MODEL ECOSYSTEM PRUNING VIA JOINT BEHAVIORAL COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

An individual redundancy audit asks whether a model can be replaced by its peers. It leaves open whether several such models can be removed together, because their replacements may depend on one another. We study *model ecosystem pruning*: finding a common retained subset and convex replacement weights that reproduce every original endpoint's declared response within a tolerance. Joint behavioral coverage preserves a family of responses rather than the performance of one aggregate predictor. We characterize the fixed convex aggregations it protects and derive a dependency-aware composition condition that bounds error accumulation along replacement chains. Experiments use 20 open-weight LLM judges on 3,724 comparison pairs under seven prompt configurations. At a predeclared tolerance of 0.20, three judges pass individual removal tests, but their joint removal fails in all five data splits. Separate exact analyses require all 20 judges at tolerances up to 0.12 for the fitting objective. Coverage-based selection improves reconstruction over accuracy-based selection, and backward-initialized 2-swap reaches the empirical optimum in 59 of 60 small search instances. Under benchmark shift, refitting weights recovers most of the gap to selecting a new subset. Joint substitutability, rather than individual redundancy alone, determines which ecosystem removals preserve the declared behavior.

Code: https://anonymous.4open.science/r/EPCRC-59AC

1 INTRODUCTION

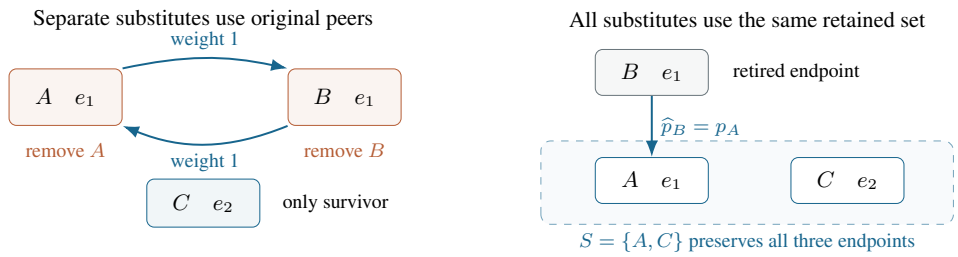
A collection of models can serve more than one purpose. Its predictions may be combined into a single decision, or its members may remain available as separate endpoints. Consider an evaluation service that exposes several LLM judges: users may inspect a particular judge, compare disagreement across judges, or combine their scores under different rules. Reducing the number of models in this service therefore raises a question beyond whether a smaller ensemble is accurate: which original responses can the surviving models still provide?

An individual redundancy audit offers a natural starting point. It asks whether a target model's responses can be reproduced from its peers, as in intervention-based model uniqueness analysis (You, 2026). A small residual suggests that the target could be replaced. Yet that conclusion assumes that its replacement peers remain available. If model A is replaced using B, and B is replaced using A, both audits can succeed even though deleting both removes the very behavior they reproduce. This dependence persists with exact observations and exact substitutes. Individual audits identify possible deletions; they do not establish that those deletions can be executed together.

We study *model ecosystem pruning*: retaining a subset that can simultaneously substitute every model in the original collection. Each deleted endpoint receives its own convex replacement, but all replacements must use the same surviving models. The output is thus both a retained set S and a substitution matrix W, rather than a ranking of redundant models. Figure 1 illustrates this shared-support requirement. We assess it through *joint behavioral coverage*, the largest replacement error across endpoints and declared evaluation contexts, with inputs averaged within each context.

The dependencies also suggest when pruning is possible. A substitute may use another substitute if the chain eventually reaches retained endpoints. We formalize this condition and bound the accumulated error. This separates whether replacements can be implemented from whether they remain accurate.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Valid audits, incompatible deletions

(b) One jointly sufficient retained set

Figure 1: **Individual redundancy does not determine a jointly removable set.** Arrows point from an endpoint to the models used by its substitute. In (a), two endpoints reproduce each other exactly, but both replacements disappear if both endpoints are removed. In (b), every replacement uses the same surviving support. The objects being preserved are declared model responses, not model parameters.

Section 4 takes the alternative route: instead of composing given witnesses, it refits them for every candidate support.

Relation to prior work. Ensemble selection and pruning retain a useful combined predictor (Caruana et al., 2004; Zhang et al., 2006); LLM routing allocates requests across models to balance quality and cost (Chen et al., 2023; Ong et al., 2025). Here the preserved object is a collection of endpoint responses, including the fixed convex aggregations that users may subsequently choose. The underlying geometry is related to approximate convex-hull selection and archetypoid representations (Sartipizadeh & Vincent, 2016; Vinué et al., 2015). Our focus is the validity of simultaneous model replacement: what shared-support coverage preserves, when individual witnesses can be composed, and how selection and reconstruction affect that validity.

LLM judge panels provide a concrete application with comparable numerical outputs (Verga et al.). Across 20 open-weight judges, some individually acceptable deletions fail together; at tighter tolerances, exhaustive enumeration shows that nearly the entire panel is required. The experiments examine which removals the specified interface and tolerance permit, rather than assume that redundancy always allows aggressive compression.

Contributions.

- **Problem.** We formulate ecosystem pruning as simultaneous endpoint substitution on a common retained support, and characterize the fixed convex aggregations that coverage preserves (Section 2).
- **Composition.** We show that exact individual redundancy can disappear under joint removal, then give a dependency-aware sufficient condition with an explicit error-amplification bound (Section 3).
- **Joint selection.** We evaluate convex reconstruction and local exchange against certified empirical optima, distinguishing search error from fixed-matrix infeasibility (Sections 4–5.3).
- **Practical consequences.** A 20-judge study identifies incompatible deletions, tight-preservation limits, and distinct effects of adapting the retained set and its replacement weights (Section 5).

2 Model Ecosystem Pruning

A shared response interface. An ecosystem contains models indexed by $\mathcal{J} = \{1, \ldots, N\}$. For input x and one of finitely many contexts $c \in \mathcal{C}$, model i exposes a response $p_i(x, c) \in \Delta^{d-1}$. A context specifies an evaluation condition, including any matched input transformation. Models receive corresponding inputs, so differences describe their responses to the same condition rather

than differences between queried examples. Their architectures need not be comparable. This response-level view follows behavioral testing and matched-intervention auditing (Ribeiro et al., 2020; You, 2026).

The interface determines what replacement means. In our application, $d = 3$ and the coordinates score "A better," "B better," and "tie." A replacement must return these scores, rather than reproduce unrestricted text. For a retained set $S \subseteq \mathcal{J}$, we define virtual endpoints

$$\widehat{p}_i(x, c) = \sum_{j \in S} W_{ij} p_j(x, c), \qquad W_{ij} \geq 0, \quad \sum_{j \in S} W_{ij} = 1. \tag{1}$$

Rows for $i \in S$ are the identity: a retained model supplies its own response. Each other row is fixed across inputs and contexts. We call (S, W) a *substitution witness*. On a new input, the service evaluates only S and computes the virtual responses numerically. The witness therefore remains executable after every model outside S has been removed. It cannot adapt its weights after observing which peer is correct on that input.

