WHEN IS ONE IMPORTANCE RANKING ENOUGH? AUDITING RANKING REUSE ACROSS DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reusing one importance ranking across lesion-based functional diagnosis, concept correction, vulnerability screening, and specialization audits compresses distinct decision utilities into a common ordering. We evaluate this decision compression by separating compatibility, consequence, auditability, and authorization. Conflict structure does not determine sharing cost, and reliable measurement alone does not authorize reuse. Controlled redundancy separates single-unit effects from group responsibility; a construction shows that identical complete strict preferences can yield different minimum sharing costs even with fixed per-decision utility ranges. On CUB, same-feature common-linear and decision-specific linear rankers incur held-out mean normalized regrets of 0.4346 and 0.0785. A penalty persists under development-only stronger search and exhaustive optimization in an eight-concept test-utility control. A five-dataset non-CBM replication supports positive raw cost but reveals a denominator-sensitive normalized reversal. A prespecified blind AwA2 audit and separately frozen exploratory CelebA replication attain family-pair ordering accuracies of 0.956 and 0.929. Using 40% of audit observations, median accuracy retains over 90% of full-audit accuracy, but only 60.0–67.7% of subsampling runs meet the declared 0.85 criterion. In a 1,000-family LAD study, frozen direct-audit and cheap-proxy policies each select 100 of 500 held-out families at a fixed 20% reuse quota. At tolerance $\epsilon = 0.40$, both record 0/100 observed violations and overall normalized regret near 0.039, versus 0.3625 under always sharing. This improvement combines reduced reuse and selection; direct-audit superiority over the cheap proxy is not established. Separate prospective studies yield both certification and abstention under risk and coverage requirements. Ranking reuse therefore requires consequence evaluation and policy-specific validation, not confidence in an importance score alone.

1 INTRODUCTION

An importance ranking becomes an operational choice when it determines what to lesion for functional diagnosis, correct, or inspect. Pruning criteria estimate removal effects (Molchanov et al., 2017), while concept bottleneck models support interventions on predicted semantic concepts (Koh et al., 2020). Reusing one ordering across these actions imposes an additional constraint: distinct decision utilities share the same prioritization. We study when this *decision compression* is adequate. **Our central claim is that conflict structure, operational cost, and evidence for reuse are not interchangeable.**

Existing frameworks distinguish attribution targets and removal mechanisms (Covert et al., 2021), and explanation disagreement is well documented (Krishna et al., 2024). We instead compare shared and decision-specific orderings under fixed models, candidate units, actions, and budgets. Two inference gaps matter. Identical complete strict preferences can yield different minimum sharing costs even with fixed per-decision utility ranges. Likewise, an informative cost measurement does not establish a selected policy's population risk or coverage. Selective prediction and learn-then-test supply risk-control tools (Geifman & El-Yaniv, 2017; Angelopoulos et al., 2025); the question here is which reuse-induced consequence and policy their guarantees concern.

We formalize *ranking-reuse adequacy* through four questions: whether strict preferences admit a common ordering (compatibility), what set-level utility is lost by sharing (consequence), whether

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.

that loss is reliably measurable on independent data (auditability), and whether evidence supports a frozen reuse policy under declared risk and coverage requirements (authorization). Exact compatibility is not necessary for acceptable reuse: conflicting preferences can have small utility differences. We combine constructive separation with matched action-level comparisons, independent measurement, and policy-specific validation.

Controlled redundancy provides the functional motivation. Across 15 seed/family replicates, the mean single-unit loss effect falls from 0.1813 to 0.0113 as redundancy increases from one to 16 copies, while the full 16-copy group lesion has effect 0.2072. Single-unit effects are averaged over copies; both interventions use the same held-out evaluation split (Appendix A.1). Unit-level marginal effects therefore need not represent group responsibility. A score can describe an isolated intervention correctly without describing the effect of acting on a set; ranking reuse must therefore be evaluated at the intended action level. CUB shared-versus-specific comparisons then quantify the operational gap between shared and decision-specific ordering classes under fixed actions. In the same-feature linear comparator, common-linear and decision-specific rankers incur outer-model-held-out mean normalized regrets of 0.4346 and 0.0785. Development-only stronger search retains a penalty; an eight-concept test-utility oracle demonstrates a penalty at the restricted empirical optimum, not generalization of the selected orderings. Non-CBM results support positive raw sharing cost but expose a denominator-sensitive normalized reversal.

A prespecified blind AwA2 audit supports measurement repeatability, with separately frozen exploratory evidence from CelebA. LAD supplies a positive operational endpoint: frozen Direct and Cheap policies each select 100 of 500 outer-model-held-out families at a fixed 20% quota and observe no tolerance violations at $\epsilon = .40$, with overall regret near .039. A matched 20% random reference has conditional expected regret .0725 on the same fixed CONF panel, supporting useful selection beyond quota reduction; Direct superiority over Cheap remains unestablished. A prospective certificate and a CERT-stage abstention further distinguish measured outcomes from sufficient evidence for a declared policy.

Our contributions are a decision-conditioned formulation of ranking reuse; a fixed-range conflict–cost separation showing why order compatibility alone is insufficient; and controlled action-level comparisons, independent measurement, and policy evaluations that delimit when reuse is supported.

2 Related Work and Positioning

Action-specific importance and intervention reuse. Pruning methods estimate the loss caused by component removal (Molchanov et al., 2019), whereas concept-intervention work studies selection and granularity (Shin et al., 2023) or learns intervention policies (Espinosa Zarlenga et al., 2023). Concept bottleneck memory models store and reapply past interventions (Steinmann et al., 2024). Reusing such feedback differs from sharing a unit ordering across actions. Rather than compare alternative explanations or improve one intervention policy, we compare shared and decision-specific orderings under fixed models, candidate units, budgets, and utilities.

Consensus and operational consequence. Rank aggregation seeks consensus among input rankings (Dwork et al., 2001), while social-choice distortion studies utility loss from ordinal information (Procaccia & Rosenschein, 2006). Building on this ordinal–cardinal distinction, we study the additional set-level loss imposed by a shared ranking at declared action budgets, relative to decision-specific rankings. Our target is this operational penalty, not agreement with input orders alone.

Risk control and reuse authorization. SelectiveNet jointly learns prediction and rejection (Geifman & El-Yaniv, 2019); risk-controlling prediction sets, Learn then Test, and Conformal Risk Control provide guarantees for specified losses under distinct assumptions (Bates et al., 2021; Angelopoulos et al., 2025; 2024). We distinguish reliable measurement of reuse-induced loss from validating a frozen reuse policy. Our contribution is the decision-specific loss and validation target, rather than a new generic risk-control procedure. Appendix B expands these comparisons.

3 FRAMEWORK AND EVIDENCE DESIGN

3.1 DECISION-CONDITIONED REUSE

Protocol Π declares candidate units, actions, budgets, utility scales, comparator classes, and sampling units. Units belong to a model; a family F is a cross-model evaluation unit with study-specific aggregation. For measurement split T, let $\widehat{U}_{F,d,b,T}(S)$ be the measured utility of set S, oriented higher-is-better. For shared ordering π_s and $\boldsymbol{\pi} = (\pi_d)_d$, define the empirical sharing contrast

$$\widehat{C}_{\Pi,T}(F; \pi_s, \boldsymbol{\pi}) = \sum_{d,b} w_{d,b} \frac{\widehat{U}_{F,d,b,T}(S_b(\pi_d)) - \widehat{U}_{F,d,b,T}(S_b(\pi_s))}{a_{F,d,b,T}}. \tag{1}$$

Weights are nonnegative and sum to one; scales are positive and follow the original protocol. Raw, menu-regret, and anchor-normalized analyses remain distinct. Fitted-comparator contrasts may be negative and are not clipped. Matched tests fix models, candidates, actions, budgets, and evaluation data. For a fixed utility problem, let $V_{F,d}(\pi) = \sum_b w_{d,b} U_{F,d,b}(S_b(\pi))/a_{F,d,b}$. The optimal sharing penalty is

$$C^*_\Pi(F) = \sum_d \max_{\pi \in \mathcal{H}_d} V_{F,d}(\pi) - \max_{\pi \in \mathcal{H}_s} \sum_d V_{F,d}(\pi) \geq 0, \tag{2}$$

for nonempty finite ordering classes with $\mathcal{H}_s \subseteq \bigcap_d \mathcal{H}_d$. Each decision uses one ordering across budgets. Exact optima and held-out fitted contrasts are different objects (Appendix C.1).

3.2 COMPATIBILITY DOES NOT DETERMINE CONSEQUENCE

Let G_F unite the declared strict unit preferences $\succ_{F,d}$. **Proposition 1 (Exact compatibility).** *An unrestricted ordering preserves these preferences exactly if and only if G_F is acyclic.* This does not guarantee representability in a restricted score class. Budget-dependent comparisons of whole ranking rules need not change their unit orders; explicitly budget-indexed unit preferences, when declared, are included in the union.

Proposition 2 (Fixed-range conflict–cost separation). *Identical complete strict preferences and identical per-decision utility ranges can yield different optimal sharing penalties.* For top-1 choices a, b, c, take $U_1 = (1, 1-\eta, 0)$ and $U_2 = (1-\eta, 1, 0)$, $0 < \eta < 1$. Preferences and full-menu ranges remain fixed, but $C^*_\Pi = \eta/2$ under equal decision weights and range normalization. Appendix C.3 proves the result; arbitrary normalization invariance is not claimed.

3.3 INDEPENDENT MEASUREMENT AND EVIDENCE ROLES

Auditability compares $\widehat{C}_{\Pi,T}$ on disjoint audit and confirmation observations under frozen comparators and normalization rules. AwA2/CelebA use the identity prediction $\widetilde{C}_{\text{confirm}} = \widehat{C}_{\Pi,\text{audit}}$, evaluated against $\widehat{C}_{\Pi,\text{confirm}}$. This tests measurement repeatability, not automatic cross-domain calibration. Direct and cheap scores predict the consequence; neither defines it. Inference respects generating model/seed or family clusters, not unit pairs or audit replicates. Prospective tests and retrospective controls retain distinct evidential roles.

3.4 EVIDENCE FOR A REUSE POLICY

Authorization is policy-form specific. For a frozen per-family map g, with $g(F) = 1$ denoting SHARE and 0 SPECIALIZE, define

$$R_\epsilon(g) = \Pr\Big(\widehat{C}_{\Pi,\text{confirm}}(F) > \epsilon \mid g(F) = 1\Big), \quad Q(g) = \Pr(g(F) = 1). \tag{3}$$

Probabilities cover the declared generator and measurement randomness conditional on its data environment; R_ϵ requires $Q > 0$. The target is a measured contrast, not latent expected utility or C^*_Π. For a policy frozen on FIT before independent, identically distributed CERT families, simultaneous bounds support authorization under declared risk α, coverage κ, and confidence failure probability δ when

$$U_R(E, M; \delta_R) \leq \alpha, \qquad L_Q(M, N; \delta_Q) \geq \kappa, \qquad \delta_R + \delta_Q \leq \delta. \tag{4}$$

Here N, M, E count CERT families, selections, and violations. Exact bounds and assumptions are in Appendix C.5 (Bates et al., 2021; Angelopoulos et al., 2025). CONF remains unopened until applicable gates pass. LAD's fixed-quota batch results require separate validation; the per-family argument does not certify their population risk or coverage.

4 When Sharing an Ordering Is Costly

Section 3 separates preference compatibility from utility loss. We examine functional effects and action-specific ranking utility, then quantify the sharing constraint. Fitted contrasts retain their signed, study-specific scales; reduced test-utility optima are reported separately.

4.1 Functional motivation and decision-specific usefulness

Controlled redundancy separates single-unit from group effects: across 15 seed/family replicates, increasing copies from one to 16 reduces the mean per-copy lesion effect from .1813 to .0113, whereas lesioning the full 16-copy group yields .2072 on the same held-out split. This is a change in intervention scope, not an equal-budget comparison.

CUB supplies a separate model comparison. The Full adaptive-minus-Standard CBM lesion-gap contrast is $+.01033\,[.00867, .01199]$ under held-out ranked lesion evaluation and $+.03478\,[.02818, .04138]$ under the current-test protocol. These quantify a model difference within each protocol, not a shared-versus-specific ranking penalty or a reversal between ranking rules.

Direct evidence of a decision reversal comes from SUN (Figure 1). Contribution-minus-Intervention regret is $-.60479\,[-.74787, -.40273]$ for lesion and $+.15097\,[.03299, .26629]$ for correction. Lower regret favors Contribution for the former and Intervention value for the latter. Thus ranking usefulness must be evaluated for the declared action. SUN uses its own frozen action operators and six-descriptor normalization contract, specified in Appendix D.2; these numerical regrets are therefore not placed on the CUB utility scale. The complete SUN audit and the separate CUB protocol results are retained in Appendix D; these summaries do not establish a between-budget interaction.

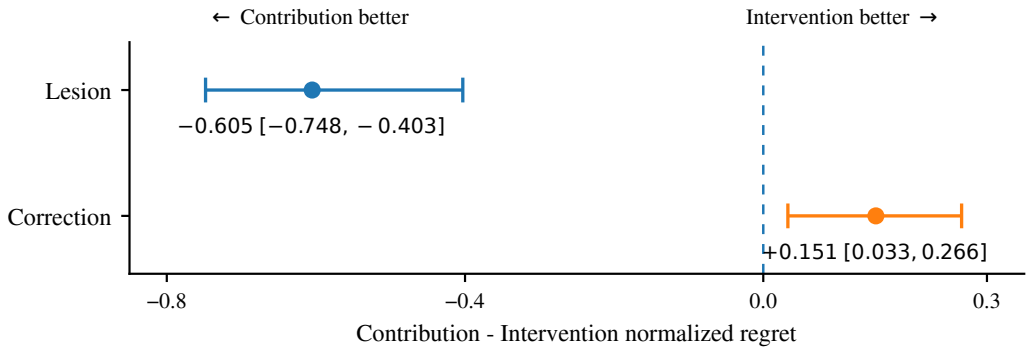


Figure 1: **A ranking's advantage reverses with the decision on SUN.** Contribution-minus-Intervention normalized regret with the reported 95% intervals. Negative values favor Contribution; positive values favor Intervention value. This is a within-decision ranking comparison, distinct from the CUB Full adaptive–Standard CBM model contrast.

4.2 Development-fitted comparisons and a reduced exact oracle

CUB evaluates top-k concept sets under lesion, ground-truth correction, vulnerability screening, and specialization, with lesion treated as a functional-damage diagnostic. Exact CUB utilities and normalization are specified in Appendix D.3; SUN uses a distinct contract in Appendix D.2.

Table 1: **Shared-ordering penalties under distinct CUB protocols.** Positive Δ favors decision-specific rankings. Development-fitted rows use different normalization menus and are not cross-row comparable. Their intervals are paired percentile-bootstrap summaries over the 15 recorded outer-heldout contrasts, conditional on the realized model panel. The oracle row optimizes and scores test utility only in a validation-selected $K = 8$ problem.

CUB comparison	Shared	Specific	Δ 95% bootstrap interval
Development-fit linear; outer-model held-out $\downarrow$	.4346	.0785	+.3562 [.3029,.4074]
Development-fit stronger search; outer-model held-out $\downarrow$	.4742	.2514	+.2229 [.1689,.2765]
Test-utility oracle, $K = 8$; score $\uparrow$	.6830	.9928	+.3098 [.2907,.3289]

Action contract. The CUB common-linear comparison uses 15 retrospective leave-one-seed/model-out folds. In each fold, validation-side functional features from the 14 development models are paired with their official-test single-unit objective targets for ridge fitting. The held-out model contributes only validation-derived features to instantiate the frozen ranking; its official-test split is reserved for evaluating the resulting action utilities. Shared and decision-specific scores use the same five functional features, development folds, target transformation, and nominal ridge parameter $\alpha_{\text{ridge}} = 1$. Because the common fit pools four equally sized objective blocks under the summed-loss implementation, equal nominal α does not exactly match per-response regularization strength. We therefore treat the linear experiment as a same-feature comparator rather than a pure coefficient-sharing intervention. Under the seven-policy range-normalized menu, mean regret is .4346 versus .0785: a paired contrast of $+.3562$ with a 95% conditional fold-bootstrap interval [.3029,.4074]. The worst-decision contrast is $+.4355$ with the corresponding interval [.3400,.5307] (Table 1).

The development-only stronger-search control fits Kemeny-style orderings to development preferences using six starts and 5,000 proposals per start. Orders are frozen before held-out validation/test evaluation. Its expanded nine-policy menu yields $+.2229\,[.1689, .2765]$ mean sharing contrast. This tests another optimization route; it establishes neither a global set-level optimum nor comparability of effect magnitudes across the two menus.

The reduced exact control has a different role. Validation selects eight concepts for each of 15 seeds; official-test utilities then intentionally optimize and score all $8! = 40{,}320$ shared and objective-specific orderings at $k = 1, 2, 3$. The specific-minus-shared score difference is $+.3098\,[.2907, .3289]$. Enumeration rules out search error within this reduced empirical problem, not out-of-sample error of its selected orders or suboptimality in the full CUB space. Reported paired seed/model bootstraps use 10,000 draws (Appendix D.3). Separately, joint accuracy/AUC matching supports none of five stronger cross-architecture divergence claims; those claims are not needed for the within-model sharing comparison.

