disjoint, explaining the complete invariant result. The image is derived from 57 exact owned strata and eight RN64-aware adverse cell products, without sampling or convexifying their union. Fills display closures with floating-point coordinates; exact strict rows and endpoint ownership are retained in the accompanying data. This is an illustration of the existing certificate, not an additional proof or the exact latent-query image. The condition also passes the stronger role-envelope check.

Table 20: Median CPU seconds for one post-hoc mechanism case under the five-second complete adapter CPU budget. In the original comparison adapters, all returns enter the complete solver after the same residual comparison remains. These are not population coverage estimates or evidence of universal speed superiority.

Method	Valid invariant returns	Adapter CPU (s)	Excluding imports (s)
Grouped image	3/3	0.376644	0.330931
Clipped image	3/3	0.680962	0.636575
Cartesian index	3/3	0.334879	0.279215
Active tables	3/3	1.803107	1.747108
Z3 (QF-LRA)	3/3	2.693760	2.466422

The complete 90-entry comparison, including qualification, takes 46.12 CPU and 40.26 wall seconds. Separate saved-record checks authenticate every call and reconstruct both original/midpoint query points; all 64 scores agree with the native seven-feature LightGBM model. No application counterexample is returned. This audit does not independently reprove invariant certificates; the completed alternative solvers provide separate computational corroboration.

Interpretation. The application evaluates useful ranking behavior, shared-information value and complete-verification value as separate questions. The identity certificates in this original study concern fixed-pool top-k membership under the declared arithmetic, not internal ordering, NDCG,

full-corpus recall or text-realizable encoder perturbations. Appendix J.7 separately derives a sufficient offline relevance-bound certificate. Query vectors recovered within quantizer cells need not be outputs of actual texts. Additional certificates and a useful ranker do not by themselves establish that verification improves retrieval quality or deployment cost.

J RETRIEVAL ROBUSTNESS AND TASK-AWARE CERTIFICATES

These experiments separate ranker quality, shared-domain precision, complete-solver increments and computational cost. Original and later validation-selected models remain distinct. The original SciFact protocol was recorded before its encoding, fitting or evaluation, transferring a fixed recipe after the NFCorpus results were known. Later model-selection and diagnostic extensions were designed after the original TEST outcomes; they are follow-up analyses, not untouched confirmations of the final workflow. Conditions sharing a query are dependent.

J.1 SCIFACT RETRIEVAL WITH A FIXED MODEL

We use the public BEIR SciFact retrieval distribution (Thakur et al., 2021; Wadden et al., 2020): 5,183 documents, 809 TRAIN queries and all 300 TEST queries, with 919/339 positive TRAIN/TEST judgments. This TEST split is not the original SciFact hidden-label test set, and the task is document retrieval rather than scientific-claim verification. The original configuration is fixed before the new outcomes and has no hyperparameter search or DEV selection.

As in Appendix I, BM25 fixes 32 candidates using $k_1 = 1.2$, $b = 0.75$, lowercased alphanumeric tokens and document-ID tie priority. There is no relevance-based candidate insertion or query removal. A frozen `all-MiniLM-L6-v2` encoder produces 384-dimensional L2-normalized vectors. One LightGBM 4.6.0 LambdaRank fit uses all TRAIN groups, 100 boosting rounds, at most 15 leaves per tree, learning rate 0.05, seed 20260922 and one training thread. Its seven features are query min–max neural similarity, BM25 score, reciprocal BM25 rank, title and abstract query-token coverage, and log document and title lengths. Only the first feature varies during verification. The encoder, candidate pool, other six features and ranker parameters remain fixed.

Table 21: Macro means over all 300 SciFact TEST queries. Ideal gains and recall denominators use the full qrels, including documents outside the candidate pool. These ranker comparisons do not measure a benefit caused by verification.

Ordering in the fixed SciFact pool	NDCG@1	NDCG@3	NDCG@10
BM25	0.533333	0.619621	0.660464
Neural similarity, full precision	0.523333	0.614476	0.662112
Original learned ranker, full precision	0.610000	0.687318	0.719613
Original learned ranker, 8-bit midpoint	0.610000	0.688072	0.719461
Original learned ranker, 16-bit midpoint	0.610000	0.687318	0.719613

All orderings have Recall@32 $= 0.858722$. The pools contain 289 of the 339 relevant query–document pairs; 38 queries have no positively judged candidate. Unjudged documents receive grade zero in evaluation, without assuming that their human relevance is known. The fixed-ranker improvement over BM25 supplies application context, not a comparison with optimized retrieval systems or an isolated effect of min–max normalization.

The 8/16-bit, $k \in \{1, 3\}$ census contains 1,200 conditions. Independent uncertainty certifies 747 invariants, the original shared marginal, global-range and joint-hull checks certify 832, and shared complete verification certifies 833. All 86 strict coupling gains pair an independent-domain counterexample with a shared-domain invariant. The one invariant beyond that original check union, `985` at 8 bits and $k = 3$, is also certified by a stronger role-envelope check (Appendix K.2). The application interface completes all 1,200 independent SMT decisions (747 invariants, 453 counterexamples) and all shared clipped decisions (833 invariants, 367 counterexamples). Its shared SMT pipeline has 833 invariants, 324 counterexamples and 43 unresolved conditions. These counts include preliminary certificates, not only full-backend proofs.

The application interface has five-second preparation and backend wall limits and two-second sufficient-check limits. It is distinct from the five-second full-adapter CPU comparison below: their timing results are not pooled. Under the latter contract, Grouped, Cartesian and Clipped each complete all 1,200 conditions and agree on 833 invariants and 367 replayed raw-zone counterexamples. All 9,600 supplied profiles pass saved original/midpoint-score checks; native boundary probes provide additional model/profile parity checks rather than independent proofs of invariance. The complete application run takes 2,033.899 wall seconds.

Two TEST texts, IDs 1292 and 870, duplicate TRAIN texts after lowercasing and whitespace normalization. The official 300-query population remains primary. A saved-result sensitivity analysis excludes these two queries, retains 1,192 conditions and still has one certificate beyond the original three-check baseline; it neither refits nor reselects the model.

J.2 An external constrained-search comparison

We adapt native VERITAS search (Devos et al., 2021) to the identical supplied-profile problem. Two balanced ordinal-cell trees encode a compared pair of complete scalar profiles. Common integer score ranks preserve comparisons and ties; a documented native predicate hook rejects states whose cells fail the same exact extrema-role feasibility oracle used by Cartesian indexing. Native search, expansion and objective bounds remain in use. This is an adapted search-framework comparison, not an independent geometric proof family or an unmodified end-to-end tree-ensemble baseline.

Each condition has three fresh-process repetitions with rotated method order. The five-second adapter CPU budget includes deferred imports, preparation, common screening, ordinal encoding, native search, exact oracle calls and witness recovery/replay. Separate worker limits are ten CPU seconds, fifteen wall seconds and one GiB RSS. Offline dependency compilation is outside the adapter budget. Completion requires success in all three repetitions; censored conditions retain their denominators.

Table 22: Conditions completed in all three repetitions. The controlled and graph rows reuse the original 52 constructed inputs; adding an external comparator does not create 52 additional independent workloads.

Reused or separately fixed cohort	Grouped	Cartesian	VERITAS + oracle
Controlled scaling	36/36	22/36	21/36
Graph metric	16/16	16/16	16/16
NFCorpus timing subset	128/128	128/128	128/128

VERITAS was added after the original five-method study. Among the 21 controlled conditions completed by both Grouped and adapted VERITAS, the median within-condition VERITAS/Grouped CPU ratio is 5.547 including imports and 5.462 after excluding imports. Ratios use the same conditions and each condition's three-run median, not a ratio of medians from different successful sets.

The real subset fixes 32 NFCorpus TEST queries, yielding 128 conditions; all methods agree on 81 invariants and 47 counterexamples. Eighty conditions finish in common screening. On the 48 conditions requiring complete solving, the paired VERITAS/Grouped ratios are 1.461 with imports and 0.980 without them. For the 16 graph conditions they are 1.426 and 0.916. A separately reported post-hoc mechanism case also completes for all methods. The advantage is concentrated in the controlled family. Post-import costs are conditional on cold-budget completion, not a measured warm-process completion rate.