Joint behavioral coverage. Let P_c be the input distribution in context c. With $\mathrm{TV}(p, q) = \frac{1}{2}\|p - q\|_1$, define

$$\mathcal{R}_P(S, W) = \max_{i \in \mathcal{J},\, c \in \mathcal{C}} \mathbb{E}_{x \sim P_c} \mathrm{TV}(p_i(x, c), \widehat{p}_i(x, c)), \tag{2}$$

$$\mathcal{E}_P(S) = \min_W \mathcal{R}_P(S, W) = \max_{i \in \mathcal{J}} \min_{w \in \Delta(S)} \max_{c \in \mathcal{C}} \mathbb{E}_{P_c} \mathrm{TV}\left(p_i, \sum_{j \in S} w_j p_j\right). \tag{3}$$

Rows of W can be optimized independently, giving the equality. Because the maximum over contexts sits inside the minimum over w, a single weight vector must work in every context at once. Allowing a different optimizer in each context would define a different replacement service. Likewise, the maximum over endpoints protects each declared response; the maximum over contexts prevents many easy conditions from masking one difficult condition. We refer to $\mathcal{E}_P(S)$ as the *coverage error* of S, and name the split when reporting it: fitting, validation, or test coverage error.

At tolerance γ, ecosystem pruning solves

$$k^\star_P(\gamma) = \min_{\varnothing \neq S \subseteq \mathcal{J}} |S| \quad \text{subject to} \quad \mathcal{E}_P(S) \leq \gamma. \tag{4}$$

The budget form minimizes $\mathcal{E}_P(S)$ over $|S| \leq k$. Cardinality counts physical endpoints. When every original judge would otherwise be called, retaining k replaces N calls by k calls plus numerical reconstruction. This is not a claim of proportional savings in heterogeneous inference time or monetary cost.

Preservation beyond one aggregate. Covering every endpoint also preserves the ecosystem's fixed convex response combinations. For any $a \in \Delta^{N-1}$, the original response $\sum_i a_i p_i$ has substitute $\sum_{j \in S} (a^\top W)_j p_j$. In fact,

$$\sup_{a \in \Delta^{N-1}} \max_{c \in \mathcal{C}} \mathbb{E}_{P_c} \mathrm{TV}\left(\sum_i a_i p_i, \sum_i a_i \widehat{p}_i\right) = \mathcal{R}_P(S, W). \tag{5}$$

Convexity gives the upper bound; choosing a coordinate vector $a = e_i$ gives equality. Thus endpoint coverage is exactly the error criterion for preserving *every fixed convex aggregation rule*. The weights a need not be chosen before pruning: the same witness supports any such later choice through $a^\top W$. This is stronger than preserving one panel average, but it does not create additional independent judgments. Total variation controls differences in event probabilities; an argmax decision near a tie may still change. Nonlinear or input-adaptive readouts require separate evaluation.

3 Composition of Substitution Witnesses

Coverage tells us what a retained set preserves, but not whether one can be found. We turn to that question. An individual witness measures a target relative to the peers that support it. Joint removal changes that support. We first isolate the resulting obstruction, then ask what additional structure allows existing witnesses to remain useful.

Proposition 1 (Exact individual redundancy can disappear jointly)**.** *There is an ecosystem and a nonempty retained set S such that each model in $R = \mathcal{J} \setminus S$ has a zero-error convex substitute using its original peers, but $\mathcal{E}_P(S) = 1$.*

Take three constant responses $p_1 = p_2 = e_1$ and $p_3 = e_2$. Models 1 and 2 replace each other exactly. Retaining only model 3 produces unit TV error for both. A sequential algorithm that rechecks joint coverage would reject the second deletion. What fails is inferring joint feasibility from the original individual errors alone.

Individual witnesses nevertheless contain more information than their residual magnitudes. Partition their weights into dependencies on removed and retained models. Stacking responses row-wise gives

$$p_R = Ap_R + Bp_S + D, \qquad A \geq 0,\ B \geq 0,\ A\mathbf{1} + B\mathbf{1} = \mathbf{1}, \tag{6}$$

where D is the row-wise residual. Let $\epsilon_i = \max_c \mathbb{E}_{P_c} \frac{1}{2}\|D_i\|_1$ for $i \in R$. The matrix A describes which substitutes themselves depend on endpoints being removed.

Theorem 2 (Dependency-aware composition)**.** *If the spectral radius satisfies $\rho(A) < 1$, then*

$$M = (I - A)^{-1}B \tag{7}$$

is a row-stochastic substitution matrix supported entirely on S. Its robust row errors satisfy

$$\max_c \mathbb{E}_{P_c} \operatorname{TV}\left(p_i, \sum_{j \in S} M_{ij} p_j\right) \leq \left[(I - A)^{-1}\epsilon\right]_i, \qquad i \in R. \tag{8}$$

Consequently, $\|(I - A)^{-1}\epsilon\|_\infty \leq \gamma$ is sufficient for joint feasibility.

The inverse is the nonnegative series $H = \sum_{t \geq 0} A^t$. Its terms expand repeated substitution: B reaches retained endpoints directly, AB passes through one removed endpoint, and A^tB passes through t. Their sum M eliminates the intermediate models while retaining a valid convex response. The same expansion accumulates approximation residuals, producing $H\epsilon$ rather than just ϵ.

Dependency depth and error amplification. The matrices also define an absorbing Markov chain: from a removed endpoint, follow weights in A within R and weights in B into S. The matrix H is its fundamental matrix (Grinstead & Snell, 1997). For a finite ecosystem, $\rho(A) < 1$ precisely when every removed endpoint has a positive-weight path to S. Cycles are therefore permitted; a closed dependency class with no route to a survivor is the obstruction. If $Z_0 = i \in R$ and τ is the first visit to S, then

$$(H\epsilon)_i = \mathbb{E}_i\left[\sum_{t=0}^{\tau-1} \epsilon_{Z_t}\right], \qquad (H\mathbf{1})_i = \mathbb{E}_i[\tau]. \tag{9}$$

Thus joint replacement must satisfy both a termination condition and an error budget. Even when every dependency eventually reaches a survivor, long chains can amplify small individual residuals. For uniform residual bounds $\epsilon_i \leq \varepsilon$, the joint bound becomes $\varepsilon \max_i \mathbb{E}_i[\tau]$. In particular, if every row assigns at most $\alpha < 1$ of its weight to removed models, then

$$\mathcal{E}_P(S) \leq \frac{\max_i \epsilon_i}{1 - \alpha}. \tag{10}$$

The condition provides a concrete reason to expect nontrivial joint pruning: accurate witnesses whose dependencies reach retained models without excessive amplification can be composed. It also explains why small individual errors alone are insufficient. The duplicate cycle in Proposition 1 has $\rho(A) = 1$ and no route to S.

These conditions concern the *given* witnesses; refitting may find other dependencies and a better joint replacement. The theorem supplies an analytical reuse rule, whereas the experiments refit each candidate support. They do not test a spectral explanation of the observed failures. Proofs and the dependency interpretation appear in Appendix A.

Geometry and its limits. For one fixed distribution or response matrix, enlarging S enlarges every convex feasible set, so $\mathcal{E}_P(T) \leq \mathcal{E}_P(S)$ whenever $S \subseteq T$. This connects the selection problem to approximate convex-hull representations (Sartipizadeh & Vincent, 2016). For evaluation, the crucial distinction is that this monotonicity concerns a *reoptimized objective on the same data or distribution*. It need not hold when weights fitted on one sample are evaluated on another. This distinction separates certified empirical infeasibility from an observed failure to generalize.