4.3 Set-level alignment and the scale of consequence

SUN's target-domain outer-heldout linear comparison yields a positive mean contrast, $+.1300\,[.0307, .2331]$, but an unresolved worst-decision contrast, $-.0092\,[-.1545, .1379]$; its study-level positive criterion is not met. Near-order results remain separate. FSD50K's retrospective set-level alignment uses 32 frozen models from four families and yields mean $+.24745\,[.22304, .27377]$ and worst $+.53616\,[.46288, .61949]$ contrasts. Family-then-seed resampling preserves that hierarchy. The aligned result does not overwrite the historical prospective `NOT_SUPPORTED` verdict (Appendix D.4).

The five-dataset non-CBM replication contains 15 dataset-by-regularization families and 45 models. Separate optimization, audit, and confirmation splits yield a positive raw contrast, $.008730\,[.003253, .014207]$, under 2,000 dataset-cluster bootstrap draws. The rowwise normalized contrast instead averages -45.1492 (95% CI approximately $[-94.2, -.00224]$). The normalization audit identifies the scale mechanism: three of 135 confirmation rows hit the 10^{-6} denominator floor, amplifying negative contrasts to values including -3296.3 and -2792.2. Positive denominators pre-

Table 2: **Measurement agreement across distinct evidence roles. AwA2 is prespecified and blind; CelebA is a separately frozen post-hoc exploratory replication. RMSE values are compared within panels only.**

Panel	Families/ models	A_{pair} [95% CI]	RMSE audit / constant
AwA2	10 / 50	.956 [.842,1.000]	.0230 / .1514
CelebA	8 / 32	.929 [.750,1.000]	.0238 / .0893

serve each row's sign but reweight mixed-sign contrasts when normalization precedes aggregation. Every reported leave-one-dataset-out point estimate remains positive on the raw scale and negative after normalization (Appendix E.1). These diagnostics explain the aggregate reversal without changing either endpoint: positive raw cost does not establish a scale-invariant penalty. Section 5 examines measurement repeatability under the declared scale.

5 Measuring Reuse Consequences

Section 4 quantifies sharing penalties, not their predictability. We test repeatability of the same signed contrast, separately from descriptor utility and policy validation.

5.1 From intervention utility to direct measurement

Intervention utility does not itself validate a sharing-cost predictor. In the supporting SUN head-compatibility study, soft correction of $k = 20$ concepts under the predicted-concept head yields gains of .02806 for Contribution, .02896 for Concept AUC, .01384 for Usage, and .02073 for Random. Contribution and Concept AUC are numerically close; Usage has a lower point estimate. Contribution Gain@20 is approximately .033 and .034 under mixed- and ground-truth-concept heads, respectively (Appendix E.2). These are head-conditional intervention outcomes, not audit-to-confirm errors or paired tests of ranking superiority. They motivate directly testing the sharing contrast, not assuming a descriptor predicts it.

5.2 Measurement repeatability and evidence roles

A prespecified blind AwA2 audit tests the identity prediction $\widetilde{C}_{\text{confirm}} = \widehat{C}_{\Pi,\text{audit}}$ against $\widehat{C}_{\Pi,\text{confirm}}$. CelebA provides a separately frozen, post-hoc exploratory replication of the same measurement pattern, not a second confirmatory test. Rankings, models, anchors, and predictions are held fixed, without confirmation-time recalibration. Table 2 pairs ordering agreement with error against a frozen constant. Inference respects ten AwA2 families and eight CelebA families, rather than pairwise comparisons, within-family seeds, or repeated audit subsamples. AwA2 passes measurement but not the separate complexity-savings criterion: repeatability implies neither comparator optimality nor threshold transport (Appendix E.3).

5.3 Measurement with fewer audit observations

The six-fraction study (5/10/20/40/60/100%) varies only audit observations, using 1,000 fixed-seed subsamples per family at each reduced fraction. Models, rankings, anchors, and confirmation data remain fixed. Median A_{pair} increases across the declared grid in both panels (Figure 2). At 40%, medians .8889 and .8571 retain approximately 93% and 92% of the respective full-audit values .9556 and .9286. Yet only 67.7% and 60.0% of subsampling runs attain $A_{\text{pair}} \geq .85$; even at 60% audit these frequencies are 82.9% and 87.5%. Thus 40% is the first declared fraction meeting both median criteria—accuracy at least .85 and over 90% retention—not a reliable-run guarantee. Percentile bands describe conditional subsampling variation, not population confidence. Reduced sample requirements do not establish end-to-end compute savings (Appendix E.4).

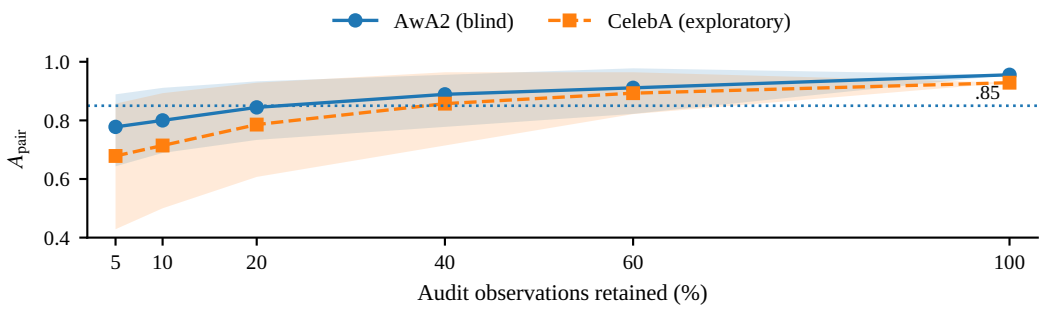


Figure 2: **Median accuracy improves before run-level reliability is uniform.** Six declared fractions: points are medians and bands join empirical 5th–95th subsampling percentiles, conditional on fixed panels, not population confidence bounds. At 40%, only 67.7%/60.0% of AwA2/CelebA runs meet the .85 ordering criterion. Full-audit bands collapse by design. AwA2 is prespecified and blind; CelebA is exploratory.

5.4 Scale sensitivity and incremental decision value

Repeatable measurement does not make normalization innocuous. The non-CBM consequence study retains a positive pooled raw sharing contrast while rowwise normalization yields a negative aggregate, with Appendix E.1 tracing the reversal to rowwise scale reweighting. A separate packaged E29 V4 audit provides unit-row ordering diagnostics, which we report separately because its aggregation unit differs from the family-level AwA2/CelebA studies (Appendix E.1).

The same secondary LAD analysis compares frozen Direct and Cheap policies at a 20% quota. Their confirmation-regret difference is $-.000219$ $[-.005769, .005389]$, establishing neither superiority nor equivalence. Strong post-hoc structure-only association is likewise not a prospective policy result (Appendix E.6). The consequence remains the validation target, not a privileged predictor. Section 6 tests whether policy-specific evidence is sufficient to support reuse.

6 Selective Reuse and Authorization Boundaries

Sections 4–5 establish operational penalties and their measurability. Neither result alone licenses a reuse policy. We now distinguish held-out decision value from the policy-specific authorization contract in Section 3.4: a violation concerns the measured sharing contrast, not latent utility or the optimal penalty C^*_{Π}.

6.1 Held-out selective reuse under a fixed quota

LAD V4B contains 1,000 families and 3,000 models, with 200 FIT, 300 protocol-CERT, and 500 CONF families. Direct and Cheap use a fixed 20% batch quota at tolerance $\epsilon = .40$. Both select 60 protocol-CERT families without observed violations; subsequent held-out CONF yields 0/100 for each policy (Table 3). Always Shared instead incurs 217/500 violations. Direct's overall normalized regret is .039011 versus .362531, a reduction of .323520 with reported paired family-bootstrap 95% interval $[.307838, .338877]$.

Writing $c_i = \widehat{C}_{\Pi,\text{confirm}}(F_i)$ and g_i for the batch's SHARE mask, the reported overall regret is

$$J(\mathbf{g}) = \frac{1}{N}\sum_i g_i c_i = \frac{M}{N}\,\bar{c}_{\text{SHARE}}, \qquad M > 0. \tag{5}$$

Specialization contributes zero additional sharing contrast by construction; reducing reuse to 20% and choosing its recipients both affect overall regret. The full reduction is therefore not an isolated selector gain or a measured compute saving. At the same quota, Direct minus Cheap regret

Table 3: **LAD held-out decision value at $\epsilon = .40$.** Regret averages all 500 CONF families. Direct and Cheap use a fixed 20% quota. The random row is a matched-quota reference on the fixed CONF panel, not an observed policy. †Values are conditional expectations under uniform sampling without replacement of 100 of 500 families; "–" denotes an undefined SHARE-conditional violation rate when $M = 0$.

Policy	SHARE M/N	Violations among SHARE	Overall sharing regret ↓
Always Specialized	0/500	–	0
Always Shared	500/500	217/500	.362531
Random 20% reference†	100/500	43.4 expected	.072506 expected
Direct	100/500	0/100	.039011
Cheap	100/500	0/100	.039229

is $-.000219\,[-.005769, .005389]$: neither superiority nor equivalence is established. The shared consequence remains the validation target, irrespective of the predictor. Moreover, a fixed batch quota is not an estimated population coverage guarantee; Section 3.4's per-family argument does not automatically certify LAD's batch selection (Appendix F.2).

Matched-quota random reference. The Always-Shared-to-Direct comparison combines lower reuse frequency with nonrandom family selection. On the fixed 500-family CONF panel, a uniform 20% quota selects each family with probability .2, giving

$$\mathbb{E}[J_{\mathrm{random},20\%}] = 0.2J_{\mathrm{all}} = 0.0725061.$$

Direct attains $J = .0390108$, which is 0.0334954 lower, or 46.2% below this matched-quota random expectation.

Likewise, with 217 violating families under Always Shared, a random 100-family subset has expected violation count 43.4, whereas Direct observes $0/100$. These are fixed-panel randomization references, not prospective population-risk certificates; they isolate evidence for useful selection beyond quota reduction without establishing Direct superiority over Cheap.

6.2 Low observed error is not sufficient evidence

Three boundary studies separate quantities that a favorable audit score can obscure. Transporting the pilot's absolute threshold to AwA2 specializes all ten families: zero sharing regret is obtained with zero reuse, not a valid conditional-risk claim on an empty SHARE set. The V6 follow-up selects 46/80 families with no observed violations, but the reported one-sided 95% exact upper bound is .0630493, exceeding the 5% risk target. This follow-up has no separate FIT/CERT/CONF authorization pipeline.

CEBaB separates a predeclared stop from later exploration. V9 stops at its CERT gate after 4/60 selected-family violations. V10 is a post-hoc $\alpha = .15$ analysis, not a repaired prospective success: Direct records 1/95 violations among 500 confirmation families, but reuse is only 19%, below the 20% requirement. Its final verdict remains unsupported. The subsequent result cannot reverse V9's stop; useful ordering and low observed error do not jointly establish the original operating claim (Appendix F.3).

6.3 Prospective authorization and the meaning of a stop

Gesture and Letter evaluate policies fixed using FIT before independent CERT, retaining $\epsilon = .10$, $\alpha = .05$, $\kappa = .20$, and $\delta = .05$. Risk and coverage must both pass: specialization alone cannot satisfy a nontrivial-reuse requirement. Their reported joint bounds lead to opposite decisions (Table 4). Gesture selects all 120 CERT families with no violations; CONF subsequently also yields 0/120 violations with full reuse. This supports the declared certificate, not zero population risk or an advantage over Always Shared. Letter's FIT-selected policy instead incurs 24/77 violations on 400 CERT families, yielding observed risk 31.17% and coverage 19.25%. Both bounds fail their requirements, so the protocol abstains and leaves CONF untouched. The finding concerns that selected policy, not the nonexistence of every feasible policy (Appendix F.4).

Table 4: **Authorization bounds for frozen per-family policies.** N, M, and E denote CERT families, selections, and tolerance violations; both risk and coverage inequalities must hold. Gesture is authorized and subsequently confirms 0/120 violations, whereas Letter abstains and leaves CONF unopened.

CERT panel	N	M	E	$U_R \leq .05$	$L_Q \geq .20$	Decision
Gesture	120	120	0	.030273	.969727	Authorize
Letter	400	77	24	.427428	.155011	Abstain

7 Discussion and Conclusion

Reuse is a decision claim, not a score property. Our results separate local importance, set-level consequence, and evidence for reuse. Redundancy distinguishes single-unit effects from group responsibility, while the fixed-range construction shows that preference structure does not determine sharing cost. Ranking reuse should therefore be evaluated through the actions induced by the ranking, not agreement alone. Exact compatibility is unnecessary, and structural features may still predict cost within a particular population.

An operational reporting principle. A reuse claim should specify its population, actions, budgets, comparator class, utility scale, tolerance, and policy form; compare matched shared and decision-specific actions; measure their consequence independently; and validate the frozen policy under the relevant risk and reuse requirements. LAD illustrates this separation: cheap predictors can support useful selective reuse, while Direct auditing shows neither superiority nor equivalence to Cheap.

Scope and limitations. Our results do not imply a universal normalized penalty or audit budget. CUB conclusions depend on declared comparator classes and normalization, and its exact oracle is restricted to a reduced empirical problem. Non-CBM results expose normalization sensitivity; AwA2 is prespecified and blind whereas CelebA is exploratory; and reduced-audit medians do not imply run-level reliability.

Policy claims are likewise specific to their inferential form. LAD's fixed-quota result is not a per-family population certificate: on the fixed CONF panel, Direct regret (.0390) is 46.2% below the matched 20% random-quota expectation (.0725), supporting selection beyond quota reduction, but not Direct superiority, population authorization, or end-to-end resource savings. Failed or stopped policies do not establish universal infeasibility. Extension to new populations requires validation of the corresponding measurement and policy contract.

Conclusion. Ranking reuse can be useful without being automatic. One ranking is enough only relative to a declared use and evidence supporting its adequacy.

REPRODUCIBILITY STATEMENT

Detailed diagnostic definitions, dataset and split specifications, model and family generators, comparator construction, training and checkpoint-selection protocols, intervention procedures, and statistical assumptions are provided in Appendices A–G.

The anonymous supplementary artifact includes machine-readable feature schemas, protocol and configuration files, seed and family manifests, canonical result tables, prospective freeze and evaluation records, and replay code for analyses whose executable implementations are preserved.

Where a historical analysis is supported by frozen protocol and result artifacts but its original low-level generator or standalone implementation is not separately archived, that replay boundary is stated explicitly. Unmapped or unrecoverable numerical endpoints are excluded from active quantitative claims rather than reconstructed from manuscript summaries.

AI USE STATEMENT

Large language models were used to assist with manuscript drafting and editing, organization, and consistency checking. They were not treated as sources of scientific evidence. All experimental designs, implementations, analyses, numerical results, citations, scientific claims, and final conclusions were reviewed and verified by the authors, who take full responsibility for the content of the paper.

ETHICS STATEMENT

This work conducts secondary analyses of existing benchmark datasets and collects no new human-subject data. We use these datasets only for the research purposes described in the paper and follow their respective access and usage requirements. The proposed diagnostics and mechanism probes are intended for research analysis rather than clinical or other high-stakes deployment.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2), 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535. URL https://arxiv.org/abs/2101.02703.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Bau_Network_Dissection_Quantifying_CVPR_2017_paper.html.

Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/c7bf0b7c1a86d5eb3be2c722cf2cf746-Abstract.html.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://www.jmlr.org/papers/v22/20-1316.html.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the International Conference on World Wide Web*, pp. 613–622, 2001. doi: 10.1145/371920.372165. URL https://www.wisdom.weizmann.ac.il/~naor/PAPERS/rank_www10.html.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375. URL https://research.ibm.com/publications/the-reusable-holdout-preserving-validity-in-adaptive-data-analysis.

Eric Enouen and Sainyam Galhotra. Debugging concept bottleneck models through removal and retraining. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/03d113a060c0ac93a5859517a0f07271-Abstract-Conference.html.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21400–21413, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/867c06823281e506e8059f5c13a57f75-Abstract-Conference.html.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/770cabd044c4eacb6dc5924d9a686dce-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/944ecf65a46feb578a43abfd5cddd960-Abstract-Conference.html.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Lijie Hu, Chenyang Ren, Zhengyu Hu, Hongbin Lin, Cheng-Long Wang, Zhen Tan, Weimin Lyu, Jingfeng Zhang, Hui Xiong, and Di Wang. Editable concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24678–24726, 2025. URL https://proceedings.mlr.press/v267/hu25u.html.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=jESY2WTZCe.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.06440.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11264–11272, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Molchanov_Importance_Estimation_for_Neural_Network_Pruning_CVPR_2019_paper.html.