J.3 Bounded screening and the original-model full census

A bounded-screen policy limits the common coarse-hull screen to $2m$ extrema-role visits per condition. Only a proved infeasible check removes a pair; exhaustion leaves the pair unresolved for the unchanged complete backend. Grouped and Cartesian both use this policy under the same five-second adapter CPU contract. They now complete all 1,292 original-model NFCorpus conditions, with 824 invariants and 468 replayed raw-zone counterexamples. The two previously capped conditions, query `PLAIN-997` at 8 bits and $k = 1, 3$, are counterexamples, not new invariant certificates. The original incomplete phase measurements do not establish the precise location of the earlier caps. Clipped was not rerun with this policy, so its earlier completion count is not a matched-backend comparison with the new counts.

On the fixed 128-condition subset with three repetitions, the predeclared median old/new CPU ratios are 1.0001 for Grouped and 1.0015 for Cartesian. A post-outcome tail-cost diagnostic sums per-condition median CPU costs: 43.682 to 36.894 seconds for Grouped and 45.277 to 38.479 seconds for Cartesian. The approximately 15% reductions concern this descriptive sum, not typical-case speed

or online latency. The change benefits both backends and is not exclusively a Grouped geometric contribution.

J.4 Lifting original-model counterexamples to common queries

A feasible raw vector in the supplied outer difference zone need not arise from one query vector in its quantization cell. We therefore apply a fixed bounded finder to audited original-model raw counterexamples. Witnesses are chosen by a fixed backend priority, not by relevance loss or ease of recovery. Both an affine-gauge LP and an adverse-cell LP run in fixed order, each with a ten-second internal limit. Their returned candidates and two fixed contractions are checked in a fixed order, with at most six search candidates per condition. Search has a separate 60-second worker CPU and 90-second wall budget; it is not charged as a five-second verifier call.

Acceptance requires all 384 rational query coordinates to belong to the original owned quantization cells and re-encode to the same codes. Exact query–document dot products, exact shared min–max normalization followed by one nearest-even binary64 conversion, all 32 native model scores, and document-ID tie ranking must reproduce a top-k membership change. Independent replay recomputes the dot products with rational summation. The first valid search candidate is representative; qrels do not select it.

Table 23: Common-query counterexample conditions for the original models. NFCorpus uses 8/10/12/14/16 bits, SciFact 8/16 bits, and both use $k = 1, 3$. Prior denotes changes already witnessed by the original query; unknown denotes searched conditions without a confirmed common-query witness.

Dataset	Population	Searched	Prior	Added	Known	Unknown
NFCorpus	3,230	890	38	521	559	331
SciFact	1,200	367	10	176	186	181

The 559 witnesses span five precisions, not just the 1,292-condition census. No search caps occurred; unsuccessful searches remain unknown. These are existence results, not attack frequencies. Vectors can be nonunit rational points off the binary32 grid, without text realizations. They add no invariants and do not transfer to the selected models.

J.5 Precision-dependent certification and offline payloads

The original NFCorpus model is also evaluated at five precisions. Each precision has all 323 queries and $k = 1, 3$, giving 646 conditions. The independent column in Table 24 is a sufficient global-range check, not complete independent-domain verification. Its difference from shared-check counts alone does not prove strict domain separation. Targets depend on each precision's midpoint, so the table is not a fixed-target monotone robustness-radius curve. Here shared checks refers to the original marginal, global-range and joint-hull union, not the stronger cascade in Appendix K.2.

Table 24: Invariant certificates at each precision for the original NFCorpus model. The 10/12/14-bit conditions produce no certificate beyond the original marginal, global-range and joint-hull checks.

Bits	Conditions	Independent check	Shared checks	Shared + complete
8	646	102	185	186
10	646	269	364	364
12	646	470	535	535
14	646	584	615	615
16	646	633	638	638

An offline policy chooses the lowest certified precision from this fixed menu, retaining full precision when required by the recorded policy. Mean ideal vector payloads for independent/shared/shared-plus-complete checks are 536.77/487.73/487.73 bytes at $k = 1$ and 606.61/562.33/562.03 bytes at $k = 3$. Adding complete verification changes no $k = 1$ choice and changes one of 323 $k = 3$ choices, saving 96 additional bytes across the entire population relative to the original checks. The stronger

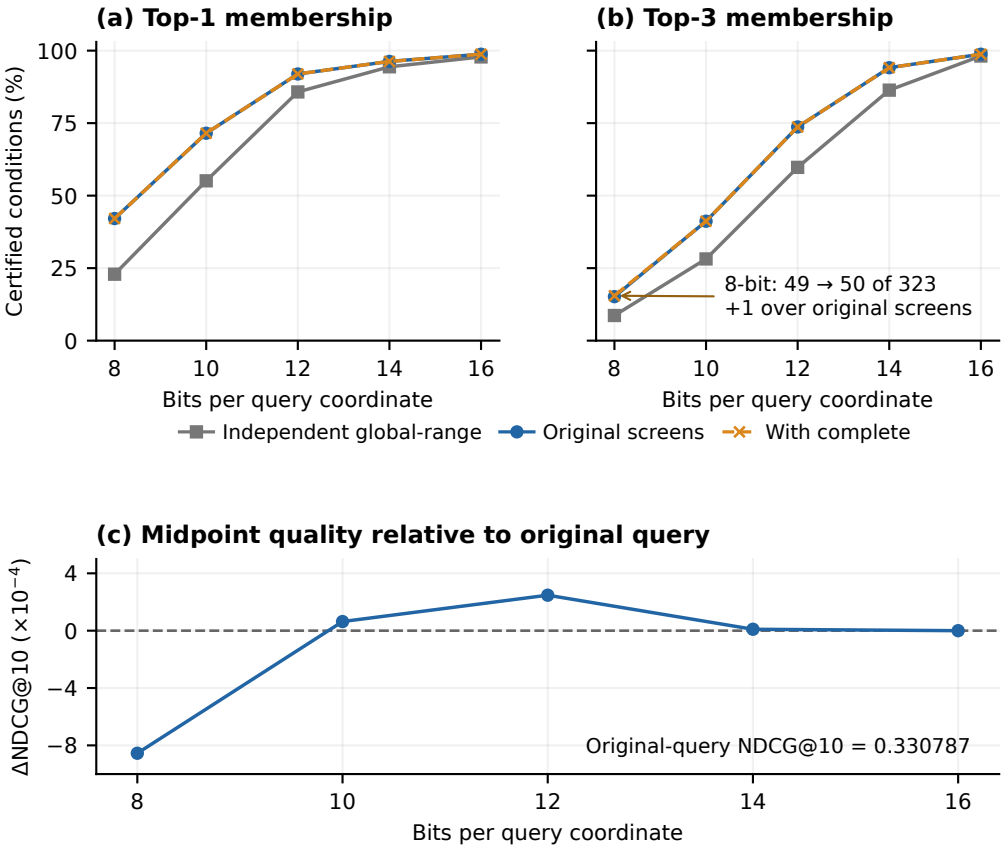


Figure 9: Certification and midpoint quality for the original frozen NFCorpus model, with all 323 queries at each k and precision. Gray is the cumulative independent global-range control; blue is the original union of shared marginal, global-range and hull checks; orange adds Grouped's certificates under that protocol. The one increment also passes the stronger role-envelope check. Panel (c) shows midpoint NDCG@10 minus original-query NDCG@10 in 10^{-4} units, not a verifier-caused improvement. Midpoint targets vary with precision, so this is not a fixed-target robustness-radius curve. Conditions sharing a query are dependent.

role-envelope check also certifies this same condition, so those 96 bytes do not require complete geometry beyond that check. Grouped and Cartesian select the same precisions. Most of the gain over the independent policy is already obtained by the original shared checks.