4 JOINT SELECTION AND RECONSTRUCTION

For each candidate S, the basic operation refits every removed endpoint using only S. On fitting data F, let

$$\widehat{W}_F(S) \in \arg\min_W \widehat{\mathcal{R}}_F(S, W), \qquad \widehat{\mathcal{E}}_F(S) = \widehat{\mathcal{R}}_F(S, \widehat{W}_F(S)). \tag{11}$$

An independent evaluation sample T measures $\widehat{\mathcal{R}}_T(S, \widehat{W}_F(S))$. We keep these two quantities distinct. A fitted witness is not, by itself, a population risk certificate.

Convex fitting and combinatorial search. The per-target minimax TV fit is a linear program: absolute-residual variables represent the loss, and an epigraph variable bounds the mean loss in every context. For three outcomes, $\mathrm{TV}(p, q) = \max_\ell |p_\ell - q_\ell|$, which reduces the number of slack variables. The experiments use a cutting-plane implementation of this objective (Kelley, 1960); Appendix B.4 gives the full formulation.

Backward selection starts with the full ecosystem and removes the model whose deletion leaves the smallest joint error. All affected substitution weights are refitted. Its budget path gives one candidate for every k. Forward selection adds the model yielding the lowest resulting error; forward–trim then removes newly unnecessary members. Every candidate is assessed against the original ecosystem, including endpoints removed at earlier steps. Otherwise an algorithm could discard a representative that still supports a previously removed model. Recomputing all replacements on the candidate support enforces this invariant; reusing an individual redundancy ranking does not.

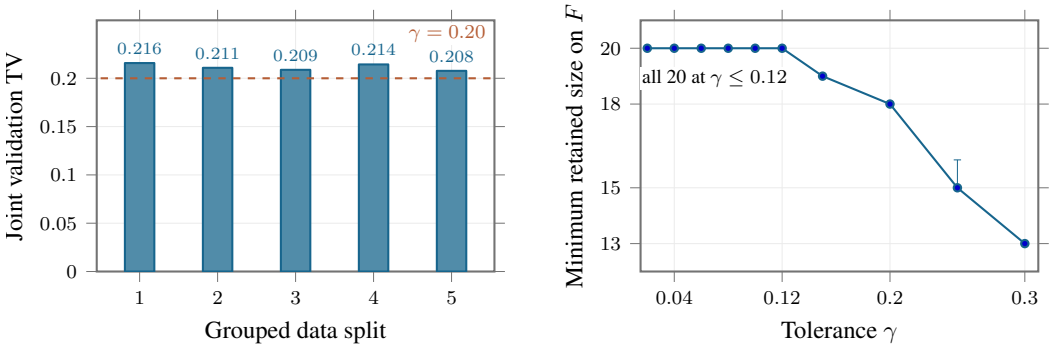
A single deletion can miss a smaller feasible set that requires changing representatives. The evaluated exchange search therefore considers compound moves that remove r members and add $r-1$ outsiders, with $r \leq 2$ or $r \leq 3$, and then rechecks the common-support condition. This differs from only exchanging one member for one other member. An optional equal-cardinality exchange is accepted only when it improves error. Each accepted move decreases $(|S|, \text{error})$ lexicographically, ensuring finite termination. Monotonicity of the reoptimized coverage objective does not supply a diminishing-returns guarantee for this search. We therefore assess its quality against enumerated small-instance optima rather than infer an approximation ratio from greedy construction. One search evaluation includes the inner reconstruction fits, so oracle counts and wall time describe different costs.

Selection, evaluation, and exactness. The main comparisons select and fit on F and evaluate on held-out T. The joint-deletion diagnostic and small-instance search fit on F but inspect validation data V (`CERT` in code); once searched, V is not an untouched test. The full-panel enumeration instead certifies $\widehat{\mathcal{E}}_F$ on F. These objectives respectively measure generalization, finite-procedure search quality, and fixed-matrix feasibility (Appendix B).

5 EXPERIMENTS ON AN LLM JUDGE ECOSYSTEM

Setting and measurements. The ecosystem contains 20 open-weight judges from nine model families. The evaluation uses 3,724 comparison pairs derived from 1,865 source tasks in RewardBench 2 (Malik et al., 2026), under seven named prompt configurations. Scores are normalized likelihoods of three label continuations, not unrestricted generated judgments or assumed calibrated posteriors. Configurations include response-order reversal, rubric changes, verbosity padding, neutral source tags, and a reference-hidden condition. Position sensitivity and rubric dependence motivate evaluating more than the clean configuration (Wang et al., 2024a; Kim et al., 2024). These are declared testing conditions, not assumed independent sources of model diversity; the reference-hidden condition can coincide with clean inputs when no reference was supplied.

All comparisons derived from one source task remain in the same partition. The primary split contains 1,866 fitting pairs, 932 validation pairs, and 926 test pairs; the selection comparison uses five prespecified repartitions of the same cached responses. Uncertainty is summarized by source-task-clustered bootstrapping with 2,000 resamples. The five repartitions assess split sensitivity within this ecosystem. Appendix B lists model revisions, context definitions, and the split used by each analysis.



(a) Joint deletion, fitted on F, scored on V

(b) Exact cardinality for the fitting objective

Figure 2: **Joint validity and empirical feasibility are different questions.** (a) Three judges pass individually at $\gamma = 0.20$, but deleting them together fails in each split. (b) A separate exact analysis of $\widehat{\mathcal{E}}_F$ reports the minimum feasible size; the range at 0.25 spans the five splits. The two panels distinguish fitted validation error from reoptimized fitting coverage.

5.1 Joint removal after individual audits

For each endpoint i, we fit a leave-one-out substitute on F and score it on V. At the predeclared tolerance $\gamma = 0.20$, the same three judges (Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, and Phi-4-mini-instruct) pass separately in every split. Refitting all removed endpoints on the remaining 17 judges yields validation error between 0.208 and 0.216, above 0.20 in all five cases (Figure 2a). Thus the failure persists after joint refitting; it is not just the mechanical reuse of weights pointing to missing models.

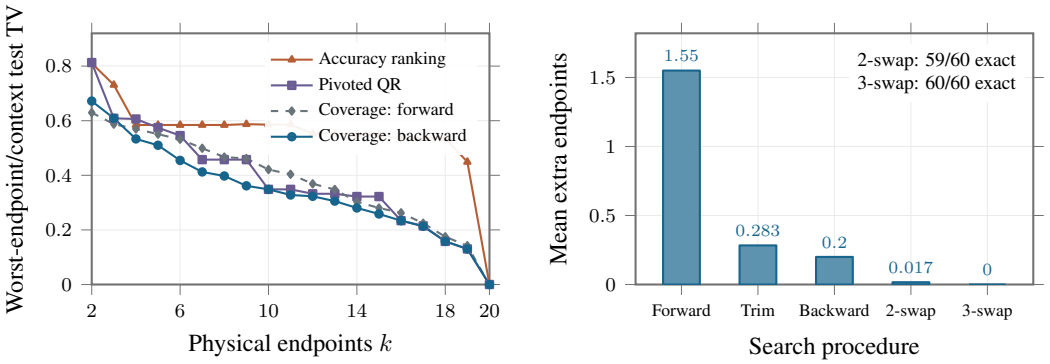
At the data-dependent leave-one-out breakpoints, 86 of 95 split–threshold configurations with at least two individually removable judges also fail after simultaneous removal. These correlated configurations trace the range of the failure; the fixed-threshold result establishes it without choosing a favorable breakpoint. The excess is modest but consistent: individual audits identify candidates, while the joint check evaluates the system that would remain.