Ari S. Morcos, David G. T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1803.06959.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971. URL https://research.dial.uclouvain.be/entities/publication/a725e7a1-a49a-403e-8774-d62d8b8e000a.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://research.ibm.com/publications/label-free-concept-bottleneck-models.

Ariel D. Procaccia and Jeffrey S. Rosenschein. The distortion of cardinal preferences in voting. In *Cooperative Information Agents X*, volume 4149 of *Lecture Notes in Computer Science*, pp. 317–331, 2006. URL https://cris.huji.ac.il/en/publications/the-distortion-of-cardinal-preferences-in-voting/.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. URL https://arxiv.org/abs/1602.04938.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31504–31520, 2023. URL https://proceedings.mlr.press/v202/shin23a.html.

David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to intervene on concept bottlenecks. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46556–46571, 2024. URL https://proceedings.mlr.press/v235/steinmann24a.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (04):6021–6029, 2020. doi: 10.1609/aaai.v34i04.6064. URL https://ojs.aaai.org/index.php/AAAI/article/view/6064.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2205.15480.

A CONTROLLED POPULATION EVIDENCE

A.1 CONTROLLED REDUNDANCY: SINGLE-UNIT AND GROUP EFFECTS

The controlled population varies the number of redundant copies over $1, 2, 4, 8, 16$. A single-unit lesion replaces one copy at a time; its reported effect is averaged over copies within the generating replicate. A group lesion replaces the entire redundant group. Both interventions use the same controlled synthetic population and held-out evaluation split. The inference unit is the seed/family replicate ($n = 15$), not the individual redundant copy. Table 5 preserves the reported means and 95% confidence intervals.

Copies	Single-unit effect	95% CI	Group effect	95% CI
1	.1813	[.1792,.1833]	.1813	[.1792,.1833]
2	.0904	[.0899,.0909]	.1950	[.1939,.1961]
4	.0452	[.0449,.0455]	.2024	[.2011,.2037]
8	.0225	[.0224,.0226]	.2050	[.2037,.2063]
16	.0113	[.0112,.0113]	.2072	[.2066,.2078]

Table 5: Controlled redundancy on a shared held-out evaluation split. The single-unit effect is a per-copy mean; the group effect removes the whole group. Intervals use the seed/family as the inference unit ($n = 15$).

Single-copy sensitivity decreases as redundancy increases, while the joint intervention remains consequential. This separates a unit's conditional marginal effect from the effect of removing its functional group. It does not make the two interventions budget-matched: they deliberately alter different numbers of copies. Nor does it show that a particular shared ranking is suboptimal. The within-model shared-versus-specific comparisons address that separate question. The result motivates evaluating the deployed set-level action instead of treating isolated-unit effects as its substitute.

A.2 KNOWN-ROLE DIAGNOSTIC RESULTS AS SUPPORTING EVIDENCE

Table 6 reports the canonical 20%-budget mean utilities for Lesion diagnostic and specialization audit. These are utilities of descriptor-selected sets, not improvements over a random baseline. Mean utilities and random-relative improvements are distinct outcomes and are not interchanged.

	Lesion diagnostic		Specialization audit	
Descriptor	Mean	95% CI	Mean	95% CI
Contribution	.0221	[.0215,.0228]	.1618	[.1558,.1679]
Intervention value	.0160	[.0125,.0191]	.1004	[.0854,.1122]
Participation/usage	.00002	[−.00003,.00007]	.0031	[.0030,.0032]
Concept quality	.0000	[.0000,.0000]	.00018	[.00015,.00020]

Table 6: Supporting known-role results at a 20% selection budget. Mean utility and intervals are reported for each descriptor–objective cell; no random-baseline subtraction or between-descriptor test is reconstructed.

These cells illustrate differences in the functional utility recovered by different descriptors. Contribution has the largest reported mean in both displayed objectives; these cells alone therefore do not establish a reversal of the winning descriptor across objectives. They remain supporting evidence, whereas controlled redundancy supplies the Introduction's direct functional motivation and the CUB/SUN comparisons establish action-dependent ranking-rule usefulness. The source display of a zero mean or interval is not upgraded to a claim of exactly zero population effect.

B Extended Related Work and Scope of the Contribution

B.1 Attribution semantics and functional evaluation

Local surrogates (Ribeiro et al., 2016), additive feature attribution (Lundberg & Lee, 2017), and axiomatic gradient attribution (Sundararajan et al., 2017) answer related but differently specified questions about model predictions. Global importance can also be defined through predictive performance: SAGE accounts for feature interactions while summarizing contributions to a global loss (Covert et al., 2020). The removal-based framework of Covert et al. (2021) explicitly separates the removal mechanism, the model behavior being explained, and the summary of feature influence. We therefore do not present the multiplicity of importance definitions as a new discovery. Our comparison asks an additional operational question: after the model, candidate units, actions, budgets, and utility definitions are declared, what is lost by constraining several decisions to use one ordering?

B.2 Reliability of explanations and their evaluation metrics

Randomization tests show that visually plausible saliency maps can be insensitive to the learned model or data (Adebayo et al., 2018). ROAR evaluates feature removal with retraining (Hooker et al., 2019), whereas infidelity and sensitivity measure specified perturbation responses (Yeh et al., 2019). Saliency-metric sanity checks further show why metric implementations and parameter choices matter to comparative conclusions (Tomsett et al., 2020). Krishna et al. (2024) document disagreement among explanation methods and the difficulty it creates for practitioners. These studies motivate behavioral validation and careful measurement. Our distinct contrast is shared versus decision-specific prioritization, evaluated through the utilities of the resulting actions. We also separate an ordinal conflict diagnostic from a set-level loss. This does not imply that structural features cannot predict that loss in a particular population; the reported Direct–Cheap comparison explicitly leaves incremental Direct superiority unestablished.

B.3 Semantic alignment, unit dependence, and pruning

Network Dissection measures alignment between individual units and semantic concepts (Bau et al., 2017), while TCAV measures model sensitivity along concept directions (Kim et al., 2018). Morcos et al. (2018) examine single-direction reliance, class selectivity, and generalization, illustrating why a semantically selective direction need not have the expected task importance. Pruning methods instead estimate the effect of removing components on a specified objective (Molchanov et al., 2017; 2019). Our controlled role and redundancy experiments serve as functional motivation within this literature. They distinguish conditional marginal effects from the responsibility of a jointly removed group; they do not establish a universal causal importance score. A lesion-based diagnostic is not automatically a measure of prediction quality. The main ranking-reuse experiments therefore retain the declared action and replacement operator as part of the estimand.

B.4 Concept bottlenecks and intervention-specific policies

Concept bottleneck models enable test-time edits of predicted semantic concepts (Koh et al., 2020). Concept Embedding Models increase the expressiveness of concept representations (Espinosa Zarlenga et al., 2022), while leakage analyses clarify that concept supervision alone does not settle what information a bottleneck conveys (Havasi et al., 2022). Post-hoc and label-free approaches broaden how concept-based predictors can be obtained (Yuksekgonul et al., 2023; Oikarinen et al., 2023). These advances motivate evaluation beyond semantic prediction quality.

Intervention-policy research is especially close to our operational setting. Shin et al. (2023) show that concept selection and intervention granularity substantially affect correction outcomes at matched intervention counts. Intervention-aware CEMs learn to benefit from and request interventions (Espinosa Zarlenga et al., 2023); intervention memories support reuse of previous interventions (Steinmann et al., 2024). Editable CBMs address model-editing requests (Hu et al., 2025), and removal-and-retraining methods address shortcut concepts that local corrections may not resolve (Enouen & Galhotra, 2026). Accordingly, neither useful intervention selection nor reuse of human feedback is new to this paper. Our object is different: the constraint of using the same unit ordering for distinct actions, including actions with different utility functions. We compare that con-

straint with allowing separate rankings, rather than proposing a better correction policy, intervention memory, or model-editing algorithm.

B.5 Rank aggregation, ordinal information, and set utility

Rank aggregation combines input rankings into a consensus ordering, including objectives based on pairwise disagreement (Dwork et al., 2001). The distinction between ordinal preferences and cardinal utility also has a direct precedent in social-choice distortion (Procaccia & Rosenschein, 2006). We do not claim to originate that distinction or the acyclicity condition for extending strict preferences. The constructive separation in this paper applies it to ranking reuse: identical strict preference structures need not determine the minimum sharing cost on the declared utility scale. This is not a claim of invariance under arbitrary normalization.

Budgeted top-k interventions introduce a further issue: utility can depend on the selected set, not only on individual positions or pairwise preferences. Classical submodular maximization supplies guarantees under explicit set-function assumptions (Nemhauser et al., 1978); we do not assume that our empirical utilities satisfy those assumptions. Our set-level alignment controls therefore evaluate the actual action utility, and our reduced exhaustive control establishes optimality only within its specified candidate subset. Neither a pairwise surrogate nor a small-instance optimum is described as a global optimum of the full empirical problem.

B.6 Selective prediction and risk-control guarantees

Selective classification studies predictions with a reject option and risk–coverage trade-offs (Geifman & El-Yaniv, 2017); SelectiveNet jointly learns prediction and selection (Geifman & El-Yaniv, 2019). Risk-controlling prediction sets provide high-probability control of population risk under their stated conditions (Bates et al., 2021). Learn then Test uses hypothesis testing to validate candidate procedures, including selection-sensitive comparisons (Angelopoulos et al., 2025). Conformal Risk Control controls expected monotone loss under its assumptions (Angelopoulos et al., 2024). These guarantees differ in the risk functional, probability statement, and admissible procedure. They are not interchangeable labels for an empirically low error rate.

Our contribution concerns the evaluation object, not a new generic confidence bound: the relevant loss concerns the consequences of sharing a ranking. A fixed per-family threshold and a batch policy selecting a fixed fraction need not share the same inferential argument. Our fixed-quota results therefore distinguish observed selected-family violations from population coverage guarantees. Similarly, a guarantee for a measured finite-sample cost is not automatically a guarantee for a latent population utility without addressing measurement uncertainty. Useful audit ordering can therefore coexist with insufficient evidence for a particular reuse claim.

B.7 Distribution shift, adaptivity, and evidence identity

Conformal prediction under covariate shift uses additional structure, including a specified reweighting relationship, to extend coverage arguments beyond the original sampling distribution (Tibshirani et al., 2019). This is different from assuming that an audit score's ordering or numerical calibration transfers to a new population. Research on adaptive data analysis also shows why repeatedly consulting a holdout can invalidate ordinary validation reasoning (Dwork et al., 2015). We do not claim to implement a reusable-holdout algorithm; instead, we distinguish frozen independent evaluations from post-hoc sensitivity analyses and preserve the original verdicts of stopped studies.

Positioning. Our contribution concerns the different roles of structure, action utility, and validation in ranking reuse. A cheap predictor can be useful without eliminating the need to define and independently validate the consequence; a reliable measurement does not by itself authorize a deployment policy.

C Framework: Definitions, Proofs, and Inferential Scope

C.1 Evaluation objects, signed contrasts, and optimal penalties

Units, families, and protocols. A unit is a candidate concept, feature, or component to which an action applies. A family is the study's designated cross-model or generator-level evaluation unit. Aggregating model seeds within a family does not make those seeds additional independent deployment families. Each study specifies its family definition, aggregation weights, and resampling unit.

Protocol Π specifies the candidates, decisions and replacement operators, budget-to-set mapping, utility direction, aggregation weights, normalization, comparator classes, and data-use rules. Percentage budgets retain their original rounding and tie-breaking conventions. An ordering π induces a nested prefix $S_b(\pi)$ at each declared budget. Set utility need not be additive, submodular, or monotone, and lesion effects do not receive a universal causal interpretation. The unit preferences in Proposition 1 must be declared; they are not a unique automatic consequence of an arbitrary set function.

Measured contrast. equation 1 explicitly compares a fixed shared ordering with a tuple of decision-specific orderings. The hat on $\widehat{U}$ and $\widehat{C}$ marks a finite-data measurement, not an unbiasedness assertion. The index T identifies the measurement split. Dependence on fitted rankers is suppressed only after the comparator pair is fixed. A shared fitted ranker may outperform a decision-specific fitted ranker on held-out data, so $\widehat{C}_{\Pi,T}$ can be negative. Negative values are retained; neither clipping at zero nor substituting oracle values is justified.

If a family contains several models, model-level contrasts are combined in the original order of operations. Averaging individually normalized quantities is not generally equivalent to normalizing averaged utilities. Historical files may call $\widehat{C}$ a sharing cost; this notation clarification does not rename or modify canonical fields, values, decisions, thresholds, or verdicts.

Raw, menu-normalized, and anchor-normalized quantities. Raw comparisons use $a_{F,d,b,T} = 1$. For a declared reference menu $\mathcal{M}$, menu-normalized regret is

$$R_{F,d,b,T}(S) = \frac{U^{\max}_{F,d,b,T} - \widehat{U}_{F,d,b,T}(S)}{U^{\max}_{F,d,b,T} - U^{\min}_{F,d,b,T}}. \tag{6}$$

The menu, extrema, and their estimation data belong to the original protocol. When both policies use the same positive denominator, their regret difference is the scaled utility difference in equation 1. Mean regret, maximum-over-decisions regret, and a mean paired contrast are noninterchangeable aggregations.

Some studies use validation anchors such as $U^{\mathrm{spec,val}}_{d,b} - \mathbb{E}U^{\mathrm{random,val}}_{d,b}$; others use separate random-anchor ranges and a denominator floor. The original anchor source, estimation split, floor, and aggregation are retained. The subscript T permits study-specific split-dependent numerical scales generated by a frozen rule; it does not require that all studies share numerical anchors. The framework uses positive scales; degenerate values receive the original protocol's handling, not an outcome-driven replacement. A normalized contrast need not lie in $[0, 1]$ and is not automatically an accuracy percentage-point loss.

An optimal penalty for a fixed utility problem. To distinguish optimization from fitted-policy performance, fix utilities $U_{F,d,b}$, weights, and positive scales independently of the candidate ordering being optimized. These utilities can describe a finite exact-control table or a specified population utility problem; an exact finite-table optimum is not thereby an infinite-data optimum. Define

$$V_{F,d}(\pi) = \sum_b w_{d,b} \frac{U_{F,d,b}(S_b(\pi))}{a_{F,d,b}}. \tag{7}$$

Let $\mathcal{H}_s$ be the shared class and $\mathcal{H}_d$ the class for decision d. They are nonempty finite sets of realizable orderings, and $\mathcal{H}_s \subseteq \bigcap_d \mathcal{H}_d$. The two optima and their difference are

$$W^*_{\text{spec}}(F) = \sum_d \max_{\pi \in \mathcal{H}_d} V_{F,d}(\pi), \tag{8}$$

$$W^*_{\text{shared}}(F) = \max_{\pi \in \mathcal{H}_s} \sum_d V_{F,d}(\pi), \tag{9}$$

$$C^*_\Pi(F) = W^*_{\text{spec}}(F) - W^*_{\text{shared}}(F) \geq 0. \tag{10}$$

Every admissible shared ordering can be used by every decision-specific class. Thus, for every $\pi \in \mathcal{H}_s$, $\sum_d \max_{\rho \in \mathcal{H}_d} V_{F,d}(\rho) \geq \sum_d V_{F,d}(\pi)$. Taking the maximum on the right proves nonnegativity. Class nesting is a sufficient condition; without it the comparison may mix the sharing constraint with unequal feasible actions, and nonnegativity is not automatic.

Each decision-specific optimum selects one ordering jointly across the declared budgets. Replacing it with a separately optimal set at every budget changes the benchmark: those sets need not be nested prefixes of one ordering. In particular, a single-decision problem with identical classes has $C^*_\Pi = 0$, even when separately optimized sets at different budgets cannot be jointly realized. This is why the budget sum is inside each decision's maximization.

A matched fitted comparison, a stronger search, and an exact optimum have different scopes. A held-out fitted contrast is neither automatically equal to, nor a lower or upper bound on, C^*_Π. Any such relation requires an additional argument controlling approximation and estimation. Exhaustive eight-concept CUB search is exact only for its specified reduced problem and optimization target. No experimental score is silently promoted to a full-space optimum.

C.2 Exact common orderability and budget scope

For a finite candidate set, let $\{\succ_{F,d}\}_{d \in \mathcal{D}}$ be the declared strict unit-preference relations. Their union graph contains $i \to j$ whenever at least one decision requires $i \succ_{F,d} j$.

Proof of Proposition 1. If an ordering preserves all edges, a directed cycle would imply $s(i_1) > s(i_2) > \cdots > s(i_r) > s(i_1)$, a contradiction. Conversely, a finite acyclic directed graph has a topological ordering. Assign distinct decreasing scores along that order; all edges, hence every declared strict preference, are preserved. □

This concerns unrestricted orderings. A particular linear score class may not represent the compatible ordering. Ties add no strict edge unless the original protocol resolves them. A predeclared margin or uncertainty rule can define an estimated graph, but the proposition about that graph is not automatically a simultaneous-confidence result about all population preferences.