Corresponding NDCG@10 values are 0.330612/0.330767/0.330767 at $k = 1$ and 0.330901/0.330750/0.330750 at $k = 3$. These descriptive values establish neither equivalence nor improvement. The policy uses an original-query counterexample precheck. Payloads assume ideal bit packing and exclude protocol overhead, profile preparation and verification costs; no online deployment or latency reduction is measured.

J.6 Selected models as context for task-aware verification

Raw neural similarity, a TRAIN-fixed affine transform, and query min–max receive the same eight-configuration search and number of fits. NFCorpus uses its official DEV split. SciFact uses a text-grouped split within TRAIN, followed by refitting on all TRAIN. An additional four-view quantization augmentation fit is retained for each transform; each view is a separately weighted query group rather than an explicit cross-view consistency objective. All six selection decisions choose ordinary training. Particular recipes are selected only on these selection splits, although the study was designed after earlier TEST outcomes were known. Equal search and fit counts do not imply equal training CPU or selected model capacity.

Table 25: Clean NDCG@10 for the validation-selected recipes. Selected min–max models provide the fixed inputs to the quality-target study below; certificates from the original models are not reused for them.

Dataset	Original min–max	Selected raw	Selected affine	Selected min–max
NFCorpus	0.330787	0.332381	0.335232	0.332550
SciFact	0.719613	0.727493	0.721000	0.728092

The selected min–max models have 200 trees with at most 15 leaves and monotonicity constraints for NFCorpus, and 300 trees with at most 31 leaves for SciFact; the originals have 100 trees with at most 15 leaves. Paired query-bootstrap 95% intervals for selected minus original NDCG@10 are $[-0.001956, 0.005884]$ and $[-0.004568, 0.022145]$, respectively. Both include zero. Neither dataset establishes a min–max quality advantage over equally searched raw/affine models, and changes from the original models do not isolate the normalization transform. These comparisons are context for model provenance, not a new training contribution.

Fresh profiles and precision-specific targets are extracted before re-verification. With the bounded screen, both Grouped and Cartesian complete all 1,292 NFCorpus conditions (820 invariants, 472 raw counterexamples) and all 1,200 SciFact conditions (813 invariants, 387 raw counterexamples). The named sufficient-check union explains 819/820 and 813/813 invariants. It combines saved internal certificates with application-side shared-global/common-hull checks on every remaining complete-solver invariant; it is not an application-only full-population benchmark or a comparison against all possible cheap checks.

There are 370 and 355 changed midpoint target sets relative to the original models. Between-model certificate differences therefore do not measure solver accuracy. In fixed 128-condition subsets, each repeated three times, the median Cartesian/Grouped adapter CPU ratio is about 1.032 on each dataset. Separate profile extraction/preparation takes 778.17 and 1,882.45 wall seconds. These costs are not included in adapter time and cannot be omitted from an end-to-end interpretation. Native replay checks every saved counterexample, not every invariant proof independently. For SciFact, excluding the two known cross-split duplicate texts leaves 1,192 conditions with 809 invariants and 383 counterexamples for both methods; the 298-query primary-quality difference interval still includes zero. No refitting is performed for this sensitivity analysis.

J.7 Offline relevance-quality verification

Identity invariance can be unnecessarily restrictive when two relevant documents exchange first place. Relevance-aware certification that permits same-class neighbor changes is established (Wu et al., 2022b). We first apply the existing verifier through a fixed-anchor sufficient condition, then give a

separate complete relevance-set extension below. These task-specific reductions change the property being certified, not the model or predictions.

Let $U = \mathcal{P}(B)$ retain all m raw coordinates and the numerical contract of Theorem 3. For original scalar profile p_i, write $s_i(r) = p_i(\mathrm{RN64}(z_i(r)))$ and $i \succ_r j$ for the total score order including tie priorities. Fix the 8-bit midpoint winner a, its qrel grade γ, the good set $H = \{i : y_i \geq \gamma\}$ and bad set $L = \{i : y_i < \gamma\}$. Let Γ be the maximum grade in the full query qrels, not just the pool.

Corollary 15 (Fixed-anchor quality lower bound)**.** *If $a \succ_r b$ for every $r \in U$ and every $b \in L$, then the winner under the original score profiles has grade at least γ everywhere in U. For $\Gamma > 0$,*

$$\mathrm{NDCG@1} \;\geq\; \frac{2^{\gamma} - 1}{2^{\Gamma} - 1}. \tag{30}$$

The fixed-anchor premise is equivalent to anchor-identity invariance after replacing each other good profile by a constant strictly below the anchor's minimum score, while retaining the original raw domain and ties.

Proof. A bad winner would outrank the anchor, contradicting the premise. The gain function is increasing and its global ideal denominator is fixed, giving the bound. Masked good candidates are always below the anchor, whereas its comparisons with bad candidates are unchanged. Thus it wins the masked problem exactly when the fixed-anchor premise holds. □

The implementation uses the finite strict binary64 floor $C = \mathrm{nextafter}(\min_{\zeta} p_a(\zeta), -\infty)$ over the supplied anchor profile; a nonfinite floor would be unsupported. Every raw coordinate and difference constraint remains present: even a masked candidate can determine normalization extrema. A zero bound contributes zero without solver calls; an empty bad set gives a separately counted trivial positive certificate when applicable. If $\Gamma = 0$, the evaluation convention gives zero. Crucially, γ is the midpoint winner's grade: the bound need not preserve a higher grade attained by the full-precision original query.

Strong sufficient-check controls and incompleteness. The anchor baseline applies shared global-range/common-hull checks to the same masked problem. A set-aware baseline retains original profiles and tries to establish

$$\forall b \in L \; \exists h_b \in H \; \forall r \in U : h_b \succ_r b.$$

Different bad candidates may have different good witnesses. Proven pair sets from completed checks are united, equivalently intersecting their unresolved-pair sets; a capped check contributes no unproved pair removal. We compare complete fixed-anchor verification against the union of these two quality-certificate baselines with identical grades and qrels access.

Neither sufficient condition is complete for the actual quality property $\forall r \in U \; \exists h_r \in H \; \forall b \in L : h_r \succ_r b$. For example, two good profiles can alternate between scores $+1$ and -1 on opposite sides of a shared threshold, while a bad profile stays at zero. A good candidate wins everywhere, although each fixed good anchor loses somewhere. A transformed counterexample therefore refutes only the anchor premise; original-profile replay is needed to detect a bad winner.

Full-population result. The diagnostic fixes the selected models above and all 623 queries, 8 bits and $k = 1$, after prior TEST and identity results were known. It performs no fitting, additional model selection or query-vector search. There are 171/105 zero-bound queries and no empty-bad positive cases in NFCorpus/SciFact. All 623 workers and all 347 nontrivial complete-adapter calls finish, yielding the full-denominator results below.

Complete verification adds two NFCorpus certificates beyond the stated fixed-anchor and set-aware sufficient checks and none on SciFact, increasing the mean certified lower bounds by 0.006192 and zero. Both gain cases have completed these checks without caps and are certified in the complete-solver phase of this pipeline. Query `PLAIN-2081` has only one good candidate and already had an identity-invariance certificate: this is a previously known joint case. Query `PLAIN-2354` instead has an earlier identity counterexample but now receives a quality lower bound of one. Its midpoint winner `MED-4303` and alternative winner `MED-2152` both have grade one. The saved identity witness swaps them. Both cheap baselines leave the lower-grade `MED-4709` uncovered, whereas complete verification excludes every bad winner. Thus only one of the two increments is new relative to the

Table 26: Positive certificates and mean certified NDCG@1 lower bounds over all 323 NFCorpus and 300 SciFact queries. Uncertified and zero-bound queries contribute zero to the means. Identity invariance is a stronger-property reference, not a same-task baseline. No predictions are changed.

Method/reference	NFC count	Sci count	NFC bound	Sci bound
Prior identity invariance	82	156	0.239422	0.520000
Anchor sufficient checks	114	159	0.336429	0.530000
Set-aware sufficient checks	114	159	0.336429	0.530000
Sufficient-check union	114	159	0.336429	0.530000
Complete fixed anchor	116	159	0.342621	0.530000
Fixed-anchor study union	116	159	0.342621	0.530000

union of these cheap certificates and saved identity certificates. These two post-outcome explanatory cases exhaust the fixed-anchor study's gains; they are not extra independent samples.