5.2 The pruning–fidelity tradeoff

We compare selectors at equal k, using the same convex reconstruction fit. Alternatives include accuracy ranking, 100 random subsets per budget, model-family coverage, measured-cost ordering, clustering, farthest-first selection (Gonzalez, 1985), pivoted QR (Businger & Golub, 1965), and leverage scores. The shared decoder isolates selection: differences arise from which physical responses remain available, rather than from giving one selector a richer reconstruction class. Averaged over $k = 2, \ldots, 19$ and five splits, backward coverage attains test coverage error of **0.363**, compared with 0.404 for pivoted QR, 0.440 for farthest-first, and 0.582 for accuracy ranking. Figure 3a shows how these differences depend on budget; the complete comparison and uncertainty summaries are in Appendix C.1.

Selection for coverage and selection for task performance favor different representatives. At $k = 10$ on the test split, accuracy-based selection achieves 78.98% logistic-readout accuracy, compared with 77.36% for the coverage-selected virtual panel. The strongest geometric baseline is not merely competitive: at this budget pivoted QR attains the same test coverage error as backward coverage to six decimal places. A selector can be useful for prediction without being a good basis for reconstructing the ecosystem.

Tight behavioral preservation is expensive. On the primary test split, retaining 10, 16, and 19 judges gives test coverage error of 0.348, 0.227, and 0.130, respectively. To distinguish search limitations from representational limits, a separate exhaustive search of the fixed fitting problem certifies that all 20 judges are needed for $\gamma \leq 0.12$, 19 for $\gamma = 0.15$, 18 for $\gamma = 0.20$, and 13 for $\gamma = 0.30$ (Figure 2b). Only at $\gamma = 0.25$ does the minimum vary across splits, ranging from 15 to 16. These two results are measured differently: the first is held-out error at a fixed budget, the second is certified



(a) Held-out coverage at equal physical budget

(b) Extra endpoints above an empirical optimum

Figure 3: **Selecting for joint coverage and improving its search.** (a) Five-split mean of the worst endpoint/context test error after fitting on F. (b) Mean cardinality gap across 60 enumerated fitted-validation problems. A 2-swap after backward selection reaches the optimum in 59 cases; 3-swap reaches it in all 60, with higher search time. Exact enumeration exposes the remaining search gap.

feasibility of $\widehat{\mathcal{E}}_F$ on the fitting split. Read together, they show that the tight-tolerance limit is not an artifact of which set the search happened to return, within the measured response matrix and fixed convex reconstruction class.

An average over all endpoints can obscure the cost of retirement because retained endpoints have zero reconstruction error. At $k = 10$, the primary-split test average of per-endpoint worst-context error, before the maximum that defines coverage error, is 0.138 across all 20 judges but 0.276 across the ten actually replaced judges. On the same split, their verdict agreement is 74.81%, compared with 87.41% when identity-reconstructed retained judges are included. Appendix C.3 reports both; a pooled zero median is not evidence that typical retired endpoints are reproduced exactly.

5.3 Exchange and Empirical Optimality

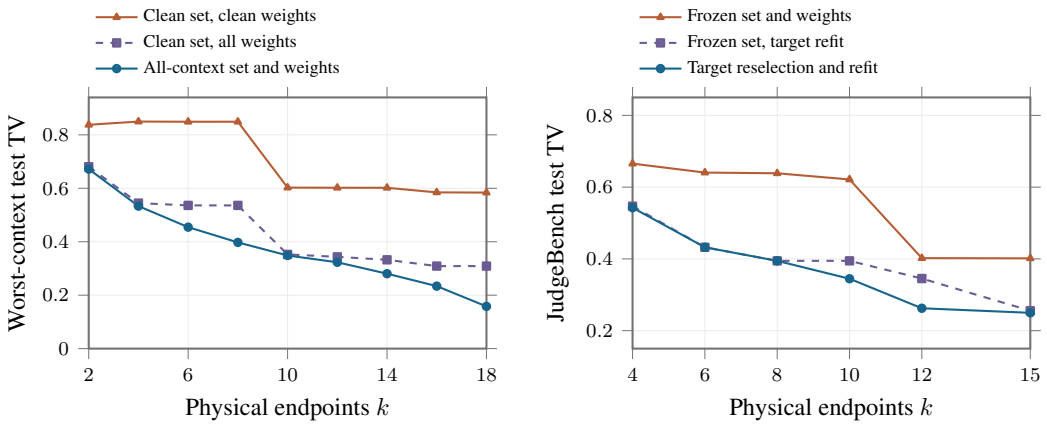
The exact search study uses ten fixed subpanels of size 12–16 and six tolerances from 0.25 to 0.50, giving 60 instances. This grid spans nontrivial feasible sizes; its exploratory selection is documented in Appendix B. Each instance fits weights on F and enumerates subsets against their score on V; enumeration therefore certifies the optimum of this finite fitted-validation procedure.

Forward selection reaches the optimum in 12 of 60 instances and retains 1.55 extra judges on average. Backward selection reaches it in 48 instances, with mean gap 0.20. A backward-initialized 2-swap closes the gap in 59 instances, leaving at most one extra judge in the remaining case. The 3-swap reaches all 60 optima. Its measured mean search time is 169.87 seconds, compared with 50.80 seconds for the 2-swap (Figure 3b; Appendix C.4). Low-order exchange therefore resolves most observed greedy errors at a smaller measured cost than the larger neighborhood. These are optimality comparisons for the specified finite procedure, rather than tests of population risk. They establish that many greedy choices can be repaired without searching the full power set.

5.4 Subset and Witness Transfer

To distinguish subset selection from reconstruction, we compare three arms: selecting and fitting on all contexts; selecting on the clean context but refitting on all contexts; and selecting and fitting only on the clean context. At $k = 10$, the fully context-aware procedure improves test coverage error by 0.254 over the clean-only procedure. Only 0.004 remains when the clean-selected subset receives all-context refitting (Figure 4a). At this budget, most of the benefit comes from learning appropriate substitution weights, not from finding different representatives.

We then transfer to 620 pairs from JudgeBench (Tan et al., 2025), with 310 fitting, 153 validation, and 157 test pairs. The three transfer arms freeze both S and W, freeze S but refit W on target fitting data, or reselect and refit on the target. At $k = 10$, their test TV values are 0.621, 0.395, and 0.345, respectively (Figure 4b). Weight refitting recovers approximately 82% of the frozen-to-reselected



(a) Selection versus context-aware refitting

(b) Frozen transfer versus target adaptation

Figure 4: **The replacement map matters independently of the retained set.** (a) Five-split mean test error; "clean set, all weights" isolates refitting without reselecting the subset. (b) JudgeBench transfer on one recorded split. The gains isolate improvements from adapting weights and from replacing the physical basis.

gap. Reselecting on the target bounds what further adaptation could buy, so the residual gap after refitting is the part that weights alone cannot recover. The result supports a narrower conclusion than domain-invariant substitution: a usable physical basis can survive a shift that invalidates its fitted replacements.

5.5 Downstream compatibility

At $k = 10$, we compare the full panel, its virtual reconstruction, and the retained physical judges used without reconstruction. Table 1 reports accuracy and agreement with the full panel. Under mean-probability aggregation, full-to-virtual accuracy changes by only -0.72 percentage points, while 16.68% of verdicts change. Under majority voting the changes are -0.37 points and 13.57%. A learned logistic readout has higher full-panel accuracy but loses 3.13 points after reconstruction. Accuracy compares each system with the external reference label; agreement compares the systems with each other. A changed verdict can correct an old error or introduce a new one, so these metrics answer different questions. The larger logistic loss also shows why the small changes for mean and majority aggregation cannot be generalized to every readout.