Unit order versus the usefulness of a ranking rule. Budget-specific utility comparisons can favor different fixed ranking rules even when those rules' unit orders never change. Formally, a rule-level relation may be

$$\pi \succ^{\text{rule}}_{F,d,b} \pi' \quad \Longleftrightarrow \quad U_{F,d,b}(S_b(\pi)) > U_{F,d,b}(S_b(\pi')). \tag{11}$$

A reversal of this relation across budgets is not a reversal of the unit order inside either rule. The reported CUB descriptor–budget comparisons evaluate selected sets and downstream regret; they are not reinterpreted here as an additional empirical budget-indexed unit graph.

If a protocol explicitly declares $\succ_{F,d,b}$ over units, its graph is $G_F = \bigcup_{(d,b) \in \mathcal{I}_\Pi} G_{F,d,b}$ for the predeclared index set $\mathcal{I}_\Pi$. The same topological-order proof applies to this union. Budget-independent relations are the special case repeated across budgets. This optional generalization does not change

the historical graph, impose new constraints after observing outcomes, or permit independent per-budget rankings in a comparison that originally used one ordering per decision.

C.3 Conflict–cost separation with fixed full-menu ranges

Proposition 2 (precise statement). There exist fixed candidate sets, two complete strict decision orderings, and a fixed feasible top-1 menu for which every decision's utility minimum and maximum are always 0 and 1, but the optimal shared penalty varies with interior utility gaps. The ratio of two positive penalties can be arbitrarily large although all strict preference relations and menu ranges are unchanged.

Proof. Fix three units a, b, c, two equally weighted decisions, budget one, and the same menu $\mathcal{M} = \{\{a\}, \{b\}, \{c\}\}$ in every instance. Both shared and decision-specific classes contain all six permutations of the three units. For $0 < \eta < 1$, define

$$\begin{array}{c|ccc} & \{a\} & \{b\} & \{c\} \\ \hline U_1 & 1 & 1-\eta & 0 \\ U_2 & 1-\eta & 1 & 0 \end{array}. \tag{12}$$

For every such η, decision 1 ranks $a \succ b \succ c$ and decision 2 ranks $b \succ a \succ c$. The union graph is always $\{a \to b, b \to a, a \to c, b \to c\}$. Its entire pairwise preference information, not just its cycle count, is unchanged. For both decisions,

$$\max_{S \in \mathcal{M}} U_d(S) = 1, \qquad \min_{S \in \mathcal{M}} U_d(S) = 0, \qquad a_{d,1} = 1. \tag{13}$$

The value $a_{d,1} = 1$ is therefore both a fixed raw scale and the actual full-menu normalization denominator. Decision-specific orders achieve utility one for each decision, so $W^*_{\text{spec}} = 1$. A shared order beginning with a or b achieves mean utility $1 - \eta/2$, while one beginning with c achieves zero. Consequently,

$$C^*_\Pi(\eta) = \min\{\eta/2, \eta/2, 1\} = \eta/2. \tag{14}$$

For any $L > 1$, choose $\eta_1 = 1/2$ and $\eta_2 = 1/(2L)$. Both stay strictly inside $(0, 1)$, so no ties alter the graph, but $C^*_\Pi(\eta_1)/C^*_\Pi(\eta_2) = L$. The ratio is unbounded as $L \to \infty$; the absolute costs remain bounded below $1/2$. □

For any fixed tolerance $\epsilon \in (0, 1/2)$, choices $0 < \eta_- < 2\epsilon < \eta_+ < 1$ put the two optimal penalties on opposite sides of ϵ without changing the menu, graph, or ranges. Thus the same structural information and range normalization do not determine top-1 reuse adequacy even at a fixed admissible tolerance.

What is fixed, and what is not. The result varies an interior utility margin, not an overall multiplicative scale or the extrema. The full feasible menu includes c throughout. If the menu were changed to contain only a, b, each range would instead be η and range-normalized mean penalty would be $1/2$; that is a different declared normalization problem. The strengthened witness consequently applies to raw utility and the stated full-menu range normalization, not arbitrary anchors, normalizers, menus, or historical experimental denominators.

With deterministic orderings, summing the two regrets or taking their maximum gives η rather than the equal-weight mean $\eta/2$. Under the mean objective, randomization over the shared first choice cannot improve the optimum: expected utility is a convex combination of the deterministic values. The construction is a mathematical comparison, not a new empirical replication or a proof that all real populations have large sharing penalties. It does not exclude useful prediction from richer structural features. The broader ordinal–cardinal information distinction has prior foundations (Procaccia & Rosenschein, 2006).

C.4 Independent measurements, data splitting, and evidence roles

Two forms of separation coexist. Within a measured family, set-optimization, audit, and confirmation observations have distinct declared roles. Across families, FIT, CERT, and CONF designate panels for policy selection, testing, and subsequent evaluation. A confirmation measurement for a FIT family is not an outcome from the reserved CONF-family panel.

In the blind AwA2/CelebA audits, rankings and validation anchors are frozen before confirmation. The identity prediction $\widetilde{C}_{\mathrm{confirm}} = \widehat{C}_{\Pi,\mathrm{audit}}$ is assessed against the independent measurement $\widehat{C}_{\Pi,\mathrm{confirm}}$. Numerical errors measure agreement; A_{pair} measures eligible-family ordering under the original tie and eligibility rules. Ordering agreement is not a calibration guarantee. The budget sweep changes audit observations, not models, rankings, anchors, or confirmation outcomes. Other studies retain their own normalization rules; the framework does not require identical numerical denominators on every split.

Raw-cost, denominator, and residualized checks are sensitivity analyses unless originally designated primary. A LAD sensitivity result does not establish the same control for AwA2 or CelebA. Learned Direct, Cheap, and structural scores are inputs to selection; their correlations do not redefine the measured functional contrast.

Inference follows the generating model/seed or family unit. Concept pairs, repeated audit samples, and multiple seeds within one family do not constitute additional independent family draws. Frozen held-out comparisons, prospective evaluations, and retrospective controls retain distinct evidential roles; a successful reanalysis never replaces an earlier prospective verdict.

C.5 Fixed-Policy Risk and Coverage Bounds

Authorization is specific to the policy form and its inferential target. The following argument conditions on a declared data environment, family generator, and outcome-independent eligibility mechanism. It requires CERT families, including their within-family measurement randomness, to be generated independently and identically under that generator, and the entire per-family map g to be frozen independently of CERT. This includes preprocessing, score fitting, thresholds, comparison rankers, and tie handling. The measurement definition, tolerance, and error levels are likewise fixed without consulting CERT outcomes. Neither a split name nor an artifact hash is by itself sufficient for these conditions.

For the Gesture and Letter studies used below, Appendix F.4 specifies the corresponding family generators and phase-freezing procedures. Their iid interpretation is conditional on the respective fixed benchmark data environments rather than a claim of universality across arbitrary datasets or generators.

Let

$$V_i = \mathbf{1}\{\widehat{C}_{\Pi,\mathrm{confirm}}(F_i) > \epsilon\}, \qquad G_i = g(F_i),$$

and define

$$M = \sum_{i=1}^{N} G_i, \qquad E = \sum_{i=1}^{N} G_i V_i. \tag{15}$$

For $Q(g) > 0$, fixed per-family selection gives

$$M \sim \mathrm{Binomial}(N, Q(g)), \qquad E \mid M = m \sim \mathrm{Binomial}(m, R_\epsilon(g)).$$

Let $B_p^{-1}(a, b)$ denote the p quantile of a Beta distribution. One-sided exact bounds are

$$U_R(E, M; \delta_R) = \begin{cases} 1, & M = 0 \text{ or } E = M, \\ B^{-1}_{1-\delta_R}(E+1, M-E), & 0 \le E < M, \end{cases} \tag{16}$$

and

$$L_Q(M, N; \delta_Q) = \begin{cases} 0, & M = 0, \\ B^{-1}_{\delta_Q}(M, N-M+1), & 1 \le M \le N. \end{cases} \tag{17}$$

For $E = 0 < M$, $U_R = 1 - \delta_R^{1/M} > 0$; for $M = N$, $L_Q = \delta_Q^{1/N}$. An empty SHARE set has undefined empirical selective risk. Setting $U_R = 1$ when $M = 0$ is a conservative gate convention rather than a population-risk estimate. For $0 < \alpha < 1$ and $0 < \kappa \le 1$, an empty SHARE set therefore cannot pass.

Joint validity and policy selection. For $Q(g) > 0$ and every $m > 0$, exact binomial coverage gives

$$\Pr(R_\epsilon(g) > U_R \mid M = m) \le \delta_R.$$

Marginalizing over M preserves this bound. Separately,

$$\Pr(Q(g) < L_Q) \le \delta_Q.$$

Hence

$$\Pr\{R_\epsilon(g) \le U_R,\ Q(g) \ge L_Q\} \ge 1 - \delta_R - \delta_Q. \tag{18}$$

No independence between the two bounds is required. Thus the probability of granting authorization when $R_\epsilon(g) > \alpha$ or $Q(g) < \kappa$ is at most δ whenever $\delta_R + \delta_Q \le \delta$.

If $Q(g) = 0$, then $M = 0$ almost surely and the gate never passes; no value is assigned to the undefined population conditional risk.

Conditioning on FIT permits a data-selected policy to be frozen using FIT and subsequently evaluated on independent CERT families. Selection among policies using CERT, repeated attempts on the same CERT panel, or data-dependent error allocations would require corresponding multiplicity or sequential control. These are standard risk-control requirements rather than new confidence machinery introduced here.

What is being certified. The certified event concerns the finite confirmation measurement of the fitted-policy sharing contrast, including its declared measurement randomness. It is not automatically a guarantee for infinite-data expected utility and is not a certificate for the oracle gap C^*_Π. Either stronger target would require an additional argument relating measurement, estimation, and optimization.

A failed certificate denies authorization to the evaluated frozen policy under the declared generator. It does not establish that no other threshold, policy class, or family generator could admit a feasible reuse policy.

C.6 Fixed quotas, population feasibility, and claim scope

LAD ranks a batch of families and selects a fixed fraction. Even when the score and selection algorithm are frozen, whether one family is selected can depend on all batch scores. The selected observations therefore do not automatically satisfy the preceding iid conditional-binomial model. A population-risk claim needs an argument tailored to batch selection. The realized batch reuse fraction is fixed by design, not an independently certified population coverage.

LAD's held-out counts, originally reported bounds, and utility outcomes retain their distinct evidential role. Its tolerance $\epsilon = .40$ is not substituted for $\epsilon = .10$ in later per-family studies. No new batch-valid certificate is asserted here.

Relative to the fitted decision-specific comparator, let $C_i = \widehat{C}_{\Pi,\mathrm{confirm}}(F_i)$ and $J_N(g) = N^{-1}\sum_i g_i C_i$. Choosing the reference action contributes zero incremental sharing contrast, not necessarily zero absolute task loss. The comparison with always sharing therefore combines lower reuse frequency with selection of which families to share. Same-quota comparisons address a different contrast. A Direct–Cheap confidence interval crossing zero establishes neither superiority nor equivalence. Reuse quota is not measured wall-clock or net resource savings.

A necessary safe-mass condition. For a per-family policy with $R_\epsilon(g) \le \alpha$ and $Q(g) \ge \kappa > 0$, define $p_{\mathrm{safe}} = \Pr(\widehat{C}_{\Pi,\mathrm{confirm}} \le \epsilon)$. Then

$$p_{\mathrm{safe}} \ge \Pr(g = 1, \widehat{C}_{\Pi,\mathrm{confirm}} \le \epsilon) = Q(g)(1 - R_\epsilon(g)) \ge (1 - \alpha)\kappa. \tag{19}$$

This necessary condition does not ensure that available scores identify a feasible subset. A FIT safe fraction estimates a population probability; nonexistence claims require valid uncertainty and any applicable multiplicity control.

For a finite declared threshold set $\mathcal{T}$, define $r^*_{\mathcal{T}}(q) = \min_{t\in\mathcal{T}:\, Q(g_t)\ge q} R_\epsilon(g_t)$, with an empty minimum equal to $+\infty$. Population feasibility in that class is equivalent to $r^*_{\mathcal{T}}(\kappa) \le \alpha$. An empirical frontier is not this population frontier. An infimum over an unrestricted class requires care about

attainment. Enlarging a nested policy class cannot remove an existing feasible policy, whereas finite-sample fitting of a more flexible policy can worsen held-out performance. Failure analyses retain these levels rather than claiming universal impossibility.

D Decision Dependence and Sharing Contrasts: Extended Results

D.1 Evidence roles and comparisons

The Section 4 evidence separates fitted-comparator results from a reduced empirical oracle. Empirical contrasts use the signed $\widehat{C}_{\Pi,T}$ of Section 3; the reduced oracle optimizes a specified finite utility problem rather than estimating the out-of-sample performance of its selected ordering. Positive differences favor decision-specific policies: shared-minus-specific for regret and specific-minus-shared for utility scores. Different menus and denominators are not pooled.

The CUB same-feature linear, CUB stronger-search, and SUN consequence studies are retrospective cross-model held-out analyses. A protocol frozen before a particular analysis does not make the dataset an untouched prospective replication. The reduced CUB control is an evaluation-only oracle using official-test utilities. FSD provides retrospective set-level alignment; and non-CBM provides split-separated replication on the stated datasets rather than untouched-dataset prospective evidence. The remaining role and crossover results are reported only within their original evidential scope.

D.2 Functional motivation and action-specific comparisons

Appendix A.1 and Table 5 report the full controlled-redundancy results. As redundancy increases from one to 16 copies, the mean per-copy lesion effect decreases from .1813 to .0113, whereas the corresponding full-group effect increases from .1813 to .2072 on the same held-out evaluation split. This contrast separates conditional single-unit sensitivity from group responsibility. Because the two interventions alter different numbers of units, it is used as functional motivation rather than as an equal-budget ranking comparison or a sharing penalty.

CUB Full adaptive versus Standard CBM. Table 7 reports two protocol-specific lesion-gap contrasts. Positive values denote Full adaptive minus Standard CBM, not task-accuracy gains, descriptor Contribution minus Participation, or shared minus decision-specific regret. The held-out result is distinct from the larger current-test result; neither is substituted for the other. Their difference is not reported as a new bias estimate or statistical test.

CUB evaluation protocol	Full adaptive − Standard CBM	95% CI
Held-out ranked lesion	+.01033	[.00867,.01199]
Current-test	+.03478	[.02818,.04138]

Table 7: CUB lesion-gap contrasts within two evaluation protocols. These are model comparisons, not ranking-rule reversals or shared-ordering penalties.

The SUN audit reuses 30 canonical models (Joint and Sequential, 15 seeds each). Its descriptor set, four decisions, three budgets, regret definition, 90 candidate crossovers, and formulation-by-seed hierarchical bootstrap were frozen before the new matrix was inspected. Of the 90 candidates, 42 have point-sign reversals and 17 satisfy both the original paired-interval and nontriviality rules. This is not a claim that all 90 candidates reverse, or that the 17 results are a new simultaneous multiplicity-adjusted discovery set. The original analysis reports 10,000 hierarchical bootstrap replicates.

SUN E15 action–utility contract. The SUN E15 audit evaluates 102 concepts for each of 30 models (Joint and Sequential, 15 seeds per formulation). Rankings are constructed from validation-side descriptor values and evaluated on the held-out test split at budgets 5%, 10%, and 20%, corresponding to $k \in \{5, 10, 20\}$.

For lesion, selected predicted concept activations are replaced jointly by their training-split predicted-concept means. Writing μ_j^{train} for the training-split mean activation of concept j,

$$U_{\text{lesion}}^{\text{SUN}}(S) = \text{Acc}_{\text{test}}(f) - \text{Acc}_{\text{test}}\left(f_{S \leftarrow \boldsymbol{\mu}_S^{\text{train}}}\right).$$

For correction, selected predictions are replaced by the corresponding soft SUN ground-truth concept values:

$$U_{\text{corr}}^{\text{SUN}}(S) = \text{Acc}_{\text{test}}\left(g(\hat{\mathbf{c}}_{S \leftarrow \mathbf{c}_S^{\text{soft}}})\right) - \text{Acc}_{\text{test}}(g(\hat{\mathbf{c}})).$$

For vulnerability screening, each selected concept is lesioned separately using the same training-mean replacement, and the selected-set utility is the mean single-concept test accuracy drop:

$$U_{\text{screen}}^{\text{SUN}}(S) = \frac{1}{|S|} \sum_{j \in S} \left[\text{Acc}_{\text{test}}(f) - \text{Acc}_{\text{test}}\left(f_{j \leftarrow \mu_j^{\text{train}}}\right)\right].$$

For specialization audit,

$$U_{\text{spec}}^{\text{SUN}}(S) = \frac{1}{|S|} \sum_{j \in S} \left(1 - \frac{H_{j,\text{test}}}{\log C}\right),$$

where $H_{j,\text{test}}$ is the concept's test class-profile entropy and C is the number of classes.