Each dataset has 36 transformed counterexamples. Original-profile replay finds a below-threshold winner in 28 NFCorpus and 35 SciFact cases; the remaining eight and one saved witnesses do not reduce the original winning grade. This latter outcome does not prove uniform quality safety. These are raw-zone witnesses, not common-query or text-realizable attacks. Three NFCorpus set-aware common-hull calls hit their two-second wall limit: `PLAIN-1441`, `PLAIN-1453` and `PLAIN-2061`. Their unproved pairs remain unresolved. Other completed checks still certify the last query; the first two receive no cheap certificate. Full-worker completion must not be read as absence of cheap-stage caps.

Cost and interpretation. Each complete adapter has the same five-second CPU cap; sufficient-check preparation has five wall seconds and each global/hull check has two. The 623-query census takes 682.340 wall seconds and independent audit 49.946 seconds. Recorded CPU totals are 190.046 seconds for anchor checks, 104.522 for incremental set-aware checks, 79.771 for complete adapters and 158.568 for native witness replay. Set-aware work is reused without a second charge when its input equals the anchor input. Parent, child and stage timings are nested; they are not summed into an online latency claim. Earlier profile preparation is reused but remains a cost of producing the supplied inputs.

All 72 transformed witnesses undergo exact normalization/profile replay and native scoring by the unmasked selected model, with 2,304 candidate rows per replay pass; the independent audit repeats this check, without re-solving all invariants. The reproducibility package includes the implementation checks and complete condition-level results.

The guarantee uses fixed offline qrels, treating missing judgments as zero; it concerns these labels within a fixed pool, not unknown human relevance or candidate generation. The gain over identity alone mainly reflects a weaker property. The limited fixed-anchor two-versus-zero gain over matched quality checks demonstrates complete inference on these cases, without changing retrieval accuracy or establishing superiority over all sufficient methods.

Complete relevance-set verification. The good winner need not remain fixed. With the same H, L, γ and original score order, the exact quality property is

$$Q_\gamma : \quad \forall r \in U \; \exists h \in H \; \forall b \in L : \; h \succ_r b. \tag{31}$$

Its failure is equivalent to $\exists b \in L \, \exists r \in U \, \forall h \in H : \; b \succ_r h$: a bad winner beats every good candidate, and any bad candidate beating every good candidate forces a bad winner under the total order. All comparisons must hold at the same raw vector. Separate pairwise counterexamples therefore do not establish a quality failure. Property (31) implies the same lower bound (30) as the fixed-anchor condition.

Lemma 16 (Owned-box feasibility in a signed closed zone)**.** *Let B be a finite, consistent, closed rational difference matrix with zero diagonal and $\min B < 0$. Fix distinct minimum and maximum roles u, v. Let $I_i \subseteq [0,1]$ be nonempty rational intervals with endpoints ℓ_i, h_i and explicit ownership, with $I_u = \{0\}$ and $I_v = \{1\}$. There exists $r \in \mathcal{P}(B)$ with $z_i(r) \in I_i$ for every i if and only if some $\delta > 0$ satisfies*

$$\delta(\ell_i - h_j) \le B_{ij} \qquad \textit{for all } i, j, \tag{32}$$

where the comparison is strict if the lower endpoint of I_i or the upper endpoint of I_j is open.

Proof. Translate a feasible raw vector to minimum zero and take its range as δ; its coordinate bounds give necessity. For closed endpoints, set $r_i = \min_j(B_{ij} + \delta h_j)$. Equation (32) and the zero diagonal put $r_i \in [\delta\ell_i, \delta h_i]$; triangle closure gives $r_i - r_j \le B_{ij}$. The two fixed roles give minimum zero and range δ. For open endpoints, choose a positive rational δ from the feasible one-dimensional interval. Shrink each open raw endpoint inward by a common rational amount smaller than half every required strict slack. The closed construction then respects the original open bounds. This also preserves closed singleton feasibility. □

For each extrema pair, bad candidate and bad score cell, retain each good candidate's cells in which that bad candidate wins, including priority ties. Pull these machine cells back through the exact RN64 preimages, then branch over the resulting owned interval unions and apply Lemma 16. Other coordinates remain in $[0, 1]$; none is removed from the normalizer. Exhaustion certifies (31); a feasible branch supplies a full raw vector for original-profile and native-scorer replay. Each box test uses $O(m^2)$ rational inequalities, but the finite cell search can be exponential. Resource exhaustion returns unknown. This separate checker uses direct zone/box feasibility, rather than the Grouped pair-image backend.

Completion of the existing diagnostic. We retain the same selected models, all 623 queries, 8-bit domains and original midpoint-grade targets. Completing all nine previously unresolved quality cases gives two certificates and seven native-replayed counterexamples. The two certificates are NFCorpus queries `PLAIN-2102` and `PLAIN-2177`; their NDCG@1 remains one. This raises fixed-target coverage from 116 to 118 of 323 NFCorpus queries, against 114 for the matched sufficient-check union. SciFact stays at 159/300. The original fixed-anchor gains and these two relevance-set gains come from different verification procedures.

For the two new cases, exhaustive fixed-protector controls find 54 pair counterexamples: 17 good candidates against each of three unresolved bad candidates in the first query, and three good candidates against one bad candidate in the second. Original native replay confirms all pair failures while the full ranking retains quality one. Thus every relevant fixed protector fails somewhere, although a good winner exists everywhere. The 54 witnesses explain the quantifier gap; they are not independent queries.

The exact role/box census finishes all nine cases under 30 wall seconds and 200,000 branch nodes per query, using 37.617 seconds total wall time. A direct SMT encoding returns unknown on all nine at the same wall budget. Post-result runs with a 300-second budget corroborate both new certificates in 147.403 seconds total; no external SMT proof trace is checked. These longer runs are correctness corroboration, not an equal-budget comparison.

Combining the completion with prior certificates, sufficient pair checks and saved attaining witnesses closes the raw-zone quality frontier for all 623 queries. NFCorpus has 203 exact minima of zero, ten of $1/3$ and 110 of one; SciFact has 141 zeros and 159 ones. Their mean minima are $20/57$ and $53/100$. The 120 positive NFCorpus minima include two additional $1/3$ bounds below their midpoint targets, derived from previously completed sufficient checks; these are not new complete-checker gains. This is an extension on already exposed data, with no fitting, changed predictions or claim of exact worst-case quality over realizable query embeddings.

K IMPLEMENTATION CONFORMANCE AND PERSISTENT VERIFICATION

These follow-up studies use the original frozen retrieval models and their TEST queries. Their protocols and inputs were fixed before the new measurements, after the earlier outcomes were known. They test numerical conformance and computational cost, not untouched predictive generalization.

K.1 POINT CONFORMANCE OF TWO FLOATING-POINT PIPELINES

The reference computes exact dot products between stored binary32 document vectors and either the original stored query vector or its exact quantizer midpoint, followed by exact pool min–max normalization and one nearest-even binary64 conversion. All tested midpoint coordinates are exactly representable in both floating formats, so every path evaluates the same input point. We compare two explicitly specified alternatives on macOS arm64 with NumPy 1.26.4, its OpenBLAS 0.3.23.dev build and one numerical thread. A C-contiguous 32×384 document matrix multiplies the query vector; minimum, maximum, range subtraction, numerator subtraction and division then execute in the chosen dtype, either float32 or float64. A rounded zero range returns zeros. The resulting first feature is converted to binary64 for the same LightGBM 4.6.0 native predictor. The other six feature columns retain their frozen binary64 values; all paths use the same model and document-ID tie rule. Thus these are two upstream arithmetic paths, not two scorer precisions.