Table 1: **Preserving usefulness is different from preserving the evaluator.** Percentages at $k = 10$, averaged across five splits and seven configurations. Agreement is with the full panel's verdict. "Virtual" reconstructs all 20 endpoints from ten physical judges; "physical" applies the readout to those ten directly.

	Accuracy			Verdict agreement	
Readout	Full	Virtual	Physical	Virtual	Physical
Mean probabilities	54.62	53.90	51.77	83.32	70.53
Majority vote	54.90	54.53	52.29	86.43	63.52
Logistic readout	80.49	77.36	77.34	87.02	87.00

Reconstruction is useful for compatibility with a fixed readout: mean-probability agreement improves from 70.53% for the physical panel to 83.32% for the virtual one. It does not create independent judgments. The virtual mean is exactly a reweighted physical mean, with effective weights $b_j = N^{-1} \sum_i W_{ij}$. Likewise, unrestricted linear readouts have the same hypothesis class on physical and linearly reconstructed features when identity coordinates are retained (Appendix A.5). The nearly identical physical and virtual logistic accuracies are consistent with this fact; regularization can still make their fitted predictors differ. Joint coverage serves endpoint compatibility, not an assurance that reproducing every old judgment maximizes accuracy.

6 Related Work

Individual representation and collective replacement. Model uniqueness auditing measures a target's residual beyond the convex responses of its peers under matched interventions (You, 2026). Convex donor combinations also underlie synthetic control (Abadie et al., 2010), although our observed-response reconstruction does not invoke its causal identification assumptions. Both connections clarify what an individual witness contains: a target-specific approximation and the peers on which it depends. Ecosystem pruning changes the operation by removing multiple targets and potential donors together. The relevant constraint is that every replacement remains supported after the same deletion.

Compression, selection, and routing. Ensemble selection and pruning optimize the performance of a combined predictor (Caruana et al., 2004; Zhang et al., 2006). Distillation learns a student to reproduce teacher or ensemble outputs (Hinton et al., 2015); model soups combine compatible model parameters (Wortsman et al., 2022). Routing and cascades select which models to invoke for a query (Chen et al., 2023; Ong et al., 2025). These approaches reduce cost through different retained objects: a predictor, a parameter vector, or a request-dependent policy. We retain existing endpoints and a fixed replacement map for the original collection. A useful aggregate and a behavior-preserving ecosystem can therefore favor different subsets, as the judge experiments show.

Geometric representatives. The fixed-response problem is closely related to approximate convex-hull selection (Sartipizadeh & Vincent, 2016). Archetypal analysis represents data through convex extreme patterns (Cutler & Breiman, 1994), with geometric and sparse-projection algorithms developed by Damle & Sun (2017) and Abrol & Sharma (2020). Importantly, archetypoid analysis already restricts representatives to observed members (Vinué et al., 2015), including functional observations (Epifanio, 2016). These methods establish the representative-selection geometry; we connect it to executable joint removal, dependency-aware composition, and cross-context endpoint risk. Approximate Carathéodory results address the support needed for an individual convex approximation (Barman, 2015); small support for each row does not imply a small union of supports across the ecosystem.

Judge quality, dependence, and replaceability. LLM evaluation includes general-purpose judges (Zheng et al., 2023; Liu et al., 2023), specialized evaluators (Wang et al., 2024b; Kim et al., 2024), and diverse panels (Verga et al.). RewardBench 2 and JudgeBench provide tasks for assessing evaluator quality (Malik et al., 2026; Tan et al., 2025); order sensitivity motivates explicit protocol controls (Wang et al., 2024a). Calibration can make individually noisy judges useful (Li, 2026), while correlated errors can reduce effective panel independence (Kohli, 2026). Neither standalone accuracy nor effective independence determines joint substitutability. An endpoint may be useful yet reproducible by others, or distinctive without improving a chosen aggregate. Our objective measures whether its declared behavior remains available after consolidation, rather than whether that behavior deserves greater weight.

7 Discussion and Conclusion

Model ecosystem pruning requires more than detecting redundant members. A replacement that depends on another deleted model is not an executable replacement. Common-support coverage makes this requirement explicit, and dependency-aware composition separates the existence of a retained-only substitute from the accumulation of its approximation errors. This provides a structural basis for joint pruning without assuming that model collections are uniformly compressible.

The judge experiments show why the distinction matters. Individual removal tests can disagree with joint feasibility; coverage and accuracy select for different properties; and strict preservation can require almost the entire measured ecosystem. Under shift, refitting the replacement map explains much of the improvement otherwise attributed to changing representatives. These findings concern one 20-judge ecosystem and its declared numerical interface. Richer decoders or other protocols may yield different pruning frontiers, and preserving an endpoint can preserve its mistakes. When task quality alone is the goal, it should be optimized directly. When original model responses must remain available, the appropriate operation is joint replacement: fewer physical models, with an explicit account of which behaviors they still support.

AI USE STATEMENT

Generative AI tools were used to help restructure and edit the manuscript, interpret and present archived experimental results, prepare LaTeX figures and tables, and write code for build and data-consistency checks. All numerical results in the paper come from archived experimental outputs; AI tools were not used to invent missing measurements. The authors are responsible for the final text, mathematical claims, experimental results, and released artifacts.

REPRODUCIBILITY STATEMENT

Numerical provenance. All experimental values in this paper come from the existing experiment outputs. The accompanying LaTeX package includes the archived summary tables, metadata, deterministic data-formatting scripts, and an independent standalone LaTeX source for each vector subfigure. These materials reproduce the tables and figures without new model inference. The proofs and dependency-aware composition result are analytical additions; they are not reported as experimentally validated mechanisms.

The supplied summaries are not a complete inference reproduction bundle. They do not contain every prompt realization, the complete per-item probability cache, or all substitution matrices. Reproducing inference or testing a new decoder requires those original experiment artifacts and the corresponding scoring environment. The present package therefore supports numerical traceability and document reconstruction, not a claim that all original GPU measurements can be regenerated from its contents alone.

Scientific scope. The tests cover a defined 20-endpoint scoring ecosystem and its subpanels. Multiple splits, configurations, and search thresholds reuse that ecosystem. The benchmark transfer changes evaluation inputs, not the set of available model families. Coverage is sensitive to the reporting interface, context family, loss, and allowed reconstruction class. It neither establishes unrestricted functional equivalence nor determines how many judges an application needs for optimal prediction.

Responsible use. Replacing a model's numerical outputs can preserve undesirable judgments as well as desirable ones. Behavioral fidelity should not be equated with fairness, factual correctness, safety, or preservation of human viewpoints. A service adopting virtual endpoints should disclose the substitution and the interface under which it was evaluated. The framework is intended to make those replacements explicit, not to hide changes to an evaluator behind a stable aggregate accuracy.

REFERENCES

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Vinayak Abrol and Pulkit Sharma. A geometric approach to archetypal analysis via sparse projections. In *International Conference on Machine Learning*, pp. 42–51. PMLR, 2020.

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):1641–1662, 2025.

Siddharth Barman. Approximating nash equilibria and dense bipartite subgraphs via an approximate version of caratheodory's theorem. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 361–369, 2015.

Peter Businger and Gene H Golub. Linear least squares solutions by householder transformations. *Numerische Mathematik*, 7(3):269–276, 1965.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 18, 2004.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

Anil Damle and Yuekai Sun. A geometric approach to archetypal analysis and nonnegative matrix factorization. *Technometrics*, 59(3):361–370, 2017.