SUN normalization and aggregation. For each formulation q, seed s, decision d, and budget b, the complete six-descriptor menu $\mathcal{R}_{\text{SUN}}$ defines

$$U_{q,s,d,b}^{\max} = \max_{r \in \mathcal{R}_{\text{SUN}}} U_{q,s,d,b}(r), \qquad U_{q,s,d,b}^{\min} = \min_{r \in \mathcal{R}_{\text{SUN}}} U_{q,s,d,b}(r),$$

and normalized regret

$$R_{q,s,d,b}(r) = \frac{U_{q,s,d,b}^{\max} - U_{q,s,d,b}(r)}{U_{q,s,d,b}^{\max} - U_{q,s,d,b}^{\min}}.$$

Each matrix cell averages normalized regret equally over the two formulations, 15 seeds per formulation, and the three declared budgets. Paired descriptor contrasts use the frozen formulation-by-seed hierarchical bootstrap with 10,000 draws. Thus SUN and CUB instantiate the same abstract ranking-reuse question under distinct frozen action operators and normalization contracts; their numerical regret scales are not pooled.

SUN descriptor	Lesion	Correction	Screening	Specialization
Participation	.567	.538	.742	.996
Specialization	.525	.776	.480	.000
Concept quality	.751	.389	.592	.372
Contribution	.168	.394	.205	.718
Vulnerability	.407	.741	.424	.793
Intervention value	.772	.243	.615	.734

Table 8: SUN E15 six-descriptor by four-objective normalized-regret matrix; lower is better. Within each formulation, seed, objective, and budget, regret is normalized by the max–min range of the complete six-descriptor menu, then averaged equally over Joint/Sequential formulations, 15 seeds per formulation, and budgets $k = 5, 10, 20$. Lesion uses training-mean concept replacement, correction uses soft ground-truth replacement, screening averages single-concept training-mean-lesion accuracy drops, and specialization averages normalized class-profile specialization. Directly aligned descriptor–objective cells are retained as sanity checks rather than independent nontrivial reversals.

The Contribution-minus-Intervention contrasts are $-.60479\,[-.74787, -.40273]$ for lesion and $+.15097\,[.03299, .26629]$ for correction. The screening contrast $-.410\,[-.595, -.221]$ remains a separate reported endpoint. These summarize action utility rather than unit-pair agreement. The original CUB correction–screening preference conflict is .738 [.725,.749], but the non-tautology audit identifies partial structural coupling for that pair and direct coupling for lesion–screening. Thus the conflict rates support the compatibility analysis; they are not sole evidence of a novel utility effect.

Additional CUB ranking-rule results. CUB compares Participation and Contribution ranking rules using their selected prefixes at 5%, 10%, and 20% budgets. Table 9 reports the rule-by-budget and pooled contrasts from this protocol. These results are distinct from the Full adaptive–Standard CBM model comparison and are not used to verify or relabel that contrast. Positive values favor Contribution because regret is lower-is-better. The reported intervals apply to the stated rule-by-budget contrasts; the sign of the 5% point estimate and exclusion of zero at 10% do not themselves constitute a tested between-budget interaction.

Decision	Budget	Participation − Contribution	95% CI
Correction	5%	+.034	[−.094, .164]
Correction	10%	−.144	[−.267, −.027]
Correction	20%	−.132	[−.182, −.078]
Correction	Pooled	−.081	[−.160, −.004]
Screening	5%	+.346	[.155, .520]
Screening	10%	+.286	[.113, .444]
Screening	20%	+.223	[.037, .406]
Screening	Pooled	+.285	[.149, .415]

Table 9: All reported CUB Participation–Contribution decision-budget contrasts.

D.3 CUB Comparators and Optimization Scope

same-feature linear comparator construction. The same-feature linear CUB analysis uses 15 leave-one-seed-out folds over Standard-CBM seeds 42–56. In each fold, the remaining 14 models form the development panel. Their validation splits provide the five functional input features and the feature-standardization statistics, whereas their official-test splits provide the unit-level objective targets used for ridge fitting. The held-out model is excluded from both operations: its validation-derived feature vectors instantiate the fitted concept scores and hence the held-out ranking, while its official-test split is used only for action-level utility evaluation.

Both the common and decision-specific rankers use the same five validation-derived functional features for each concept: (i) *participation*, the mean predicted concept activation; (ii) *specialization*, one minus normalized class-profile entropy; (iii) *contribution*, the increase in cross-entropy loss under single-concept zero lesion; (iv) *vulnerability*, the corresponding single-concept zero-lesion accuracy drop; and (v) *intervention value*, the accuracy gain obtained by replacing the predicted concept value with its binary ground-truth CUB attribute under the model's original concept-to-class classifier. All five feature coordinates are computed from validation-side measurements; no official-test outcome from the held-out model enters feature construction or ranking formation.

Feature means and standard deviations are estimated from the 14 development models only. Standard deviations below 10^{-12} are replaced by one. Every ridge model therefore receives the same development-standardized five-dimensional representation and uses the same nominal ridge parameter,

$$\alpha_{\text{ridge}} = 1.$$

Objective-specific fitting targets. Let $x_{mj} \in \mathbb{R}^5$ denote the validation-derived five-feature vector of concept j in development model m, and let $z(x_{mj})$ denote its development-only standardized representation. The ridge responses are constructed separately from the corresponding official-test single-unit objective arrays of the development models:

$$y_{mj}^{\text{lesion}} = L_{\text{off}}(f_{j\leftarrow 0}) - L_{\text{off}}(f), \tag{20}$$

$$y_{mj}^{\text{corr}} = \text{Acc}_{\text{off}}\left(g(\hat{\mathbf{c}}_{j\leftarrow c_j})\right) - \text{Acc}_{\text{off}}(g(\hat{\mathbf{c}})), \tag{21}$$

$$y_{mj}^{\text{screen}} = \text{Acc}_{\text{off}}(f) - \text{Acc}_{\text{off}}(f_{j\leftarrow 0}), \tag{22}$$

$$y_{mj}^{\text{spec}} = 1 - \frac{H_{mj}}{\log C}, \tag{23}$$

Here "off" denotes the frozen official-test split of a development model. L_{off} is cross-entropy loss, Acc_{off} is classification accuracy, $\hat{c}$ is the predicted concept vector, c_j is the binary ground-truth CUB

attribute, g is the model's original concept-to-class classifier, and H^{off}_{mj} is the official-test class-profile entropy. Because the four targets have different native numerical scales, each objective is transformed separately to its within-objective percentile ranks before ridge fitting. We denote the resulting target by $\rho_d(y_{mjd})$.

For decision d, the decision-specific ranker fits an intercept and a decision-specific coefficient vector:

$$(\widehat{b}_d, \widehat{\beta}_d) = \arg\min_{b,\beta} \sum_{m \in \mathcal{M}_{\text{dev}}} \sum_{j=1}^{212} \left[\rho_d(y_{mjd}) - b - z(x_{mj})^\top \beta\right]^2 + \alpha_{\text{ridge}} \|\beta\|_2^2. \tag{24}$$

The common ranker instead fits one intercept and one coefficient vector to the pooled percentile-ranked responses from all four objectives:

$$(\widehat{b}_{\text{common}}, \widehat{\beta}_{\text{common}}) = \arg\min_{b,\beta} \sum_{d=1}^{4} \sum_{m \in \mathcal{M}_{\text{dev}}} \sum_{j=1}^{212} \left[\rho_d(y_{mjd}) - b - z(x_{mj})^\top \beta\right]^2 + \alpha_{\text{ridge}} \|\beta\|_2^2. \tag{25}$$

In both fits, the intercept is estimated but not regularized, no sample weights are used, and $\alpha_{\text{ridge}} = 1$ is the nominal ridge parameter. Because the common fit pools four equally sized objective blocks under the summed-loss implementation while retaining the same nominal α, its ridge penalty is weaker relative to the data-fit term than in each decision-specific fit.

Each objective contributes the same number of concept-level responses to equation 25. Combined with the objective-wise percentile transformation, this gives the four objective blocks equal implicit weight in the pooled fit; no additional objective-specific loss weight is applied.

The common and decision-specific comparators share the candidate concepts, five validation-derived input features, development folds, feature scaling, linear model family, nominal ridge parameter $\alpha_{\text{ridge}} = 1$, and target-scale transformation. Because the common fit pools four equally sized objective blocks under a summed-loss parameterization while retaining the same nominal α, its ridge penalty is weaker relative to the data-fit term. Accordingly, the canonical linear analysis is interpreted as a same-feature, same-nominal-regularization comparison rather than a pure coefficient-sharing intervention with exactly matched per-response regularization.

Fitting targets versus held-out action utilities. The unit-level quantities above are fitting targets used to learn an ordering; they are not, in general, identical to the set-level utilities used to evaluate the selected top-k prefixes.

For a selected concept set S, the held-out CUB utilities are

$$U_{\text{lesion}}(S) = \text{Acc}(f) - \text{Acc}(f_{S \leftarrow 0}), \tag{26}$$

$$U_{\text{corr}}(S) = \text{Acc}(g(\hat{\mathbf{c}}_{S \leftarrow \mathbf{c}_S})) - \text{Acc}(g(\hat{\mathbf{c}}))\,, \tag{27}$$

$$U_{\text{screen}}(S) = \frac{1}{|S|} \sum_{j \in S} \left[\text{Acc}(f) - \text{Acc}(f_{j \leftarrow 0})\right], \tag{28}$$

$$U_{\text{spec}}(S) = \frac{1}{|S|} \sum_{j \in S} \left(1 - \frac{H_j}{\log C}\right). \tag{29}$$

In particular, lesion fitting uses single-concept zero-lesion cross-entropy increase, whereas the final lesion utility measures the accuracy drop produced by jointly zeroing the selected set. Likewise, vulnerability screening averages single-concept zero-lesion accuracy drops rather than jointly lesioning the selected set. These distinctions are part of the declared action contract rather than additional differences between the common and decision-specific rankers.

Held-out evaluation, normalization, and aggregation. For each held-out model, the fitted score induces an ordering of the 212 concepts. Rankings are evaluated at budgets 5%, 10%, and 20%, corresponding to

$$k = 11,\ 21,\ 42.$$

CUB policy	Mean normalized regret	Worst-decision regret
Common linear	.434629684	.640614650
Decision-specific linear	.078476445	.205079259
Single descriptor	.3387	.7881

Table 10: CUB same-feature linear policy results. The common and decision-specific rows use identical candidate concepts, input features, development folds, standardization, linear model family, nominal ridge parameter, target transformation, held-out action utilities, budgets, and menu normalization. Because the common fit pools four objective blocks with the same nominal ridge penalty, per-response regularization is not exactly matched; this table is therefore not interpreted as a pure coefficient-sharing intervention.

Objective	Common	Specific	Single descriptor
Lesion diagnostic	.5588	.1215	.1882
Correction	.3173	.0419	.1973
Vulnerability screening	.4772	.1478	.1814
Specialization audit	.3852	.0027	.7881

Table 11: CUB objective-level same-feature linear normalized regret. Lower is better. These rows decompose the same-feature linear comparison rather than constituting separate robustness experiments.

Within every decision–budget cell, the common and decision-specific rankings are evaluated on the same held-out official-test utility. The frozen seven-policy evaluation menu defines

$$U^{\max}_{F,d,b,T} = \max_{\pi \in \mathcal{M}_{\mathrm{CUB}}} U_{F,d,b,T}(S_b(\pi)), \qquad U^{\min}_{F,d,b,T} = \min_{\pi \in \mathcal{M}_{\mathrm{CUB}}} U_{F,d,b,T}(S_b(\pi)),$$

and normalized regret is

$$R_{F,d,b,T}(\pi) = \frac{U^{\max}_{F,d,b,T} - U_{F,d,b,T}(S_b(\pi))}{U^{\max}_{F,d,b,T} - U^{\min}_{F,d,b,T}}. \tag{30}$$

The common and decision-specific comparators therefore use the same normalization denominator within each matched decision–budget cell. No matched-random baseline or random-relative normalization is used in this CUB comparison.

Mean normalized regret weights the four decisions and three budgets equally,

$$w_{d,b} = \frac{1}{12},$$

and hence averages the resulting 4×3 cells. Worst-decision regret first averages over the three budgets within each decision and then takes the maximum across the four decisions.

The paired common-minus-specific mean-regret contrast is

$$0.3561532394 \quad [0.3028546592,\ 0.4073974664],$$

and the corresponding worst-decision contrast is

$$0.4355353913 \quad [0.3400179456,\ 0.5306786044].$$

Positive values favor decision-specific ranking because normalized regret is lower-is-better.

Development-only stronger-search control. A second control uses the same 15 outer folds but replaces the linear shared/specific rankers with development-only Kemeny-style ordering searches over the 212 concepts. Six fixed starts each receive 5,000 proposals, with temperature decreasing from .05 to .001, followed by adjacent-swap cleanup. Only development preferences enter the search; the resulting orderings are frozen before held-out validation and official-test evaluation.

The expanded evaluation menu contains nine policies. Shared and decision-specific mean regrets are respectively .4742330281 and .2513687102, giving

$$\Delta_{\mathrm{mean}} = 0.2228643180 \quad [0.1689487274,\ 0.2765084126].$$

Worst-decision regrets are .7097585204 and .5557571789, giving

$$\Delta_{\text{worst}} = 0.1540013415 \quad [0.0291516897,\ 0.2779279869].$$

This control does not claim global optimality. It asks whether the same-feature linear result disappears under a substantially stronger development-only shared-order search. Because both the search procedure and evaluation menu differ from the same-feature linear analysis, the absolute regrets across these analyses are not interpreted as points on one optimization trajectory.

Reduced exact test-utility control. The reduced exact control reuses the 15 Standard-CBM models without new training. For each evaluated seed, a validation-side proxy selects eight concepts. Official-test utilities then provide both the optimization objective and the reported score. All

$$8! = 40{,}320$$

possible orderings are enumerated for the shared ordering and for each decision-specific ordering at budgets

$$k = 1, 2, 3.$$

The shared and decision-specific utility scores are .6830079574 and .9927981664, respectively, yielding

$$\Delta_{\text{specific}-\text{shared}} = 0.3097902090 \quad [0.2906585062,\ 0.3289076790].$$

The corresponding worst-objective penalty is

$$0.6821673929 \quad [0.5808893987,\ 0.7797521656].$$

Enumeration removes optimizer error within this reduced finite empirical problem. It does not establish a full-212-concept optimum, a latent population-utility optimum, or out-of-sample performance for orderings selected using official-test utility.

All 15 seed definitions, the eight validation-selected concepts per seed, the budgets $k = 1, 2, 3$, and the enumeration runner are retained in the reproduction package. The standalone utility cache is not archived separately; replay proceeds from the frozen inputs and runner. This is a reproducibility boundary of the reduced oracle rather than a change to its estimand.

Inferential target and fold-level stability. The reported CUB intervals are paired percentile-bootstrap intervals over the 15 recorded held-out seed/model contrasts, using 10,000 draws and RNG seed 2027. The bootstrap resamples these recorded outer-fold contrasts only; it does not retrain models, refit ridge rankers, or rerun ranking selection within bootstrap replicates.

The inferential target is therefore variation of the observed outer-heldout comparison conditional on the trained model panel, the declared fold construction, and the fitted analysis procedure. It is not the sampling distribution obtained by retraining, refitting, and reselecting the complete pipeline on a newly generated model panel.

For the mean-regret contrast, all 15 recorded outer folds are positive: 15/15 are greater than zero, with no zero or negative contrasts. The fold-level contrasts range from 0.142844 to 0.522317, with median 0.387068 and mean 0.356153.

As a descriptive influence diagnostic, we delete one recorded outer-fold contrast at a time without refitting any model or ranker. The resulting leave-one-fold-out mean contrasts range from 0.344284 to 0.371390. The minimum occurs when omitting held-out seed 42, and the maximum when omitting held-out seed 46.

The worst-decision contrast shows the same sign stability: all 15 recorded folds are positive, with fold-level values ranging from 0.008950 to 0.800583, median 0.404331, and mean 0.435535. Its leave-one-fold-out mean ranges from 0.409461 to 0.466006; the minimum again occurs when omitting seed 42 and the maximum when omitting seed 46.

These leave-one-fold-out calculations are influence diagnostics on the recorded contrasts, not reruns of the outer training-and-selection procedure and not additional confidence intervals. Because the outer folds reuse overlapping development models, the 15 fitted folds are not interpreted as 15 independent training experiments.

Accordingly, the bootstrap intervals and influence diagnostics support stability of the observed held-out comparison within the declared CUB panel and comparator construction. They are not claimed to provide distribution-free coverage for the effect of retraining the complete procedure on a new model population.

D.4 Set-Level Alignment and Distinct SUN Protocols

SUN target-domain consequence study. The SUN consequence study is a retrospective target-domain outer-heldout comparison over 30 Joint and Sequential models (15 seeds per formulation, seeds 42–56). For each held-out target model, the comparator is fitted using the remaining 29 models and then evaluated on the held-out target. This protocol is distinct from the earlier zero-refit transfer study in which CUB-fitted rankers were transported to SUN without target-domain refitting.