Every original query and the 8/10/12/14/16-bit midpoints are retained for NFCorpus; SciFact retains every original query and the 8/16-bit midpoints. The original representation is the stored binary32 query, not another quantization midpoint. This gives 1,938 and 900 query–representation points, respectively, or 90,816 candidate feature rows for each floating pipeline. Exact recomputation agrees with saved feature bits and native reference scores for every row. Identity-keyed records retain raw similarities, normalized features, native leaves, score bits and selected document IDs.

Table 27: Finite point-conformance checks against the exact reference. Each pipeline is evaluated on the same 90,816 candidate rows and 2,838 query–representation points. Feature differences count unequal binary64 encodings after any float32-to-float64 cast. Top-k counts concern set membership at the same point, not invariance across quantization.

Difference from reference	Float32 path	Float64 path
Normalized feature rows	85,110	48,811
Maximum absolute feature difference	3.79×10^{-7}	6.66×10^{-16}
Native leaf-path rows	0	0
Native score-bit rows	0	0
Top-1 / top-3 sets	0 / 0	0 / 0

Both floating paths preserve all evaluated native leaf decisions and score bits, as well as top-1/top-3 membership and top-3 order. Feature values nevertheless differ frequently (Table 27). The same frozen midpoint is compared across arithmetic paths; these checks do not claim that its ranking equals the original query's ranking.

Separate deterministic fixtures cover constant and nearly constant raw ranges, tied extrema, signed zero and quantizer endpoints. Near-zero ranges can collapse after floating conversion. A second fixture probes the predecessor, value and successor of each of 78 semantic thresholds on the first NFCorpus TEST fixed feature row. Casting these already-normalized feature probes to float32 changes native leaves and scores on 51 of 234 rows. These are direct threshold probes, not 234 retrieval queries or realizable query perturbations.

The point study therefore supports compatibility on the measured corpus, while the fixtures show why the numerical contract remains necessary. It establishes neither uniform equivalence on a quantization cell nor certification of an arbitrary floating preprocessing chain. Such a guarantee requires implementing the reference by construction or incorporating sound rounding enclosures for the entire upstream computation, including the near-zero-range case. All discrepancies and all tested points are retained; no model is retrained and no certificate count changes.

K.2 PERSISTENT PROCESSES AND FULL VERIFICATION COSTS

Workload and reuse boundary. The primary protocol retains all 323 NFCorpus and 300 SciFact TEST queries from the original frozen models. One request prepares and verifies four conditions for a query: 8/16-bit quantization and top-1/top-3 set identity. Thus the measured unit is a four-condition query bundle, not a single serving decision. Grouped, Cartesian, Clipped and adapted native VERITAS run in separate persistent workers, with one active request at a time. Each worker reuses imported modules, authenticated embedding/feature arrays and parsed models. Every request freshly extracts all 32 scalar profiles, checks native boundary behavior, constructs both exact difference zones, and checks reference and midpoint scores. Profiles are shared within the four conditions; no profiles, zones, decisions or witnesses are cached between query requests. Candidate retrieval, text encoding, corpus embedding and training precede this input boundary and are excluded.

All methods receive identical regenerated mathematical inputs, checked against the frozen case bank within the measured request. Each condition pays for input validation, closure and the same marginal and bounded common-hull checks, with at most $2m$ extrema-role visits in that hull check. Unresolved conditions proceed to the respective complete backend. Every returned counterexample undergoes exact raw-zone/profile replay and, in the primary study, native LightGBM score-bit and selected-set replay. This is a raw-zone witness check, not a common-query lifting test. VERITAS retains its exact Cartesian role-feasibility oracle; agreement does not make it an independent proof of the pair-image geometry.

Scheduling, budgets and accounting. Measurements use an Apple M1 Pro with 16 GiB RAM on macOS arm64. The protocol fixes three full rounds, a hash-based query order and rotated method order before measurement. The 2,492 conditions therefore give 29,904 scheduled method–condition attempts. Each condition has a five-second CPU and fifteen-second wall limit, including parity checks and witness replay. Fresh query preparation has a separate 60-second CPU/120-second wall limit. Workers have a 1 GiB RSS limit; native VERITAS search additionally has a 512 MiB memory limit. A supervisor enforces phase deadlines independently of Python signal handling. On a condition cap, the worker stops; a fresh worker reprepares the query and attempts only its remaining conditions. The capped condition is retained and is not retried. Restart and repeated-preparation costs are charged to that query, including a restart after its last condition. Preparation caps retain unreturned condition slots without retrying that query. Unexpected failures invalidate the campaign while preserving visited and unvisited denominators.

Query CPU begins before request parsing and ends before final response emission. CPU totals combine these timed worker bodies with interpreter startup to readiness; final response serialization and process shutdown are outside that CPU sum. Supervisor-observed request wall time includes transport and response handling. Supervisor CPU within requests is recorded separately. For either reported clock, stream cost is the sum of request costs plus startup costs, subtracting recovery startup already charged within requests. Preparation totals contain their component phases and are not added to those phases. The supervisor's reference-bank authentication, expected-hash construction, saved-label audit, result writing and post-run analysis lie outside this verifier-service boundary. Reported service wall time therefore differs from elapsed campaign wall time and does not measure complete retrieval latency. RSS is the maximum observed lifetime peak of a worker, not a per-request allocation. Four workers remain resident, and numerical libraries use one thread. No concurrent training or sustained numerical experiment is scheduled; host activity records document the shared workstation context.

Comparison units and secondary workloads. Completion requires a valid decision in every repetition. All-attempt costs retain caps and recovery. Request-latency quantiles pool every observed query-bundle wall time across the three rounds, including all recovery segments; unvisited requests have no latency and are counted separately. Paired speed ratios use within-query medians across repetitions on jointly complete queries, with completion denominators shown separately. Descriptive bootstrap intervals for ratios of summed query medians resample queries, not the four correlated conditions or technical repetitions. They characterize these already examined queries, not an untouched population. The real tasks have fixed $m = 32$, so they cannot establish a candidate-count crossover. Separate persistent runs retain the previously fixed 128-condition real subset and all 52 constructed cases. These secondary runs start from supplied zones/profiles and retain exact witness

replay; their timings do not include fresh query preparation or native-model replay. Constructed scaling and observed real-query profile variation have distinct interpretations.

Full-population results. Every method completes all 1,292 NFCorpus conditions and all 1,200 SciFact conditions in every round. All 29,904 scheduled decisions return without a cap, restart, missing slot or verdict conflict. The unique outcomes remain 824/833 invariants and 468/367 counterexamples; repetitions do not add independent certificates. Table 28 reports all observed request latencies, including fresh preparation and replay.

Table 28: Actual persistent query-bundle wall time in seconds. Each method has 969 NFCorpus and 900 SciFact requests (three rounds); every request verifies four conditions. Totals exclude the separately reported one-time startup. All methods complete every condition in every round.

	NFCorpus			SciFact		
Method	p50	p95	Total	p50	p95	Total
Grouped	2.256	2.623	2180.96	1.822	2.142	1668.36
Cartesian	2.259	2.703	2203.07	1.782	2.088	1636.47
Clipped	2.373	3.083	2359.94	1.882	2.534	1758.88
VERITAS	2.231	2.653	2159.41	1.766	2.083	1622.17

Grouped's total request wall time is $1.010\times$ that of the fastest NFCorpus comparator, VERITAS, and $1.028\times$ that of the same comparator on SciFact. The predeclared engineering target—at least 20% lower NFCorpus cost with equal completion and at most 5% higher SciFact cost—is not met. It also required the NFCorpus query-bootstrap interval for comparator/Grouped to lie above one. No numerical acceptance threshold was specified for p95 or RSS; they remain descriptive. Paired ratios of summed query medians, expressed as VERITAS/Grouped, are 0.9895 [0.9864, 0.9926] and 0.9723 [0.9704, 0.9741], with query-bootstrap 95% intervals. For Cartesian they are 1.0090 [1.0054, 1.0128] and 0.9814 [0.9796, 0.9833]; for Clipped, 1.0822 [1.0676, 1.0994] and 1.0550 [1.0467, 1.0638]. Ratios above one favor Grouped. These modest differences do not establish a general real-task speed advantage.