Irene Epifanio. Functional archetype and archetypoid analysis. *Computational Statistics & Data Analysis*, 104:24–34, 2016.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38:293–306, 1985.

Charles M Grinstead and J Laurie Snell. Markov chains. *Introduction to probability*, pp. 405–470, 1997.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

James E Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine llm evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026.

Yanran Li. Calibrate, don't curate: Label-efficient estimation from noisy llm judges. *arXiv preprint arXiv:2605.09702*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah Smith, Hanna Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 144839–144866, 2026.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, Mohammed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, volume 2025, pp. 34433–34448, 2025.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Hossein Sartipizadeh and Tyrone L Vincent. Computing the approximate convex hull in high dimensions. *arXiv preprint arXiv:1603.04422*, 2016.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pp. 63277–63303, 2025.

P Verga, S Hofstatter, S Althammer, Y Su, A Piktus, A Arkhangorodsky, M Xu, N White, and P Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models, 2024. *arXiv preprint arXiv:2404.18796*.

Guillermo Vinué, Irene Epifanio, and Sandra Alemany. Archetypoids: A new approach to define representative archetypal data. *Computational Statistics & Data Analysis*, 87:102–115, 2015.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024a.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *International Conference on Learning Representations*, volume 2024, pp. 43573–43593, 2024b.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. Pmlr, 2022.

Lei You. Quantifying model uniqueness in heterogeneous ai ecosystems. *arXiv preprint arXiv:2601.22977*, 2026.

Yi Zhang, Samuel Burer, W Nick Street, Kristin P Bennett, and Emilio Parrado-Hernández. Ensemble pruning via semi-definite programming. *Journal of machine learning research*, 7(7), 2006.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Proofs and Operational Consequences

A.1 Coverage under fixed convex aggregation

For a fixed witness (S, W), the convexity of the ℓ_1 norm yields, pointwise,

$$\mathrm{TV}\left(\sum_i a_i p_i, \sum_i a_i \widehat{p}_i\right) = \frac{1}{2}\left\|\sum_i a_i (p_i - \widehat{p}_i)\right\|_1 \tag{12}$$

$$\leq \sum_i a_i \,\mathrm{TV}(p_i, \widehat{p}_i). \tag{13}$$

Taking expectations in any context, and then the maximum over contexts, bounds this quantity by $\mathcal{R}_P(S, W)$. The bound holds uniformly for every fixed $a \in \Delta^{N-1}$. Since the index and context sets are finite, take $(i^\star, c^\star)$ achieving the maximum in (2); choosing $a = e_{i^\star}$ attains that value. This proves (5). The substituted aggregation is implementable on S because $b = a^\top W$ is nonnegative and sums to one.

There is no interchange of a supremum over input-dependent policies with the expectation. A policy $a(x, c)$ can emphasize different endpoint errors on different inputs, potentially producing $\mathbb{E}\max_i \mathrm{TV}(p_i, \widehat{p}_i)$ instead of $\max_i \mathbb{E}\,\mathrm{TV}(p_i, \widehat{p}_i)$. The equality also does not apply directly to majority voting, a trained nonlinear aggregator, or preservation of individual argmax decisions. Those are separate response interfaces or downstream measurements.

A.2 Failure of independent witnesses

Proof of Proposition 1. Let all inputs and contexts produce $p_1 = p_2 = e_1$ and $p_3 = e_2$ in a simplex with at least two coordinates. The peer-only witnesses $p_1 = p_2$ and $p_2 = p_1$ have zero residual. Set $R = \{1, 2\}$ and $S = \{3\}$. The only convex response on S is e_2. Its TV distance to either removed response is one, so $\mathcal{E}_P(S) = 1$. In particular, arbitrarily small individual removal errors do not force a small joint error. TV is bounded by one; the failure is maximal, not unbounded. □

The example identifies a dependency problem, not a claim that independent audits are useless. They can screen candidates or initialize a joint search. If a sequential pruning method refits and checks all removed endpoints after each deletion, it explicitly prevents this failure. Rechecking only the newly deleted endpoint would still leave earlier witnesses unsupported.

A.3 Dependency-aware composition

Proof of Theorem 2. Write $H = (I - A)^{-1}$. Since $A \geq 0$ and $\rho(A) < 1$, the Neumann series converges and gives $H = \sum_{t=0}^{\infty} A^t \geq 0$. Consequently $M = HB \geq 0$. The stochasticity of the original witnesses implies

$$B\mathbf{1} = (I - A)\mathbf{1}, \qquad M\mathbf{1} = H(I - A)\mathbf{1} = \mathbf{1}. \tag{14}$$

Thus each row of M is a convex weight vector over retained models. Rearranging (6) gives, for each input and context,

$$p_R - Mp_S = HD. \tag{15}$$

Because H is nonnegative, the triangle inequality gives

$$\mathbb{E}_{P_c}\frac{1}{2}\|(p_R - Mp_S)_i\|_1 \leq \sum_{r \in R} H_{ir}\mathbb{E}_{P_c}\frac{1}{2}\|D_r\|_1 \tag{16}$$

$$\leq \sum_{r \in R} H_{ir}\epsilon_r. \tag{17}$$

Taking the maximum over c proves (8). Identity rows have zero residual for retained endpoints, so appending them yields a witness for the entire ecosystem with risk at most $\|H\epsilon\|_\infty$. Optimizing over all witnesses cannot make this bound worse. □

Derivation of (10). If every row sum of A is at most $\alpha < 1$, then $\|A\|_\infty \le \alpha$ and $\rho(A) \le \alpha$. The same series gives $\|H\|_\infty \le \sum_{t\ge 0} \alpha^t = (1-\alpha)^{-1}$. Hence $\|H\epsilon\|_\infty \le \max_i \epsilon_i/(1-\alpha)$. Since TV is at most one, one may also truncate this upper bound at one. □

The sufficient condition is about the given witness matrix. A cycle with eventual positive flow to a retained model can still have spectral radius below one; the presence of a cycle alone does not prove failure. Conversely, a closed cycle may simply reflect a poor choice of individual substitutes even when a different retained-only reconstruction exists. For exact witnesses satisfying $D = 0$ and $\rho(A) < 1$, the composed witness is exact. For approximate witnesses, H exposes the accumulation of residuals along substitution chains. This is a direct application of a nonnegative matrix series to the ecosystem replacement problem; it is not an approximation guarantee for subset search. The experiments do not estimate $H\epsilon$ as a causal explanation of the observed failures.

Dependency paths and accumulated residuals. For completeness, augment the witness blocks to a Markov transition matrix with retained states absorbing,

$$Q = \begin{pmatrix} A & B \\ 0 & I \end{pmatrix}. \tag{18}$$

Starting at $i \in R$, the probability of being at $r \in R$ at step t is $(A^t)_{ir}$. When $\rho(A) < 1$, summing these probabilities gives H_{ir}, the expected number of visits to r before absorption, including a visit at time zero. Multiplying by the fixed nonnegative residual bound ϵ_r and summing over r proves the first identity in (9); taking all costs equal to one proves the second. This uses the standard fundamental-matrix interpretation (Grinstead & Snell, 1997, Section 11.2) to interpret the replacement error bound, not a stochastic assumption on the models' score outputs.