The linear comparator uses fixed ridge regularization $\alpha_{\text{ridge}} = 1$. The near-order control uses six fixed starts and 5,000 proposals per start over the 102 SUN concepts. The two analyses are reported separately because they use different ranking classes rather than representing successive points on one optimization trajectory.

For both analyses, the primary study criterion requires support for the declared mean and worst-decision consequence endpoints. Although the mean contrasts are positive, the required worst-decision endpoint is not established. The SUN study therefore retains its original `NOT_SUPPORTED` verdict.

Uncertainty is computed using the declared formulation-by-seed hierarchical resampling scheme with 10,000 draws and RNG seed 2027. These intervals summarize variation within the recorded 30-model SUN panel under that resampling design; the bootstrap draw count does not create additional independently generated target-domain model populations.

Protocol	Endpoint	Contrast	95% CI
SUN linear	Mean	+.130022545	[.030695704,.233080728]
SUN linear	Worst	−.009197227	[−.154514474,.137903763]
SUN near-order	Mean	+.153832995	[.035909458,.252896200]
SUN near-order	Worst	+.146792738	[−.024146685,.336291359]
FSD aligned retrospective	Mean	+.247453129	[.223039023,.273772120]
FSD aligned retrospective	Worst	+.536159520	[.462875709,.619491719]
FSD historical prospective	Mean	+.063103422	[−.001454948,.131402920]
FSD historical prospective	Worst	−.097435962	[−.196011322,.032677924]

Table 12: SUN and FSD set-level consequence contrasts. The SUN linear and near-order analyses are distinct retrospective target-domain comparators. The aligned FSD result is a retrospective estimand-aligned analysis; it does not replace or revise the historical prospective `NOT_SUPPORTED` result.

FSD aligned set-level analysis. The aligned FSD analysis contains four top-level families with eight model seeds nested within each family, yielding 32 frozen models. Validation anchors and the previously frozen Set-Level Alignment V2 policies are held fixed, and cached official-test probabilities and labels supply the action utilities. No additional model training, policy fitting, threshold selection, or target-dependent recalibration is performed in this retrospective audit.

For each model, mean consequence averages the 12 declared objective–budget cells. The worst-decision endpoint first averages over budgets within each objective and then takes the largest objective-level consequence.

Inference is interpreted at the family level. The hierarchical bootstrap resamples the four recorded families and, within each sampled family, its eight seed replicates, using 10,000 draws and RNG seed 2027. Thus the top-level independent panel has

$$n_{\text{family}} = 4,$$

not 32; the eight seeds are within-family replicates, and the number of bootstrap draws does not increase the number of independent families.

Accordingly, the positive aligned FSD result is interpreted as retrospective set-level evidence within this declared four-family panel. It is not presented as precise population coverage for arbitrary new FSD families and does not overwrite the historical prospective NOT_SUPPORTED verdict.

Optimizer-quality check. The aligned FSD analysis includes 35 optimizer-quality checks, all of which satisfy the declared tolerance: the approximate solution is within 1% of the corresponding exact score on the checked instances. This check assesses optimization quality on the recorded audit instances; it is not a proof that every larger search problem is solved to global optimality.

The optimizer-quality evidence therefore supports use of the recorded set-level solutions for the retrospective aligned analysis while preserving the distinction between finite checked instances and a full-space global-optimality claim.

D.5 Functional motivation and the joint-quality boundary

Controlled redundancy provides the primary functional motivation; the known-role study in Appendix A.2 is supporting evidence, not a requirement for the present action-level comparison. In the redundancy control, one copy has single-unit and full-group drops of .1813; with 16 copies the corresponding drops are .0113 and .2072. The quantities measure conditional marginal and group dependence, respectively. They are not new demonstrations that all functional quantities are statistically independent, nor universal causal attribution scores.

The separate joint task-accuracy/Concept-AUC matching control tests a stronger cross-architecture statement. It uses 165 existing CUB models across 11 methods and 15 seeds, 825 same-seed cross-method candidate pairs, and 165 retained matched pairs. Against the outcome-free same-method seed reference, none of five endpoints has the positive excess required by the frozen success rule (0/5). All five reported excess intervals are below zero. This does not establish that quality causally determines functional organization; it limits the claim of irreducible cross-method divergence under that matching analysis. It does not replace or invalidate within-model shared-versus-specific ranking tests. The full mechanism and matching analyses belong with the population-diagnostics appendix; they are not independent replications of the CUB ranking comparison.

E Measurement, Replication, and Evidence Roles

E.1 Non-CBM Consequence Results and Scale Dependence

Five-dataset panel and split contract. The non-CBM replication covers five OpenML datasets: 28 (optdigits), 60 (waveform-5000), 182 (satimage), 250 (BNG(mfeat-fourier)), and 252 (BNG(mfeat-karhunen)). The panel contains 15 dataset-by-regularization families with three seed replicates per family, yielding 45 trained models. These are therefore not 15 unrelated algorithm families.

Within each seed, the data are partitioned into 55% train, 15% optimization, 15% audit, and 15% confirmation splits. Shared and decision-specific orderings are fitted on the optimization split, the audit split supplies the audit measurement, and the confirmation split supplies the final held-out evaluation. This is split-separated replication on fixed declared datasets rather than an untouched-dataset prospective study.

Shared and decision-specific searches use the same five starts, four rounds, and 48 deterministic swap proposals per round. The optimizer audit additionally contains ordered-prefix exhaustive checks on deterministic eight-feature subsets. Uncertainty is computed by resampling the five datasets as clusters for 2,000 bootstrap draws with seed 2027; the three seeds remain within-family replicates.

The pooled raw sharing contrast is

$$0.008729804 \quad [0.003252703,\ 0.014206904],$$

whereas the row-normalized confirmation contrast is

$$-45.149204088 \quad [-94.151578502,\ -0.002243975].$$

The latter interval is wholly negative and is therefore not a near-zero positive-consequence result.

Dataset-level inference scope. The five OpenML datasets are the top-level clusters for this analysis: 28, 60, 182, 250, and 252. The 2,000 dataset-cluster bootstrap draws resample this recorded five-dataset panel; they do not increase the number of independent dataset clusters beyond $n_{\text{dataset}} = 5$.

The three model seeds within each dataset-by-regularization family are within-family replicates rather than additional dataset-level units. Accordingly, the non-CBM result is interpreted as split-separated replication across these five declared datasets, not as a precise population guarantee over arbitrary new tabular domains.

The raw-positive and normalized-negative conclusions are therefore reported together with the denominator, floor, and leave-one-dataset-out sensitivity analyses rather than extrapolated beyond the observed five-dataset panel.

Raw and normalized consequences. The normalized endpoint does not satisfy the positive-consequence criterion. This does not contradict the positive raw aggregate because the empirical quantity is a signed fitted-comparator contrast rather than the nonnegative oracle penalty C^*_Π.

Let x_r be the raw contrast numerator for confirmation row r and let $a_{r,T} > 0$ denote its declared split-specific normalization scale. Normalization is performed rowwise:

$$c_r^{\text{norm}} = \frac{x_r}{a_{r,T}}, \qquad \bar{c}^{\text{norm}} = \frac{1}{135}\sum_{r=1}^{135} \frac{x_r}{a_{r,T}}, \tag{31}$$

rather than by dividing an aggregate numerator by a single aggregate scale.

Because every denominator is positive, normalization preserves the sign of each individual row. Across rows, however, inverse-denominator weighting changes the relative contribution of positive and negative signed contrasts. The raw and normalized aggregates therefore answer different weighted empirical questions.

The confirmation-scale distribution has minimum 10^{-6}, median .148962, maximum .484629, three near-zero denominators, three floor/clipping hits, and no negative denominators. The three floor hits use a floor of 10^{-6}.

Waveform-5000 contains two of the floor hits and satimage contains one. Their minimum normalized rows are -3296.296 and -2792.170, respectively, showing how near-zero scales amplify magnitude. However, datasets 250 and 252 also have negative normalized means without floor hits. Thus floor amplification explains the extreme magnitude but is not necessary for sign reversal; the broader issue is rowwise scale reweighting of signed contrasts.

No alternative floor or normalization rule is selected after observing these outcomes.

Leave-one-dataset-out sensitivity. Removing one dataset at a time leaves the raw aggregate positive: .0074, .0091, .0064, .0105, and .0102 for deletions of datasets 28, 60, 182, 250, and 252, respectively. The corresponding normalized means remain negative: -56.44, -25.80, -30.64, -56.43, and -56.42.

These deletions are point-estimate sensitivity checks rather than independent replications or new confidence intervals. They show that no single reported dataset deletion removes the raw-positive/normalized-negative distinction, while the magnitude of the normalized endpoint remains sensitive to scale.

The consequence results above and the packaged E29 V4 unit-row measurement audit have different aggregation contracts; quantities from the two are therefore not combined into a single measurement endpoint.

Separate packaged E29 V4 unit-row audit. Unlike the five-dataset consequence panel above, the final packaged E29 V4 measurement audit covers OpenML datasets 28, 60, and 182. It contains

$$3 \text{ datasets} \times 3 \text{ split seeds} \times 3 \text{ model families} \times 5\ C\text{-values} = 135$$

unit rows.

$$A^{\text{raw}}_{\text{pair}} = 0.8189054726, \qquad A^{\text{norm}}_{\text{pair}} = 0.8561636263.$$

These values are distinct from the legacy $A_{\text{pair}} = 0.9333333333$ manuscript value. Auditing the current canonical package does not reproduce 0.9333333333 under the tested 15-row or 45-row collapse schemes; the closest audited 15-row collapse is 0.9238095238. Because the artifact producing the legacy value has not been recovered, we exclude 0.9333333333 from active quantitative claims rather than merge it with the current packaged aggregation.

The unit-row pair accuracies are descriptive measurement diagnostics. They are not pooled with the family-level AwA2/CelebA results, and they do not alter the raw-positive/normalized-negative consequence analysis above.

Excluded archival development panel. An earlier development analysis (E28) is excluded from the present quantitative evidence set because its original protocol, canonical results, bootstrap output, and complete reconstruction artifacts are unavailable in a form that supports reconstruction of the reported endpoint. No E28 numerical value is used in the present claims or pooled with the E29 results.

E.2 Intervention-Priority Descriptors and Head Compatibility

This supporting study evaluates SUN concept correction rather than prediction of sharing cost. The evaluated setting uses soft correction at $k = 20$ with the predicted-concept head under the Sequential CBM.

The mean intervention gains are:

$$0.02806\ [0.02315, 0.03298]$$

for Contribution,

$$0.02896\ [0.02504, 0.03288]$$

for Concept AUC,

$$0.01384\ [0.00651, 0.02116]$$

for Usage, and

$$0.02073\ [0.01753, 0.02392]$$

for Random.

At this operating point, Concept AUC and Contribution have similar point estimates, whereas Usage has a smaller point estimate than Random. The marginal intervals do not establish paired superiority or equivalence among ranking rules. Changing the downstream concept head also changes the evaluated model, so these results should not be interpreted as an isolated causal effect of the ranking rule or of the head architecture.

The study is therefore used only as supporting evidence that intervention-priority descriptors are setting-dependent. It is not an additional audit-to-confirm replication and is not used to establish sharing-cost predictability.

Structural-proxy analysis. An earlier structural-proxy analysis is not included in the quantitative comparison because its previously quoted prediction-error endpoint could not be mapped to a canonical result artifact. We therefore exclude that numerical endpoint from all active claims. This provenance limitation does not affect the SUN intervention results above or the independent audit-to-confirm measurements reported in the following sections.

Ranking	Mean intervention gain	95% CI
Contribution	.02806	[.02315,.03298]
Concept AUC	.02896	[.02504,.03288]
Usage	.01384	[.00651,.02116]
Random	.02073	[.01753,.02392]

Table 13: SUN predicted-concept head, soft correction at $k = 20$. Each row reports an intervention gain in the stated setting, not a gain relative to Random and not a sharing-cost prediction error. The marginal intervals are not paired comparison intervals.

E.3 Measurement contracts, evidence roles, and reported endpoints

Target and prediction. The target is the finite-data sharing contrast $\widehat{C}_{\Pi,T}$ defined in Section 3, with the original comparator policies, actions, budgets, and normalization rule. For the prespecified blind AwA2 study and the separately frozen post-hoc exploratory CelebA replication, the prediction is the audit measurement itself, $\widetilde{C}_{\text{confirm}} = \widehat{C}_{\Pi,\text{audit}}$. The confirmation measurement is evaluated without fitting a new mapping or recalibrating this prediction on confirmation outcomes. A high correlation alone would establish only ordering association; RMSE and MAE additionally report numerical disagreement on each study's own scale.

AwA2. AwA2 uses 37,322 examples with a frozen train/setopt/audit/confirm split of 22,392/5,599/4,665/4,666. Its no-bypass model predicts 85 concepts from fixed features and feeds them to a linear 50-class head. Ten families with five seeds each yield 50 models. An initial feasibility version stops because the Sequential head is underoptimized; the reported pre-functional revision changes its head training from 8 to 30 epochs while retaining thresholds, splits, and concept training choices. Models, rankings, validation anchors, audit values, and policy decisions are frozen before confirmation utilities are opened. All 10 families and 50 models are retained in the reported result. The reported MAE improvement over the frozen constant is $[.0609, .1589]$ at 95% confidence. The audit-to-confirm measurement endpoint satisfies its declared criterion, while the broader AwA2 study remains PARTIAL because its separate complexity-savings criterion is not satisfied. Measurement success is therefore not treated as an all-endpoints success verdict.

CelebA. CelebA is a post-hoc exploratory replication under a separately frozen protocol; it is not Level-3 confirmatory evidence. It contains eight families with four seeds each, hence 32 models. Its model panel, data splits, ranking optimizer, budgets, anchors, audit values, predictions, and constant baseline are frozen before confirmation unlock. All 32 models are retained. This is an exploratory independent-split measurement study on semantic attributes, not an ordinary feature-ranking counterpart to a concept-level AwA2 experiment.

Metric	AwA2	CelebA
A_{pair}	.9560	.9290
Pearson	.9660	.9530
Spearman	.9760	.9520
Audit RMSE	.0230	.0238
Constant RMSE	.1514	.0893
Audit MAE	.0203	.0176
Constant MAE	.1301	.0796
Calibration slope	.9370	1.2220

Table 14: Reported endpoints on study-specific scales. No pooled estimate is formed.

Pair-ordering accuracy. Let a_i and c_i denote the family-level audit and confirmation sharing contrasts. Each study retains its frozen pair-eligibility and tie convention rather than retrospectively harmonizing them. For AwA2, confirmation-tied pairs are excluded. For every remaining pair (i, j),

let

$$s_{ij} = \begin{cases} 1, & (a_i - a_j)(c_i - c_j) > 0, \\ 1/2, & a_i = a_j,\ c_i \neq c_j, \\ 0, & (a_i - a_j)(c_i - c_j) < 0. \end{cases}$$

Then

$$A_{\text{pair}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} s_{ij}.$$

For CelebA, the frozen implementation assigns score one iff

$$(a_i - a_j)(c_i - c_j) > 0,$$

and zero otherwise. Thus audit ties do not receive half credit under the CelebA convention.

These slightly different tie conventions are historical study definitions and are retained explicitly rather than silently standardized.

Reported uncertainty and inference units. For AwA2, $A_{\text{pair}} = .956\,[.842, 1.000]$; Pearson and Spearman permutation p-values are .00001 and .00003. For CelebA, $A_{\text{pair}} = .929\,[.750, 1.000]$, Pearson is $.953\,[.871, .996]$, Spearman is $.952\,[.620, 1.000]$, and the calibration slope is $1.222\,[.711, 1.896]$; permutation p-values are .00012 and .00048. AwA2 Pearson/Spearman confidence intervals are not available for this analysis, so none are synthesized from a presumed sample count. Family pairs, seeds within a family, and repeated measurements are not counted as independent family populations. A_{pair} follows the original eligible-pair and tie rules. Exact reproduction additionally requires the frozen family/model mappings, eligibility masks where applicable, original resampling code, and frozen baselines.

Non-CBM unit-row measurement diagnostics are reported separately in Appendix E.1 because their aggregation contract differs from the family-level AwA2/CelebA studies.

Interpretation. These studies support repeatability of a protocol-defined comparison across independent evaluation samples. They do not establish an unbiased estimate of a latent infinite-data utility, an optimal sharing penalty C^*_Π, a normalization-invariant quantity, or a portable absolute deployment threshold. Same-estimand repeatability and learned cross-domain prediction are different evidential roles.

E.4 Full audit-budget distributions on fixed panels

The archived audit-budget analysis covers all six declared fractions: 5, 10, 20, 40, 60, and 100%. Each reduced fraction uses 1,000 fixed-seed subsamples per family, while the full-audit record is fixed. Models, rankings, anchors, family definitions, and confirmation observations remain unchanged. The analysis contains 12 domain-by-fraction summaries and 90,018 family-level audit-cost rows. These repeated observations do not increase the number of independent families. The prespecified blind AwA2 panel and the post-hoc exploratory CelebA panel retain their distinct evidence roles.