With startup charged once, combined service wall times are 3852.12, 3841.93, 4121.25 and 3784.00 seconds for Grouped, Cartesian, Clipped and VERITAS, respectively. Their startup wall times are 2.80, 2.38, 2.43 and 2.41 seconds; peak worker RSS is 219.05, 219.86, 220.03 and 222.66 MiB. For Grouped, fresh preparation accounts for 50.7% of measured request-body CPU on NFCorpus and 46.8% on SciFact. Native profile checks alone account for 41.6% and 37.9%; these are components of preparation, not additional costs. Per-condition zone preparation accounts for a further 23.7% and 29.0%. Grouped image construction plus profile comparison accounts for only 2.6% and 2.7%. Thus reducing that geometric component alone cannot remove the dominant costs at this workload boundary. The protocol includes fresh native boundary validation; a deployment that qualifies the adapter separately could have a different cost balance, which these timings do not measure.

The bounded runtime screens settle 821 NFCorpus and 832 SciFact conditions, leaving 471 and 368 for complete solving. The NFCorpus screen is weaker than the union of inexpensive checks used in Table 3 (823). Its extra solver calls do not constitute additional certificates beyond those stronger checks. The full-population result supports correctness agreement and identifies preparation costs; it does not supply a practical speedup, a candidate-count crossover or an untouched confirmation dataset.

Comparison with stronger sufficient checks. The three-check baseline in Table 3 does not exhaust the available sufficient checks. A targeted comparison adds compatible-order, endpoint-role and per-adverse-cell role-envelope filters to the common cascade, followed by exhaustive role-hull screening. The role-envelope filter is part of our implementation: it uses the same common-scale and extrema-role reduction as our role verifier, relaxing joint role compatibility through coordinatewise maxima without constructing the complete pair image. Sharing this sufficient filter across backends isolates the additional contribution of their complete solvers. All four backends certify NFCorpus `PLAIN-3211` at 8 bits and $k = 1$, NFCorpus `PLAIN-741` at 8 bits and $k = 3$, and SciFact `985` at 8 bits and $k = 3$ in the role-envelope phase, before complete solving. The latter two are the identity certificates beyond the original three-check baseline. This comparison retains the original models, domains,

profiles and targets and the 2,048-bit arithmetic limit. It is a targeted correctness comparison, not a population coverage or timing measurement; the original table and runtime totals retain their original screen policies. The analytical separation in Appendix B.14 still establishes incompleteness of its stated filters, independently of these empirical cases.

Persistent comparisons on supplied inputs. On the fixed 128-condition real subset, every method again completes all 81 invariants and 47 counterexamples in every round. Combined service wall times, including startup, are 107.37, 107.60, 125.72 and 105.86 seconds for Grouped, Cartesian, Clipped and VERITAS. Thus removing fresh query preparation does not reveal a general Grouped advantage on that subset.

The geometric advantage persists on the constructed cohort (Table 29). Grouped finishes all 52 cases in every round, whereas the other methods retain capped conditions. All returned decisions agree. Over all 624 scheduled calls, 528 finish and 96 reach a resource cap; none is failed, missing or unvisited. Restart cost is included even when the cap occurs on the worker's last requested condition. These results demonstrate a favorable constructed regime, not a realistic retrieval population or a universal speedup.

Table 29: Persistent verification of the 52 supplied constructed cases, three rounds. Completion requires success in all rounds. CPU and wall totals include every attempt, startup and recovery; request p95 wall includes caps and charged recovery. CPU denotes timed request bodies plus startup-to-ready process CPU. A request here contains one condition.

Method	Complete	Capped calls	CPU (s)	Wall (s)	p95 (s)	RSS (MiB)
Grouped	52/52	0	99.17	100.50	2.041	180.64
Cartesian	38/52	42	373.67	385.71	7.628	206.42
Clipped	48/52	12	238.40	242.78	7.370	208.38
VERITAS	38/52	42	428.60	440.94	7.676	210.52

Where the constructed crossover occurs. A descriptive reaggregation retains every original family and factorial setting. On the invariant profile-axis cases ($m = 12$, $k = 2$, four residual comparisons), Grouped/Cartesian/VERITAS request-wall totals are 3.291/1.214/1.350 seconds at $P = 16$ and 3.811/9.234/10.191 seconds at $P = 64$. Each total covers both original seeds and all three repetitions. At $P = 256$ and 1024, Grouped still completes both invariant cases, while Cartesian and VERITAS cap on both; Clipped completes all four profile settings. For the paired counterexamples, every method completes both cases through $P = 256$; at $P = 1024$, Grouped and Clipped complete both and the other methods cap on both. Here P is the generator's profile-size parameter: some candidates have constant profiles, so it is not the total number of cells. This is a constructed profile-size crossover, not an estimated universal threshold. The graph-metric settings show mixed timing wins and full completion. The residual-pair axis also changes m, k and total profile cells, so it does not isolate the effect of comparison count. Natural queries have fixed $m = 32$; their total profile-cell ranges are 1,132–1,568 for NFCorpus and 812–920 for SciFact. They therefore do not supply a natural candidate-count crossover.

K.3 A SUFFICIENT TRANSFER TO A SPECIFIED BINARY32 PIPELINE

We specify a numerical target separately from the once-rounded reference. Let $q \in Q \cap \mathbb{F}_{32}^{384}$, where Q is a query cell, and let $v_i \in \mathbb{F}_{32}^{384}$ be fixed document vectors; all entries are finite. There is no input cast or query renormalization. With RN_{32} denoting nearest-even rounding, compute, in coordinate order,

$$a_{i0} = 0, \qquad p_{i\ell} = \mathrm{RN}_{32}(v_{i\ell} q_\ell), \qquad a_{i\ell} = \mathrm{RN}_{32}(a_{i,\ell-1} + p_{i\ell}), \quad \ell = 1, \ldots, 384.$$

Products and sums are separately materialized, without fusion. Set $\widehat{r}_i = a_{i,384}$, $\mu = \min_i \widehat{r}_i$ and $\nu = \max_i \widehat{r}_i$, and compute

$$\widehat{d} = \mathrm{RN}_{32}(\nu - \mu), \qquad \widehat{n}_i = \mathrm{RN}_{32}(\widehat{r}_i - \mu), \qquad \widehat{z}_i = \mathrm{RN}_{32}(\widehat{n}_i / \widehat{d}).$$

The zero-range branch returns zeros; the gates below exclude that branch. Assume correctly rounded binary32 primitives, gradual underflow, finite intermediates and a stable rounding environment

throughout execution. Fix the scorer's other inputs, branch and accumulation order, and total tie priority π. Its complete original machine profile F_i must describe the actual scorer and treat both zero signs identically. The feature is widened exactly to binary64 before scoring.

An enlarged raw zone. Write $\overline{Q} = \prod_\ell [l_\ell, h_\ell]$ with rational endpoints. Require a finite, consistent, closed rational B satisfying the exact support checks $\sum_\ell \max\{(v_{i\ell} - v_{j\ell})l_\ell, (v_{i\ell} - v_{j\ell})h_\ell\} \le B_{ij}$ for every pair. Thus the exact dots $r_i = v_i^\top q$ belong to $\mathcal{P}(B)$. Set $u = 2^{-24}$, $\alpha = 2^{-150}$, $L = 768$, and evaluate exactly

$$A_i = \sum_\ell |v_{i\ell}| \max\{|l_\ell|, |h_\ell|\}, \qquad \epsilon_i = \frac{LuA_i + L\alpha}{1 - Lu}, \qquad B^+_{ij} = \begin{cases} 0 & i = j, \\ B_{ij} + \epsilon_i + \epsilon_j & i \neq j. \end{cases} \tag{33}$$

The standard rounding model (Rump, 2012), $\mathrm{RN}_{32}(x) = x(1+\delta) + \eta$, $|\delta| \le u$, $|\eta| \le \alpha$, expanded over the L operations, gives $|\widehat{r}_i - r_i| \le \epsilon_i$ uniformly on the cell. Indeed, coefficient errors are at most $Lu/(1 - Lu)$ and the at most L local absolute errors are amplified by at most $1/(1 - Lu)$. The conservative check $2\max_i(A_i + \epsilon_i) \le M_{32}$, where M_{32} is the largest finite binary32 number, rules out product, partial-sum and subtraction overflow. The endpoint additions preserve closure, and $\widehat{r} \in \mathcal{P}(B^+)$. Require $\rho = -\min_{ij} B^+_{ij} > 0$ and $(1 - u)\rho > \alpha$; otherwise abstain. These gates supply a uniform positive raw range.