The graph condition follows directly in the finite case. If each removed state has a positive-weight path to S, there are a finite L and a uniform $\beta > 0$ such that absorption within L steps has probability at least β from every state in R. Consequently $\|A^L\|_\infty \le 1 - \beta$, which implies $\rho(A) < 1$. If some removed state cannot reach S, the set of states it can reach contains a closed class within R. The associated stochastic block has eigenvalue one, giving $\rho(A) = 1$ because A is substochastic. Thus cycles with an eventual exit are allowed; closed classes are not. The bound $H\epsilon$ additionally requires that this exit occurs quickly enough relative to the approximation residuals.

Beyond probability-valued endpoints. The composition proof only uses nonnegativity of the witness weights and the triangle inequality. It therefore extends to responses in a normed vector space by replacing $\frac{1}{2}\|\cdot\|_1$ with the declared norm. Scalar responses with absolute loss are a special case. The simplex assumption supplies the probability-preservation and total-variation interpretations; it is not needed for the algebra of common-support substitution.

A.4 Monotonicity on one objective

Lemma 3. *If $S \subseteq T$ and the response functions, context family, distributions, loss, and optimization rule are unchanged, then $\mathcal{E}_P(T) \le \mathcal{E}_P(S)$. The same holds for the reoptimized empirical fitting objective $\widehat{\mathcal{E}}_F$.*

Proof. Extend every feasible weight vector over S to T by assigning zero mass to $T \setminus S$. The reproduced responses are unchanged. Any newly retained model can instead use its identity row and incur zero loss. Thus every approximation achievable using S is matched or improved using T. □

This argument concerns the optimum value of the same objective. A minimizer $\widehat{W}_F(T)$ need not be the extension of $\widehat{W}_F(S)$, and its evaluation risk on a different sample can be higher. Consequently a held-out error at $k = N - 1$ is not a monotonicity-based lower bound on all smaller fitted panels. The exact fitting analysis in Figure 2b is deliberately distinguished from held-out frontiers.

A.5 Why virtual readouts do not create information

Flatten the retained probabilities into a feature vector $z \in \mathbb{R}^{|S|d}$. The virtual feature vector is $\widehat{z} = Tz$, with $T = W \otimes I_d$ under a consistent indexing convention. Any linear predictor $B\widehat{z}$ equals $(BT)z$,

so every linear readout of virtual features is a linear readout of physical features. Conversely, the virtual representation includes the identity coordinates of S. There is a coordinate-selection matrix Q with $QT = I$, so every physical readout Cz equals $CQ\widehat{z}$. The unrestricted linear readout classes therefore coincide; adding a bias term preserves this equality.

Regularization in parameter space need not be invariant under this feature transformation, and finite-sample optimization need not return the same classifier. The claim is about representational classes, not equality of all training procedures. For the simple panel mean, the equivalence is pointwise and requires no fitting:

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{p}_i = \sum_{j\in S}\left(\frac{1}{N}\sum_{i=1}^{N}W_{ij}\right)p_j. \tag{19}$$

A virtual mean should therefore be interpreted as an effective reweighting of the retained judges, not as N independent votes. This identity also explains why comparing only against an *unweighted* physical mean isolates compatibility with the original aggregation rather than new predictive information.

B Experimental Interface and Protocol

B.1 The measured ecosystem

Table 2 specifies the 20 endpoints. The experiment records fixed model revisions; the abbreviated hashes below are the reported prefixes. Different sizes within a family are distinct endpoints. The inference record reports 20 judges $\times$ seven configurations $\times$ 3,724 pairs, or 521,360 scored model–configuration–pair cells, without imputing missing scores. This describes the original measurement run; the accompanying paper package contains its numerical summaries and figure data, not that complete score tensor.

Table 2: The 20-judge ecosystem and recorded revision prefixes. IDs are used only to cross-reference results and data; they are not experiment names.

ID	Model endpoint	Revision prefix
J01	Qwen/Qwen2.5-3B-Instruct	`aa8e72537993`
J02	Qwen/Qwen2.5-7B-Instruct	`a09a35458c70`
J03	Qwen/Qwen2.5-14B-Instruct	`cf98f3b3bbb4`
J04	Qwen/Qwen3-4B	`1cfa9a720891`
J05	Qwen/Qwen3-8B	`b968826d9c46`
J06	Qwen/Qwen3-14B	`40c069824f42`
J07	meta-llama/Llama-3.2-3B-Instruct	`0cb88a4f764b`
J08	meta-llama/Llama-3.1-8B-Instruct	`0e9e39f249a1`
J09	google/gemma-3-4b-it	`093f9f388b31`
J10	google/gemma-3-12b-it	`96b6f1eccf38`
J11	microsoft/Phi-4-mini-instruct	`cfbefacb9925`
J12	microsoft/phi-4	`2db69c1c3e91`
J13	microsoft/Phi-3.5-mini-instruct	`2fe192450127`
J14	mistralai/Mistral-7B-Instruct-v0.3	`c170c708c41d`
J15	mistralai/Mistral-Nemo-Instruct-2407	`04d8a90549d2`
J16	ibm-granite/granite-3.3-8b-instruct	`51dd4bc2ade4`
J17	tiiuae/Falcon3-7B-Instruct	`1e57a0ecd176`
J18	tiiuae/Falcon3-10B-Instruct	`8799bc6aec01`
J19	CohereLabs/aya-expanse-8b	`5062468bf9bc`
J20	allenai/OLMo-2-1124-7B-Instruct	`470b1fba1ae0`

Score semantics. The interface normalizes likelihoods of three answer-label continuations. If ℓ_A, ℓ_B, ℓ_T denote the protocol's recorded continuation scores, the numerical endpoint is

$$p = \mathrm{softmax}(\ell_A, \ell_B, \ell_T). \tag{20}$$

These probabilities are used as a fixed scoring interface. We do not assert calibration to human probabilities, equivalence to free-form judge explanations, or invariance to a different continuation-scoring protocol. A virtual judge evaluates all members of S and combines these numerical outputs.

A stochastic router that samples only one retained judge would match their average response but would add routing variability, and is not the operational experiment reported here.

B.2 CONTEXTS, MATCHED INPUTS, AND GROUPING

Table 3: Named context configurations. Each row describes the recorded transformation, not a claim that every transformation is nontrivial on every source example.

Configuration	Description
Clean	Baseline pairwise-comparison prompt and evaluation rubric.
Order swap	Swap responses A and B; map the resulting A/B coordinates back to the canonical response order before comparison.
Correctness rubric	Emphasize correctness rather than general response preference.
Bias-resistant rubric	Use the alternate rubric designed to discourage superficial preference cues.
Verbosity padding	Add the configured padding to response A. This tests the recorded perturbation, not all forms of verbosity bias.
Neutral source tags	Add neutral source identifiers without disclosing the true generator identity.
Reference hidden	Remove the reference field where present. On examples without a supplied reference, this operation may coincide with clean.

All models see corresponding items in all configurations. One row of substitution weights must work across the entire configured family. The reference-hidden arm is retained as a recorded configuration, but it is not counted as independent evidence that a distinct evidence-removal mechanism was tested. In particular, equality with the clean condition where no reference is present is not evidence of robustness to removing substantive evidence.

The primary RewardBench 2 partition uses seed 20260818. It contains 935 source tasks / 1,866 pairs in F, 467 tasks / 932 pairs in V, and 463 tasks / 926 pairs in T. Thus the pair counts are approximately 50/25/25 percent, not thirds. The five recorded split seeds are 20260818, 20260819, 20260820, 20260821, and 20260822. Data augmentation, alternate ordering, and comparisons derived from a common source task stay within a group. The transfer study uses the recorded 620-pair JudgeBench construction and a single 310/153/157 fit/validation/test partition. Absence of gold tie labels does not remove the tie coordinate from judge outputs.