Median criteria versus run-level reliability. The medians increase across the six reported fractions in each panel; no monotonicity is claimed for every subsampling run or untested fraction. Forty percent is the first declared fraction satisfying both median criteria in both panels: $A_{\text{pair}} \geq .85$ and at least 90% of full-audit accuracy. However, only 67.7% of AwA2 and 60.0% of CelebA subsampling runs meet .85 there. The 5th–95th percentile ranges at 40% are [.7778,.9556] and [.7143,.9643], respectively. At 60%, pass frequencies increase to 82.9% and 87.5%; neither tested reduced-budget operating point establishes a 95%-of-runs criterion. The .85 line concerns ordering accuracy, not false-share risk, family coverage, or reuse authorization.

Precision and full-budget values. Using the displayed full-grid summaries, the 40%/100% ratios are $.8889/.9556 \approx 93.02\%$ and $.8571/.9286 \approx 92.30\%$. The main text rounds these to approximately 93% and 92% rather than claiming unavailable full-precision ratios. Earlier abbreviated summaries used .9560 and .9290 for the full-audit values; the new grid reports .9556 and .9286.

Panel	Audit	$Q_{.05}$	$Q_{.25}$	Median	$Q_{.75}$	$Q_{.95}$	$\widehat{p}_{.85}$
AwA2	5%	0.6433	0.7111	0.7778	0.8222	0.8889	0.113
AwA2	10%	0.6889	0.7556	0.8000	0.8667	0.9111	0.256
AwA2	20%	0.7333	0.8000	0.8444	0.8889	0.9333	0.418
AwA2	40%	0.7778	0.8444	0.8889	0.9111	0.9556	0.677
AwA2	60%	0.8222	0.8667	0.9111	0.9333	0.9778	0.829
AwA2	100%	0.9556	0.9556	0.9556	0.9556	0.9556	1.000
CelebA	5%	0.4286	0.5714	0.6786	0.7500	0.8571	0.060
CelebA	10%	0.5000	0.6429	0.7143	0.7857	0.8929	0.136
CelebA	20%	0.6071	0.7143	0.7857	0.8571	0.9286	0.274
CelebA	40%	0.7143	0.8214	0.8571	0.8929	0.9643	0.600
CelebA	60%	0.8214	0.8571	0.8929	0.9286	0.9643	0.875
CelebA	100%	0.9286	0.9286	0.9286	0.9286	0.9286	1.000

Table 15: **Full audit-budget distributions on fixed panels.** Quantiles and $\widehat{p}_{.85}$, the empirical fraction with $A_{\text{pair}} \geq .85$, are the archived subsampling summaries. They do not give population confidence bounds. The full-audit row has no subsampling variation. AwA2 is prespecified and blind; CelebA is an exploratory replication.

Table 2 retains the rounded full-study metrics and their separately reported family-level intervals. Ratios of A_{pair} are not chance-corrected information retention.

What the bands and frequencies describe. The quantiles and pass frequencies summarize audit-sampling variation conditional on the fixed models, rankings, anchors, and confirmation panel, not uncertainty over new families or a new deployment distribution. At 100%, every displayed quantile coincides and $\widehat{p}_{.85} = 1$ because the full audit is fixed. This zero subsampling variance does not remove the family-level uncertainty of the full study.

The archived replicate summaries support these distributional endpoints, while exact low-level replay additionally requires the original eligibility masks, family aggregation, tie rules, and synchronized replicate identifiers.

Panel	Audit	Spearman	RMSE	Inversion
AwA2	20%	.8182	.0458	.1556
AwA2	40%	.8788	.0332	.1111
AwA2	100%	.9760	.0230	.0444
CelebA	20%	.7143	.0421	.2143
CelebA	40%	.8571	.0316	.1429
CelebA	100%	.9520	.0238	.0714

Table 16: Auxiliary operating-point metrics at 20%, 40%, and 100% audit fractions.

The six-fraction analysis provides complete A_{pair} distribution summaries. The auxiliary Spearman, RMSE, and inversion metrics are available only at 20%, 40%, and 100% and are not imputed at unreported fractions.

E.5 LAD Normalization and Common-Factor Sensitivity

We conduct a secondary post-hoc sensitivity analysis on the 500-family LAD confirmation panel to examine whether the observed audit–confirmation associations can be explained entirely by recorded background variation or normalization-scale effects. This analysis does not modify the frozen LAD V4B experiment or its deployment policies.

We distinguish three quantities throughout. The *Direct risk score* is the output of the fitted Direct risk model; *audit raw* is the unnormalized audit measurement; and *audit normalized* is the corresponding normalized audit measurement. These quantities have different roles and are not treated as interchangeable predictors or measurements.

Comparison	Pearson	Spearman
Direct risk / normalized CONF	.9165	.9276
Direct risk / raw CONF	.7399	.7691
Audit normalized / normalized CONF	.8873	.8824
Audit raw / raw CONF	.9798	.9795
Denominator / normalized CONF	.6195	.7224
Denominator / Direct risk	.6820	.7619
Direct / normalized CONF, residual	.6665	.6594
Audit normalized / CONF, residual	.5836	.5859
Audit raw / raw CONF, residual	.8091	.7835

Table 17: Post-hoc LAD sensitivity analysis on the 500-family confirmation panel. The first six rows report raw associations among fitted risk scores, audit measurements, confirmation measurements, and normalization denominators. The final three rows report associations after residualizing the corresponding variables against the declared common-factor controls. These analyses are secondary sensitivity checks rather than primary estimands.

Common-factor adjustment. The residual analysis adjusts for the recorded model- and training-level covariates used in the sensitivity analysis: task accuracy, concept AUC, anchor-only, sequential and low-rank configuration indicators, concept-loss weight, concept dropout, concept-label noise, concept-supervision fraction, intervention probability, normalization equalization, learning rate, and weight decay. It additionally includes the sum, mean, minimum, and maximum of the confirmation-denominator statistics.

The adjustment set is used to test a specific alternative explanation: that the observed association is induced entirely by these recorded background and denominator variables. It is not assumed to contain every feature used by the separately fitted structure-only predictor, and conditioning on confirmation-side denominator summaries is used only for explanatory sensitivity analysis, not as a deployable feature specification.

Before adjustment, raw audit and raw confirmation measurements are strongly associated,

$$r = .979772,$$

whereas the association between the Direct risk score and raw confirmation is

$$r = .739908.$$

These are different comparisons and should not be interpreted as alternative estimates of the same quantity.

After adjustment for the declared common-factor controls, the raw-audit/raw-confirmation association remains

$$r = .809099,$$

and the Direct-risk/normalized-confirmation association remains

$$r = .666472.$$

The corresponding rank associations are also positive, as reported in Table 17.

These residual associations weaken the explanation that LAD audit–confirmation agreement is solely induced by the recorded background or denominator variables. They do not identify a causal effect, establish independence from unmeasured structure, or imply invariance to arbitrary normalization choices.

Normalization-scale diagnostics. Because the LAD consequence is normalized by family- and objective-specific scales, we additionally inspect the empirical distribution of the anchor denominators.

The denominator is itself associated with normalized confirmation cost (Spearman $\rho = .722373$) and with Direct risk ($\rho = .761945$). Vulnerability screening has the highest low-range frequency under both anchor sources, with fractions .729333 and .723333. Normalization scale is therefore a meaningful source of variation in this panel and is not treated as negligible.

Anchors	Objective	Median	Minimum	Low-range fraction
A	Correction	.01605	.01	.3640
A	Lesion diagnostic	.03319	.01	.3187
A	Specialization	.01373	.01	.1587
A	Screening	.01000	.01	.7293
B	Correction	.01685	.01	.3347
B	Lesion diagnostic	.03461	.01	.3040
B	Specialization	.01390	.01	.1233
B	Screening	.01000	.01	.7233

Table 18: LAD normalization-scale diagnostics. Each anchor–objective row contains 1,500 observations. The low-range indicator follows the frozen diagnostic definition and is reported separately from the observed minimum; it is not interpreted as an exact floor-hit count.

Score	Features	Pearson	Spearman	Overall regret
Background only (post-hoc)	15	.78532	.77788	.04182
Structure only (post-hoc)	8	.93867	.94447	.03778
Background + structure (post-hoc)	23	.94044	.94419	.03773
Direct (frozen)	123	.91647	.92758	.03901
Cheap (frozen)	29	.91335	.92426	.03923

Table 19: LAD score comparisons at the same 20% reuse quota ($N = 500$, $M = 100$). The first three rows are secondary post-hoc score models; Direct and Cheap are frozen deployment policies. Point differences across these rows are not by themselves causal attributions to individual feature groups.

At the same time, the persistence of substantial residual association after adjustment shows that the observed audit–confirmation relation is not explained solely by the recorded denominator summaries. This is a within-LAD sensitivity result. It does not establish normalization invariance, does not transfer the same adjustment to AwA2 or CelebA, and does not replace the separate non-CBM normalization analysis in Appendix E.1.

E.6 Predictive Substitutability versus Incremental Decision Value

The LAD decision-value comparison holds the reuse quota fixed at 20% on the 500-family confirmation panel. Each score therefore selects 100 families under the declared tolerance $\epsilon = .40$. Post-hoc score models and the frozen Direct/Cheap policies retain distinct evidential roles.

The background-only policy has four tolerance violations among its 100 selected families. The structure-only, background-plus-structure, Direct, and Cheap rows each have zero violations among their 100 selected families. These counts do not establish a statistically significant superiority contrast between the corresponding policy classes.

The frozen Direct policy uses `linear_q90_a001` at the 40% audit stage, whereas Cheap uses `linear_q95_a001` without direct-audit-stage features. Their confirmation regrets are

$$J_{\text{Direct}} = .0390107779, \qquad J_{\text{Cheap}} = .0392294094,$$

giving

$$J_{\text{Direct}} - J_{\text{Cheap}} = -0.0002186315.$$

Paired bootstrap. The paired confidence interval conditions on the realized frozen Direct and Cheap decisions. The resampling unit is the confirmation family ($N = 500$). For each of $B = 10{,}000$ bootstrap replicates, confirmation-family rows are sampled with replacement and the saved `direct_decision` and `cheap_decision` masks are carried with the sampled rows. A family drawn multiple times therefore contributes with the corresponding bootstrap multiplicity.

The policy scores are not re-sorted within bootstrap replicates and the 20% quota is not reapplied. Consequently, this bootstrap quantifies uncertainty in the regret contrast conditional on the two

frozen realized policy masks; it is not an estimate of the sampling distribution of refitting or rerunning the fixed-quota selection procedure.

The resulting 95% paired family-bootstrap interval is

$$[-0.0057686657,\ 0.0053889306].$$

Because this interval contains zero, the analysis does not establish an incremental regret advantage for Direct over Cheap. It also does not establish equivalence, because no equivalence margin or equivalence test was prespecified.

Equal reuse quota does not imply equal measurement resources, and the two frozen policies use different fitted risk models and input schemas. The comparison is therefore a matched-quota deployment-policy comparison rather than a controlled causal estimate of the value of one feature group.

The study records audit fractions, sample counts, and model counts but does not measure wall-clock, GPU-time, monetary, or human-review costs. No end-to-end efficiency conclusion is inferred from these proxies.

E.7 Historical prediction targets are not extra audit replications

Earlier CUB/SUN/FSD/Civil studies predict vulnerability ordering from a frozen participation-based mapping. They are not identity predictions of the sharing contrast measured on another split. These vulnerability-prediction results do not constitute additional replications of identity-based sharing-cost measurement and are not pooled into Table 2.

Panel	A_{pair} [95% CI]	Evidential role
CUB original	.698 [.400, .900]	Blind discovery, imprecise
CUB V2	.5511 [.4687, .6455]	Confirmation attempt, unresolved
Civil	.8444 [.668, .884]	Frozen external confirmation
SUN fresh	.6159 [.4689, .7289]	Post-unlock seed reproduction
FSD50K V2	.5967 [.4718, .6877]	Coarse regime; quality overlap

Table 20: Historical vulnerability-prediction evidence, kept separate from sharing-cost measurement. Intervals are retained on their original study-specific scales.

F Selective Reuse and Authorization Details

F.1 Policy forms, measured losses, and denominators

The deployment comparison and certificate comparison have different targets. Let $c_i = \widehat{C}_{\Pi,\text{confirm}}(F_i; \pi_s, \boldsymbol{\pi})$ denote the protocol-specific signed contrast of Section 3.1. For a reported policy mask g_i, its overall additional sharing regret is

$$J = \frac{1}{N}\sum_{i=1}^{N} g_i c_i = \frac{M}{N}\,\overline{c}_{\text{SHARE}} \quad (M > 0).$$

SPECIALIZE contributes zero additional sharing contrast relative to the same decision-specific reference. This is not a claim of zero task error, zero intervention cost, or optimal decision-specific performance. At $M = 0$, $J = 0$ by construction, but the conditional violation rate E/M and $R_\epsilon(g)$ are undefined. The symbol C^*_Π is reserved for the optimal-value difference; the deployed fitted-policy contrast need not equal it.

For a batch policy, g_i can depend on scores of other batch members. An exact 20% quota specifies the realized fraction selected, not a lower confidence bound on the coverage of a fixed per-family map. The bounds for a frozen per-family policy in Section 3.4 therefore do not automatically extend to such a batch rule. Numeric risk limits carried by historical reports are retained with their original roles, not converted into new simultaneous guarantees. No risk rates, tolerances, or regret scales are pooled across LAD, CEBaB, V6, and the finite-family certificate studies.

The below-tolerance FIT count is also a different denominator: it counts families satisfying the measured tolerance, regardless of which threshold selects them. In the final summaries, entries such as "69 safe / 150" must not be read as $M = 69$ decisions made by the stopped policy. Likewise a safe fraction or fixed quota is not L_Q merely because it appears in a general coverage column.

F.2 LAD V4B: Held-Out Fixed-Quota Deployment

Dataset and attribute vocabulary. The LAD study uses the Large-scale Attribute Dataset (LAD), containing 230 classes and an original annotation vocabulary of 359 attributes. Before any FIT/CERT/CONF sharing-cost outcome is accessed, we construct a fixed 85-attribute vocabulary using only the official training and validation annotations. We retain attributes whose empirical prevalence lies in $[0.05, 0.95]$ and whose certainty is at least 0.50, rank the remaining attributes by entropy times certainty, break ties by source attribute ID, and retain the top 85. The official LAD test split is not used for this selection.

The resulting source attribute IDs, names, and used-column indices are frozen for all subsequent families and are provided in the supplementary artifact. We do not infer any relationship to the AwA2 85-attribute vocabulary from the coincident dimensionality.

Visual backbone and feature extraction. We instantiate the torchvision ResNet-50 architecture with `weights=None` and manually load the checkpoint `resnet50-11ad3fa6.pth`, whose archived provenance identifies it as ImageNet-1K pretrained. We replace the final fully connected layer with the identity, yielding 2048-dimensional visual features.

Each image is cropped to its official LAD bounding box, clipped to the image dimensions, resized to 256 pixels, center-cropped to 224×224, converted to a tensor, and normalized using ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. The exact torchvision pretrained-weight enum was not retained, so we report the archived checkpoint identifier rather than infer an enum retrospectively.

Pre-unlock scope. Feature extraction before protocol unlock covers 54,610 images. The 23,407 official-test images remain locked at that stage, and no pre-unlock official-test features are computed. Thus neither the 85-attribute vocabulary nor the cached pre-unlock visual features are selected using official-test outcomes.

Population construction. A family denotes one frozen model-and-training configuration rather than one individual training seed. All LAD models use the fixed 85-dimensional semantic bottleneck without an additional latent bypass.

Families vary in formulation (`joint` or `sequential`), concept-head type (`dense`, `lowrank16`, `lowrank24`, `lowrank32`, or `lowrank48`), and eight continuous training or intervention factors: concept-loss weight in $[0.35, 2.00]$ on a log scale, concept dropout in $[0, .30]$, concept-label noise in $[0, .10]$, concept-supervision fraction in $[.65, 1]$, intervention probability in $[0, .45]$, normalization equalization in $[.004, .045]$ on a log scale, learning rate in $[3 \times 10^{-4}, 1.2 \times 10^{-3}]$ on a log scale, and weight decay in $[10^{-6}, 3 \times 10^{-4}]$ on a log scale.

The eight continuous factors are generated by a fixed 1,000-point Latin-hypercube design with generator seed 20271019. Each family is instantiated using three training-seed replicates, 2800, 2801, and 2802, yielding 3,000 trained models from 1,000 family configurations. The three models within a family are therefore repeated trainings of the same family configuration rather than three distinct architectures.

Family indices 0–199 form FIT, 200–499 form protocol-CERT, and 500–999 form CONF. This partition is fixed by generator index.

Policy form and inferential scope. LAD evaluates a fixed-quota batch policy rather than the frozen per-family threshold policy covered by the certificate in Section 3.4. The declared reuse quota is .20 and the tolerance is $\epsilon = .40$. Direct and Cheap therefore each select exactly 20% of the evaluated batch for SHARE.