Normalization and machine-score envelopes. Let $t_i = \mathcal{N}_i(\widehat{r})$ denote exact normalization. For its exact numerator $0 \le n \le d$, with $d \ge \rho$, monotonicity of rounding gives $0 \le \widehat{n} \le \widehat{d}$. The subtraction errors imply $|\widehat{n}/\widehat{d} - n/d| \le (2ud + 2\alpha)/((1 - u)d - \alpha)$; rounding the quotient adds at most $u + \alpha$. Consequently

$$|\widehat{z}_i - t_i| \le \sigma(\rho) := \frac{2u + 2\alpha/\rho}{1 - u - \alpha/\rho} + u + \alpha, \qquad R = \sigma(\rho) + 2^{-54}. \tag{34}$$

The second term in R covers $|\,\mathrm{RN}_{64}(t_i) - t_i|$ on $[0, 1]$. For machine inputs $x \in \mathbb{F}_{64} \cap [0, 1]$, define

$$H^-_i(x) = \min_{\substack{w \in \mathbb{F}_{64} \cap [0,1] \\ |w - x| \le R}} F_i(w), \qquad H^+_i(x) = \max_{\substack{w \in \mathbb{F}_{64} \cap [0,1] \\ |w - x| \le R}} F_i(w).$$

These finite profiles can be constructed exactly. An original machine cell with first and last values a, b contributes precisely for machine centers in $[a - R, b + R] \cap [0, 1]$: activate at the directed ceiling of $a - R$, and deactivate after the directed floor of $b + R$. The final source cell includes 1; the others end at the predecessor of the next start. All bounds and event cuts require exact or proved outward arithmetic.

Proposition 17 (Conditional binary32 certificate)**.** *Under the preceding premises and gates, fix a size-k target set S, $0 < k < m$, and define $G_i = H^-_i$ for $i \in S$, and $G_i = H^+_i$ otherwise. If exact verification over $\mathcal{P}(B^+)$ certifies S for $G_i(\mathrm{RN}_{64}(\mathcal{N}_i(r)))$ with priority π, then the specified binary32 pipeline has top-k set S for every $q \in Q \cap \mathbb{F}_{32}^{384}$.*

Proof. For every allowed query, $\widehat{r} \in \mathcal{P}(B^+)$ and $H^-_i(x_i) \le F_i(\widehat{z}_i) \le H^+_i(x_i)$, where $x_i = \mathrm{RN}_{64}(\mathcal{N}_i(\widehat{r}))$. Each certified selected lower score outranks every outsider upper score, including ties resolved by π. This is the score-bound ordering argument of Proposition 14, applied pointwise on the joint domain. □

The target remains fixed, for example the original reference midpoint set. Theorem 3 is complete for the supplied surrogate; this transfer is only sufficient for the numerical pipeline. A surrogate counterexample may combine unattainable rounding errors and yields abstention, as do failed premises or resource caps. Neither equivalence to a BLAS reduction nor empirical certificate retention follows from this result.

Implementation and bounded application check. An arm64 implementation uses separate NumPy 1.26.4 binary32 multiply/add calls and the stated normalization. Inputs, sources and binaries are pinned. A read-only same-thread floating-control-register guard rejects unsupported modes. Eight native fixtures match an exact rounding oracle, including 32-row vectorized and 33-row tail cases;

arithmetic and envelope fixtures also pass. These finite checks supplement the primitive assumptions and cannot exclude transient external mode changes.

The full 623-query, 2,492-condition schedule, original targets, sources and cost rule were frozen before a fixed identity-hash subset of two queries per dataset was evaluated. Each query freshly checks original native profiles, frozen input identities, exact cell inclusion and the preceding gates. All ten reference invariants among these 16 conditions retain conditional certificates, all through marginal checks. The six surrogate counterexamples yield abstention, not counterexamples for the actual numerical pipeline.

The four bundle wall times, including profile validation, enclosures and four decisions, are 8.380, 9.098, 7.806 and 7.559 seconds. All 16 slots return without caps or conflicts. The predeclared cost projection scales each dataset's largest bundle time to its full population and adds startup: 5,283.85 seconds exceeds the 4,320-second threshold, so no census is run under this initial protocol. Total qualification wall time is 37.88 seconds, including startup and supervision. This four-query check establishes neither population retention, an increment from complete joint solving, nor a practical speedup. A separate protocol with envelope reuse, additional sufficient controls and a larger registered budget is described in Appendix L.

L POPULATION RETENTION FOR THE SPECIFIED BINARY32 PIPELINE

This study retains all 623 previously examined NFCorpus/SciFact queries and 2,492 conditions, with unchanged candidate pools, models and reference midpoint targets. Retention uses all 1,657 reference invariants; coverage includes all conditions, including 835 reference counterexamples. A separate protocol permits 10,800 seconds for the census, with an 8,640-second launch projection threshold; the initial cost-gated study in Appendix K.3 is preserved.

Lower/upper score envelopes are constructed once per candidate and precision and reused across top-1/top-3 surrogate assembly. Qualification checks exact equivalence to the original builder on all 16 earlier conditions. The actual specified sequential binary32 pipeline is checked on the original query and both exact midpoints during query preparation; a preparation cap remains uncertified. These finite points check the reference targets and native profile interface; the uniform guarantee still depends on Proposition 17 and its numerical premises.

On each identical enlarged zone and one-sided score surrogate, we compare marginal, global-range and common-hull sufficient checks with complete Grouped verification. Their preparation and individual costs are charged. A capped check does not remove residual comparisons. Completed sufficient certificates exit early; otherwise the original complete engine is called, with its own bounded internal checks recorded separately. An increment beyond the stated checks requires completed negative control results and an invariant returned in the complete-solver phase. This three-check union does not include the broader role-envelope controls evaluated on the exact reference cases in Appendix K.2. Surrogate counterexamples give abstention, not attacks on the binary32 pipeline. Failed premises, caps and unvisited slots remain in the denominator. Neither these controls nor this numerical contract certify arbitrary BLAS or GPU reductions.

Full-population outcomes. The census finishes all 623 queries and 2,492 conditions in 5,738.17 seconds, within the new 10,800-second campaign budget. It retains 1,654 of the 1,657 reference invariants (99.82%), with coverage of 1,654/2,492 (66.37%) over all conditions (Table 30). There are 838 surrogate-counterexample abstentions: all 835 reference counterexamples and three reference invariants. These three losses are transfer abstentions, not demonstrated binary32 ranking failures. They are NFCorpus `PLAIN-23` and SciFact `75/514`, all at 16-bit, top-3; their original and corresponding midpoint executable checks still preserve the target. No condition has a final cap, premise rejection, failure or missing outcome, and no certificate contradicts a reference counterexample or a checked executable-pipeline point.

Table 30: Conditional certificates for the specified binary32 pipeline on the entire previously examined population. Retention is relative to reference invariants; all conditions remain in the coverage and abstention denominators.

Dataset	Queries	Conditions	Reference INV	Certified	Abstentions
NFCorpus	323	1,292	824	823	469
SciFact	300	1,200	833	831	369
Total	623	2,492	1,657	1,654	838

The 1,654 certificates comprise 1,631 marginal, 20 global-range, one combined three-check residual-ledger and two complete-geometry outcomes. The last two follow completed negative results from all three controls; they do not establish gains over the broader role-envelope cascade tested on exact reference inputs. They are the original NFCorpus `PLAIN-741` and SciFact `985`, both at 8-bit, top-3, rather than new query cases. Nineteen common-hull checks hit their resource limit and contribute no exclusion to the sufficient union. Every remaining complete call in these conditions returns a surrogate counterexample on a reference-counterexample condition. Thus the absence of final caps does not imply that every intermediate check finishes.