The realized reuse fraction is fixed by the batch rule rather than estimated as the population coverage of an independently applied per-family map. Accordingly, LAD is interpreted as held-out deployment evidence. Its fixed-quota counts and selected-subset binomial summaries are not reinterpreted as Section 3.4 population risk or coverage certificates.

Prediction target. Direct and Cheap predict the same frozen family-level calibration target. For family F,

$$Y_F^{\mathrm{cal}} = \mathrm{BalancedCost}(\mathrm{SharingCosts}(D_{\mathrm{cal}}; \pi_{\mathrm{shared}}, \boldsymbol{\pi}_{\mathrm{specific}}, A_B), A_B)\,,$$

where D_{cal} is the frozen `calibration_target` split and A_B denotes the frozen anchor-B normalization contract. No CERT or CONF sharing-cost outcome enters target construction, risk-model fitting, or FIT model selection.

The feature `setopt_crossfit_target` is a pre-audit proxy feature, not the supervised response:

$$x_F^{\mathrm{xfit}} = \frac{1}{2}\left[C(A \mid B) + C(B \mid A)\right],$$

where $C(A \mid B)$ denotes the cost from anchor-A quantities evaluated under anchor-B normalization, and conversely for $C(B \mid A)$. The occurrence of "target" in the feature name does not indicate use of CERT or CONF target outcomes.

At deployment, lower predicted risk is safer. Families are ordered lexicographically by

$$(\text{risk score}, \text{family ID}),$$

and the lowest-scoring 20% of the evaluated batch are assigned SHARE. This deterministic tie break yields exactly 60/300 selections on protocol-CERT and 100/500 on CONF.

Direct and Cheap inputs. Direct and Cheap are trained against the same `calibration_target`, evaluated on the same family population, and deployed at the same 20% quota. They differ in the information available to the fitted risk model and in the frozen model selected on FIT, rather than in the prediction target, deployment population, or reuse fraction.

The selected Cheap policy is `linear_q95_a001` ($q = .95$, $\alpha = .01$). It uses the following 29 pre-audit inputs:

- `task_accuracy`, `concept_auc`, `anchor_only`, `old_gamma`, `local_gamma`, and `setopt_crossfit_target`;
- `concept_loss_weight`, `concept_dropout`, `concept_label_noise`, `concept_supervision_fraction`, `intervention_probability`, `norm_equalization`, `learning_rate`, and `weight_decay`;
- `sequential`, `lowrank16`, `lowrank24`, `lowrank32`, and `lowrank48`;
- `anchors_a_denom_mean`, `anchors_a_denom_std`, `anchors_a_denom_min`, `anchors_a_denom_max`, `anchors_a_denom_sum`, `anchors_b_denom_mean`, `anchors_b_denom_std`, `anchors_b_denom_min`, `anchors_b_denom_max`, and `anchors_b_denom_sum`.

The selected Direct policy is `linear_q90_a001` ($q = .90$, $\alpha = .01$) at the 40% audit stage. It contains the same 29 pre-audit inputs plus audit-derived features available from the 20% and 40% audit stages. The audit-derived quantities include `c_audit_balanced`, `c_lesion_pruning`, `c_concept_correction`, `c_vulnerability_screening`, `c_specialization_audit`, `objective_mean_std`, and `objective_mean_range`, with the applicable summaries drawn from mean, standard deviation, q_{50}, q_{80}, q_{90}, q_{95}, minimum, maximum, and range.

The selected Direct model contains 123 inputs in total and does not use 60%-audit features. The exact expanded 123-column schema is provided in the supplementary artifact.

Stage	Policy	N	M	E	Overall J
CERT	Direct	300	60	0	.040653
CERT	Cheap	300	60	0	.041059
CONF	Direct	500	100	0	.039011
CONF	Cheap	500	100	0	.039229

Table 21: LAD stage-specific fixed-quota results. M is the number of SHARE decisions and E the number whose measured confirmation contrast exceeds $\epsilon = .40$. Because selection is a batch fixed-quota rule, these counts are not treated as Section 3.4 per-family population certificates.

Held-out results. The protocol-CERT panel contains 300 families and the CONF panel contains 500 families. At the fixed 20% quota, both Direct and Cheap select 60/300 families on protocol-CERT and 100/500 families on CONF.

On CONF, Always Shared has overall additional sharing regret

$$J_{\text{always}} = .3625307066$$

and exceeds the tolerance on 217 of 500 families. Direct and Cheap obtain

$$J_{\text{Direct}} = .0390107779, \qquad J_{\text{Cheap}} = .0392294094.$$

The Always-Shared-minus-Direct contrast is

$$0.3235199286 \quad [0.3078380246,\ 0.3388765420],$$

where the interval uses the fixed-mask paired family bootstrap described below.

Matched-quota random reference. The Always-Shared-to-Direct comparison combines lower reuse frequency with nonrandom family selection. On the fixed 500-family CONF panel, uniform selection of exactly 100 families gives each family selection probability .2, and hence

$$\mathbb{E}[J_{\text{random},20\%}] = 0.2 J_{\text{always}} = 0.0725061.$$

Direct attains $J = .0390108$, which is 0.0334954 lower, or 46.2% below this matched-quota random expectation.

Likewise, with 217 violating families under Always Shared, a uniform random 100-family subset has expected violation count 43.4, whereas Direct observes $0/100$. These are conditional references on the fixed CONF panel, not prospective population-risk certificates. They show that the observed Direct subset improves on what would be expected from reducing reuse to 20% alone, without establishing Direct superiority over Cheap.

At the same 20% quota, the Direct-minus-Cheap contrast is

$$-0.0002186315 \quad [-0.0057686657,\ 0.0053889306].$$

This establishes neither an incremental advantage for Direct nor equivalence.

Fixed-panel randomization distribution. To characterize variation around the matched-quota reference, we uniformly sample 100 of the 500 CONF families without replacement while holding the observed family-level confirmation contrasts fixed. Across $B = 100{,}000$ draws (RNG seed 2027), the random-quota regret has mean 0.07251369 and empirical 2.5–97.5% interval $[0.06839399,\ 0.07654652]$. None of the 100,000 sampled random subsets attains regret as low as Direct (0/100,000; Monte Carlo resolution 10^{-5}).

For tolerance violations, the corresponding count follows $\text{Hypergeometric}(500, 217, 100)$, with expectation 43.4. The exact fixed-panel probability of selecting zero violating families is

$$\frac{\binom{283}{100}}{\binom{500}{100}} = 1.630 \times 10^{-29}.$$

These quantities describe uniform random 20%-quota selection on the realized CONF panel; they are not population-risk certificates or uncertainty from regenerating families.

Bootstrap semantics. The reported LAD policy intervals use a paired family bootstrap over the 500 recorded CONF families with 10,000 draws and RNG seed 2027. Each resampled family row carries its observed confirmation contrast together with the previously realized Direct and Cheap SHARE indicators. Rows are sampled with replacement, and repeated rows contribute according to their bootstrap multiplicity.

The bootstrap does not refit either risk model, recompute policy scores, rerank the resampled families, or reapply the 20% batch quota within each bootstrap replicate. The 20% quota therefore characterizes the original frozen CONF deployment; an individual bootstrap replicate need not itself contain exactly 100 selected row instances because the frozen selection indicators are resampled together with their family rows.

For the Always-Shared-minus-Direct and Direct-minus-Cheap contrasts, the same resampled family indices are used for both terms within each replicate, preserving the paired comparison. The reported intervals therefore quantify variation of the realized fixed-mask policy contrasts on the recorded CONF panel, conditional on the frozen Direct and Cheap selections. They do not estimate uncertainty from regenerating families, refitting the risk models, rescoring a new batch, and rerunning top-quota selection.

Interpretation and resource boundary. The full Always-Shared-to-Direct reduction combines the effect of lowering reuse from 100% to 20% with the effect of selecting which families remain shared; it is therefore not itself an isolated selector effect. The matched-quota random reference separately shows that the observed Direct subset has lower sharing regret than expected from quota reduction alone.

The study does not measure wall-clock time, GPU time, monetary cost, or human-review cost. Audit fractions, reuse rates, and feature availability are therefore not interpreted as end-to-end efficiency or resource-savings measurements. The reported LAD endpoints concern decision value under the declared sharing-cost contract.

F.3 Transport, finite-sample resolution, and exploratory transfer

AwA2 transported threshold. The threshold $\tau = .1640390867563019$ transported from the pilot assigns SPECIALIZE to all ten unseen AwA2 families. Overall sharing regret becomes zero from $.267381$ under Always Shared, but shared coverage and complexity savings are also zero.

The measurement endpoint is supported, but the transported threshold policy remains NOT SUPPORTED—NO SHARED COVERAGE. No separate finite CERT stage is claimed.

V6 fresh follow-up. V6 contains 80 families and 240 models; Direct selects 46/80 at its canonical $.40/.05$ operating point and observes no false shares. It is a V5-informed fresh follow-up, not a separate FIT/CERT/CONF certification pipeline. Its reported one-sided 95% exact binomial upper limit is .0630493. The zero-error calculation $1 - .05^{1/46}$ reproduces that value and exceeds the 5% target even before a joint risk–coverage allocation. This is a finite-sample resolution example, not an assertion of applicability of a new per-family certificate to the historical selection process. No population-coverage certificate is asserted for V6 under Section 3.4.

CEBaB V9 and V10. V9 is the predeclared transfer attempt: its CERT gate stops after four violations among 60 selected families, with reported upper limit .146097. This selected-set tally does not supply an overall-coverage result for V9. V10 is a separate post-hoc $\alpha = .15$ boundary analysis on 500 confirmation families and 1,500 models. Direct selects 95 families and observes one violation; the reported upper limit is .0489622, while the realized reuse fraction is $.19 < .20$ and the coverage criterion is false. The final study verdict remains `NOT_SUPPORTED`.

These reported limits preserve their source procedures. A small V10 number neither repairs V9's predeclared stop nor becomes a fresh 5% population-risk certificate after the fact. The two versions cannot be merged into a single prospective FIT/CERT/CONF chain. Their useful measurement or ranking signal is retained without asserting that calibration, risk, and coverage criteria jointly authorize deployment.

F.4 Frozen Per-Family Certificates: Gesture and Letter

Inferential population. The Gesture and Letter certificate studies evaluate frozen per-family policies under the risk–coverage criterion of Section 3.4. Their iid interpretation is conditional on a fixed benchmark data environment.

Within each study, FIT, CERT, and CONF families are generated from the same predeclared family generator using independent family/model/data sampling randomness. Phase assignments are fixed before sharing-cost outcomes are accessed. Policy selection uses FIT only; CERT is not used to refit the score or choose a new threshold. CONF remains unopened until the protocol reaches the corresponding stage.

The resulting certificate therefore concerns the declared family generator conditional on that fixed environment. It is not a guarantee for arbitrary new datasets, family generators, or policy classes.

Gesture family generator. For Gesture, a family is generated by selecting the model family from $\{\texttt{l2}, \texttt{l1}, \texttt{elasticnet}\}$, sampling

$$C = 10^U, \qquad U \sim \mathrm{Uniform}(-2, 2),$$

selecting a training fraction from

$$\{.50, .65, .80, 1.00\},$$

and selecting a shift standard deviation from

$$\{.50, 1.00, 1.50\}.$$

Train, set-optimization, audit, and confirmation sampling use independent seeds under the frozen generator. The final panel contains 60 FIT, 120 CERT, and 120 CONF families.

The certificate uses

$$\alpha = .05, \qquad \kappa = .20, \qquad \delta = .05, \qquad \epsilon = .10,$$

with $\delta_R = \delta_Q = .025$.

On CERT,

$$N = 120, \qquad M = 120, \qquad E = 0.$$

The observed reuse fraction is therefore 1.0 and the observed selected violation rate is 0. The one-sided simultaneous bounds are

$$U_R = .0302729726, \qquad L_Q = .9697270274.$$

Thus

$$U_R \le \alpha \qquad \text{and} \qquad L_Q \ge \kappa,$$

so the frozen Gesture policy is authorized under the declared certificate.

A subsequent CONF panel again selects all 120 families and observes zero tolerance violations. The mean confirmation sharing contrast on the shared subset is $.0156965046$. This subsequent agreement corroborates the frozen policy in the declared environment; it is not a claim that all future families are safe, nor a selector advantage over Always Shared.

Letter family generator. The Letter study uses the fixed UCI Letter Recognition environment with 20,000 examples, 16 predictors, and 26 classes. Families vary the model family (`l2`, `l1`, or `elasticnet`), training-set size

$$\{2500, 4000, 6000\},$$

and shift standard deviation

$$\{0.5, 1.0, 1.5\}.$$

Train, validation, set-optimization, audit, and confirmation sampling use independent seeds.

The regularization parameter C is selected using TRAIN/VAL only from the frozen grid

$$\{0.03, 0.1, 0.3, 1, 3, 10, 30\}.$$

The final panel contains 200 FIT, 400 CERT, and 400 reserved CONF families.

FIT selects a frozen threshold of

$$\tau = 0.0750849382.$$

The FIT empirical target is risk at most .05 with coverage at least .20. No FIT policy satisfies both conditions; the frozen procedure therefore uses its prespecified fallback rule (`fallback_min_risk_at_target_coverage`), yielding FIT coverage .20 with 11 violations among 40 selected families. The policy remains frozen and proceeds to CERT under the declared V3 protocol.

On CERT,

$$N = 400, \qquad M = 77, \qquad E = 24,$$

giving observed coverage

$$77/400 = .1925$$

and observed selected risk

$$24/77 = .3116883.$$

The simultaneous one-sided bounds are

$$U_R = .4274275267, \qquad L_Q = .1550109451.$$

Both required conditions fail:

$$U_R > .05, \qquad L_Q < .20.$$

The resulting decision is therefore ABSTAIN at CERT.

The reserved Letter CONF panel remains unopened. This is a prospective abstention for the frozen evaluated policy. It does not establish that no alternative threshold, conditional rule, or richer policy class could satisfy the risk–coverage target.

Scope of the certificate. The Gesture and Letter bounds use the exact one-sided formulas of Appendix C.5 with $\delta_R = \delta_Q = .025$. Both studies use $\epsilon = .10$; this tolerance is distinct from LAD's $\epsilon = .40$ fixed-quota deployment analysis.

The numerical bounds are meaningful together with the generator and freezing conditions above: FIT-only policy selection, independent CERT generation under the declared environment, a fixed per-family rule, and a measurement target frozen before CERT outcomes are accessed.

F.5 FIT STOPS: MASS, EMPIRICAL SEPARATION, AND DESIGN MARGIN

The final FIT-stop summaries give the same selection gate: empirical selected risk at most .015, coverage at least .30, and at least 50 selected families. For $N_{\mathrm{FIT}} = 150$, the minimum count makes the effective coverage requirement $50/150 = 1/3$; for MiniBooNE's 200 it is .30. Neither a safe fraction nor an operating point at only 20% coverage by itself satisfies this full gate.

FIT domain	Below tolerance / N	Best risk at $\geq 20\%$	Max cov. at $\leq 1.5\%$
MiniBooNE	25/200	–	0
Covertype	69/150	.156250	.093333
Sensorless	103/150	.022727	.080000
HIGGS	17/150	.700000	.013333

Table 22: FIT diagnostics for stopped studies. Below-tolerance counts are not SHARE decisions.

MiniBooNE and HIGGS show low measured safe fractions. Covertype has more below-tolerance families, yet the reported threshold frontier does not reach low selected risk at 20% coverage. Sensorless reaches .022727 at at least 20% coverage, which is empirically compatible with a looser .05 risk target; it still fails the frozen .015 margin and the complete FIT gate. Relaxing that gate after outcomes would be a different analysis, and no conclusion about an unrun CERT follows.

These are empirical diagnostics on FIT, not confidence bounds on population safe mass or proofs of population infeasibility. No new one-sided safe-mass tests or joint cross-domain significance claims are introduced here. For MiniBooNE, the saved frozen policy determines the reported FIT outcome. Covertype, Sensorless, and HIGGS likewise stop at FIT with CERT/CONF unopened. Under the predeclared two-of-three V2 success rule, three FIT stops leave that rule unmet; they are not three completed failed CERT evaluations.

G Protocol Evolution and Diagnostic Boundaries

G.1 Versions and original stopping decisions

LAD V1 through V4B and finite-certification V1/V2/V3 are protocol histories, not mutually exchangeable independent replications. Scientific conclusions follow the declared protocol and observed stage outcomes, not completion of auxiliary checks. The V9 CEBaB stop remains separate from the V10 post-hoc analysis, and all unopened stages remain unopened.

G.2 Historical design defects and conditional-policy replay

Wall Robot has a design defect: 40 FIT families cannot satisfy the requirement of at least 50 selections. It is therefore not counted as a completed positive certificate. Gas Drift was not activated and is not treated as a replication.

Retrospective Letter conditional-policy diagnostics require exact replay of the saved rule, including grouping and tie handling, on the already accessed FIT/CERT data. Because exact replay is unavailable, no numerical result from these diagnostics is used here. Any retrospective update must remain separate from the prospective Letter verdict, leave CONF unopened, and not replace the frozen policy.