The executable checks cover all 1,869 query–representation points, with native score/profile parity and target agreement for every retained certificate. Timed request-body CPU totals 5,641.52 seconds; startup-to-ready CPU adds 2.73 seconds. These clocks exclude supervisor CPU, final response emission and shutdown, whereas the campaign wall clock includes supervision and output. The maximum worker RSS is 228.44 MiB. This single census is a numerical retention study, not a complete-backend speed comparison or an independent timing replicate.

M SCOPE BEYOND MIN–MAX NORMALIZATION

We record elementary reductions that clarify the scope of the verification framework. Affine post-processing preserves the min–max reduction. Softmax and nonnegative ℓ_1 normalization instead admit separate exact feasibility formulations. The latter formulations do not establish the linear pair-image complexity, construction bounds, or machine-arithmetic guarantees proved for min–max normalization. All statements below concern exact real arithmetic.

M.1 FIXED AFFINE POSTPROCESSING

Proposition 18 (Affine profile reduction)**.** *Let $z_i(r)$ be the min–max feature defined in Eq. (1), including its constant-input convention. Fix rational $a_i > 0$ and b_i, and set $y_i(r) = a_i z_i(r) + b_i$. Suppose G_i is a piecewise-constant score profile on $[b_i, b_i + a_i]$ with a finite rational interval partition, exact endpoint ownership, and exactly comparable scores. Then*

$$F_i(t) = G_i(a_i t + b_i), \qquad t \in [0, 1],$$

satisfies the supplied-profile contract of Theorem 3, and $G_i(y_i(r)) = F_i(z_i(r))$ for every feasible r. Under the remaining hypotheses of Theorem 3, the min–max verification results apply with the resulting profile sizes and unchanged tie priority.

Proof. The inverse image of an interval with endpoints l, u has rational endpoints $(l - b_i)/a_i$ and $(u - b_i)/a_i$. Since $a_i > 0$, endpoint inclusion and order are preserved. These inverse images partition $[0, 1]$ and preserve all profile scores. The score identity follows by substitution, so the selected set and every adverse comparison are unchanged. For a constant raw vector, $z_i = 0$ maps to $y_i = b_i$, as required by the composed convention. □

For example, normalization to $[-1, 1]$ corresponds to $a_i = 2$, $b_i = -1$. For a floating-point implementation, the profile must represent the entire composed evaluation, including any additional rounding operations. The real-arithmetic identity alone does not certify an arbitrary machine pipeline.

M.2 SOFTMAX UNDER DIFFERENCE CONSTRAINTS

Proposition 19 (Exact softmax image)**.** *Let $m \geq 2$, let $B \in \mathbb{R}^{m \times m}$ have finite entries, and suppose*

$$\mathcal{P}(B) = \{r \in \mathbb{R}^m : r_i - r_j \leq B_{ij} \text{ for all } i, j\}$$

is nonempty. For a fixed temperature $\tau > 0$, define

$$s_{\tau,i}(r) = \frac{\exp(r_i/\tau)}{\sum_{v=1}^{m} \exp(r_v/\tau)}.$$

Then

$$s_\tau(\mathcal{P}(B)) = \mathcal{S}_\tau(B) := \left\{ p \in \mathbb{R}^m_{\geq 0} : \sum_{i=1}^{m} p_i = 1, \quad p_i \leq e^{B_{ij}/\tau} p_j \text{ for all } i, j \right\}. \tag{35}$$

The set $\mathcal{S}_\tau(B)$ is a nonempty compact polytope contained in the strictly positive simplex. Every $p \in \mathcal{S}_\tau(B)$ has a raw preimage $r_i = \tau \log p_i$.

Proof. If $r \in \mathcal{P}(B)$ and $p = s_\tau(r)$, then $p_i > 0$, $\sum_i p_i = 1$, and

$$\frac{p_i}{p_j} = e^{(r_i - r_j)/\tau} \leq e^{B_{ij}/\tau}.$$

Thus $p \in \mathcal{S}_\tau(B)$.

Conversely, take $p \in \mathcal{S}_\tau(B)$. If $p_j = 0$ for any j, finiteness of all B_{ij} gives $p_i \leq e^{B_{ij}/\tau} p_j = 0$ for every i, contradicting $\sum_i p_i = 1$. Hence all coordinates are positive. Define $r_i = \tau \log p_i$. Taking logarithms yields

$$r_i - r_j = \tau \log(p_i/p_j) \leq B_{ij},$$

so $r \in \mathcal{P}(B)$. Moreover, $\exp(r_i/\tau) = p_i$ and $\sum_i p_i = 1$, implying $s_\tau(r) = p$. Finally, $\mathcal{S}_\tau(B)$ is the intersection of the compact simplex with finitely many closed linear halfspaces. Nonemptiness follows from the forward inclusion and nonemptiness of $\mathcal{P}(B)$. □

Consequence for score-cell feasibility. For fixed scalar score profiles, attainability of a cell pair I_i, I_j reduces exactly to feasibility of

$$p \in \mathcal{S}_\tau(B), \qquad p_i \in I_i, \qquad p_j \in I_j.$$

These are linear constraints with the original open or closed cell boundaries. A proposed top-k set is invariant exactly when no selected– outsider pair can occupy cells where the outsider wins, including the specified tie priority. A feasible full p recovers a raw witness through $r_i = \tau \log p_i$. This is an exact mathematical reduction, not a claim that the current min–max implementation solves these systems.

Scope of the reduction. Even for rational B and τ, coefficients $e^{B_{ij}/\tau}$ and recovered raw coordinates need not be rational. Exact comparisons, strict cell boundaries, and machine-level exponential evaluation therefore require separate treatment. No min–max construction bound or rational bit-complexity claim is transferred. Existing work develops convex bounds for softmax verification (Wei et al., 2023); the identity above concerns the exact image of a supplied difference-constraint zone, which need not equal the reachable logits of a neural network. Direct ranking of softmax probabilities preserves raw-logit order; the score-cell formulation concerns downstream candidate-specific profiles.

M.3 Nonnegative ℓ_1 normalization

Proposition 20 (Positive-scale feasibility formulation)**.** *Let $B \in \mathbb{R}^{m\times m}$ have finite entries, and define*

$$\mathcal{D}_B^+ = \{r \in \mathbb{R}^m_{\geq 0} : r_i - r_j \leq B_{ij} \text{ for all } i, j, \quad \sum_i r_i > 0\}.$$

For $r \in \mathcal{D}_B^+$, define $n_i(r) = r_i / \sum_v r_v$. Then

$$n(\mathcal{D}_B^+) = \left\{ p \in \mathbb{R}^m_{\geq 0} : \sum_i p_i = 1, \quad \exists t > 0 :\ p_i - p_j \leq t B_{ij} \text{ for all } i, j \right\}. \tag{36}$$

Any feasible (p, t) has a raw preimage $r = p/t$.

Proof. For a feasible raw vector, set $t = 1/\sum_v r_v$ and $p = tr$. Then $t > 0$, $p \geq 0$, $\sum_i p_i = 1$, and $p_i - p_j = t(r_i - r_j) \leq tB_{ij}$. Conversely, a feasible (p, t) gives $r = p/t \geq 0$, $\sum_i r_i = 1/t > 0$, and $r_i - r_j \leq B_{ij}$. It also gives $n(r) = p$. □

This is a specialization of classical fractional scaling (Charnes & Cooper, 1962). Cell-pair feasibility has a linear formulation in (p, t), with $t > 0$ retained as a strict constraint; replacing it by $t \geq 0$ can add unattainable limit points. The statement concerns only nonnegative inputs and exact division by their sum. It does not cover signed ℓ_1 normalization, ℓ_2 normalization, LayerNorm, RMSNorm, or an epsilon-clamped denominator, and supplies no counterpart of the sharp min–max pair-image complexity bound.