B.3 WHICH OBJECTIVE EACH ANALYSIS EVALUATES

Table 4: Data roles must be read together with every claim. F supplies weight fitting, V is the separate validation partition called `CERT` in the implementation, and T is test. Repeated partitions share the same underlying inference cache.

Analysis	Objective and use of data	Repetition
Independent / joint audit	Fit on F, inspect joint error on V	Five splits
Exact full-panel cardinality	Optimize and evaluate $\widehat{\mathcal{E}}_F$ on F	Five splits
Primary detailed frontier	Select and fit on F; evaluate on T	Primary split
Selector comparison	Select and fit on F; evaluate on T	Five splits
Local-exchange reference	Fit on F; search / enumerate scores on V	60 instances
Context refitting	Select / fit in specified contexts on F; test on T	Five splits
Downstream readouts	Frozen selected features; test predictions on T	Five splits
Benchmark transfer	Frozen or target-fitted witnesses; target T	One split
Risk-bound diagnostic	Repeated fit / validation / test assignments	120 repartitions

The original execution plan anticipated tighter compression than observed. The larger tolerance grid in the small-instance exchange study and the interpretation of accuracy versus decision agreement are exploratory analyses. The joint-deletion result at $\gamma = 0.20$ is on the predeclared tolerance grid. We preserve this distinction rather than retrospectively treating every analysis as confirmatory.

B.4 Optimization and search details

For target i and fixed S, let x_{ct} denote the tth fitting item in context c, and write $r_{ct\ell} = p_{i\ell}(x_{ct}, c) - \sum_{j\in S} w_j p_{j\ell}(x_{ct}, c)$. A general finite-dimensional TV formulation is

$$\min_{w,u,z} \quad u \tag{21}$$

$$\text{subject to} \quad w \geq 0, \quad \mathbf{1}^\top w = 1, \tag{22}$$

$$z_{ct\ell} \geq r_{ct\ell}, \quad z_{ct\ell} \geq -r_{ct\ell}, \tag{23}$$

$$\frac{1}{2n_c} \sum_{t,\ell} z_{ct\ell} \leq u \quad \text{for every } c. \tag{24}$$

For three outcomes, the residual coordinates sum to zero and $\frac{1}{2}\|r\|_1 = \|r\|_\infty$. The implementation can therefore replace the coordinate slacks with one $z_{ct} \geq \pm r_{ct\ell}$ per item and impose $n_c^{-1} \sum_t z_{ct} \leq u$. The experimental solver evaluates the equivalent convex piecewise-linear objective using cutting planes. A singleton retained set is valid and has the unique weight one; an empty retained set is infeasible.

Subset search is more expensive than one convex fit. With s retained and $N - s$ excluded models, a remove-r/add-$(r-1)$ neighborhood contains $\binom{s}{r}\binom{N-s}{r-1}$ candidates. An enabled equal-size swap adds $\binom{s}{r}\binom{N-s}{r}$ candidates. Each candidate can require multiple target fits, and several search rounds may be needed. We therefore report observed search time and cardinality gap rather than equating a neighborhood-count power of N with end-to-end runtime.

The full fitting-objective analysis uses monotonicity-based subset pruning and a capped search, retaining exact labels only for completed instances. The small-subpanel reference enumerates cardinality layers against the fitted validation score. Its cache can reuse an error lower bound across tighter tolerances for the *same subset*; it does not require subset monotonicity of held-out error. A feasible solution gives an upper bound on minimum size, while exhausting all smaller layers is needed for exactness. In the independent-audit diagnostic, the capped minimum-size search does not establish an exact composition gap; the direct joint-error measurements are sufficient for the stated failure result.

C Additional Results and Their Interpretation

C.1 The full selection comparison

Table 5: Worst-endpoint/worst-context TV, averaged over $k = 2, \ldots, 19$ and five splits. The final column retains the archive's averaged pointwise bootstrap endpoints; it is not a newly constructed simultaneous 95% interval for the complete sweep. Random is the mean of 100 draws, not the best draw.

Selection rule	Mean error	Mean bootstrap endpoints
Joint coverage: backward	**0.363**	[0.352, 0.379]
Joint coverage: forward	0.401	[0.391, 0.419]
Pivoted QR	0.404	[0.392, 0.420]
Farthest first	0.440	[0.430, 0.456]
Hierarchical	0.456	[0.445, 0.473]
Correlation clustering	0.459	[0.446, 0.478]
k-medoids	0.473	[0.459, 0.491]
Cheapest first	0.487	[0.475, 0.505]
Leverage scores	0.530	[0.517, 0.548]
Random	0.540	[0.526, 0.560]
Descending measured cost	0.555	[0.542, 0.577]
One per family	0.557	[0.544, 0.575]
Top accuracy	0.582	[0.566, 0.604]

All methods in the selection comparison share the same fitted convex reconstruction class; they differ in the physical set selected. The lowest average error is therefore evidence that the joint objective

is useful for finding representatives, not that a new decoder has been given an unfair advantage. The alternative selectors span individual utility, measured inference cost, family diversity, pairwise geometry, and linear-algebraic representatives. The descending-cost baseline uses measured scoring time as a proxy, not a direct ordering by parameter count.

The archive also contains paired source-task bootstrap comparisons for each split–budget–method cell. Among 1,080 recorded cells, coverage backward has a worse point estimate in 57, and 31 pointwise intervals exclude zero in that direction. Those 31 occur at $k \leq 4$. These are correlated, unadjusted comparisons and should not be interpreted as familywise significance or as proof that every difference at larger k is zero. At very small budgets, accumulated deletion choices can make backward selection inferior to another heuristic; the number of prior deletions is large, not small.

The paired random comparison uses the best of 100 draws selected after seeing test performance. That post hoc diagnostic is a stronger reference than ordinary random selection and is not a deployable rule without additional selection data. It is deliberately not substituted for the ordinary random row in Table 5. Likewise, equal-budget inference counts should not be presented as equal GPU time when endpoints differ in cost.

C.2 Exact empirical cardinality and mandatory endpoints

An endpoint is *mandatory* at a fixed empirical objective and tolerance when it belongs to every optimal retained set. Appearance in one solver-returned optimum does not establish this property. The recorded fitting-objective search enumerates the relevant optimal sets. At $\gamma = 0.20$, the optimum has size 18 in all five splits. Four splits have three optimal sets, with 17 mandatory and three optional endpoints; the remaining split has four optimal sets, with 16 mandatory and four optional endpoints. The extra optional endpoint there is J20. Across all five splits, 16 endpoints are always mandatory at that tolerance.

These membership claims belong to the fitting objective $\widehat{\mathcal{E}}_F$, whereas the independent-removal diagnostic is evaluated on V. The exact size 18 cannot therefore be used to declare that the validation diagnostic overdeletes by *exactly one*. Its direct observation is that the selected 17-endpoint set violates the validation tolerance. The full fitting frontier is separately useful because it distinguishes empirical infeasibility from a heuristic failure to find a feasible set.

C.3 Retired-only reconstruction metrics

Suppose every endpoint is weighted equally in a reported mean and retained endpoints reconstruct themselves. If k of N endpoints remain, an all-endpoint mean error $\bar{e}$ and agreement $\bar{a}$ imply

$$\bar{e}_{\text{retired}} = \frac{N}{N-k}\bar{e}, \qquad \bar{a}_{\text{retired}} = \frac{N\bar{a}-k}{N-k}, \qquad k < N. \tag{25}$$

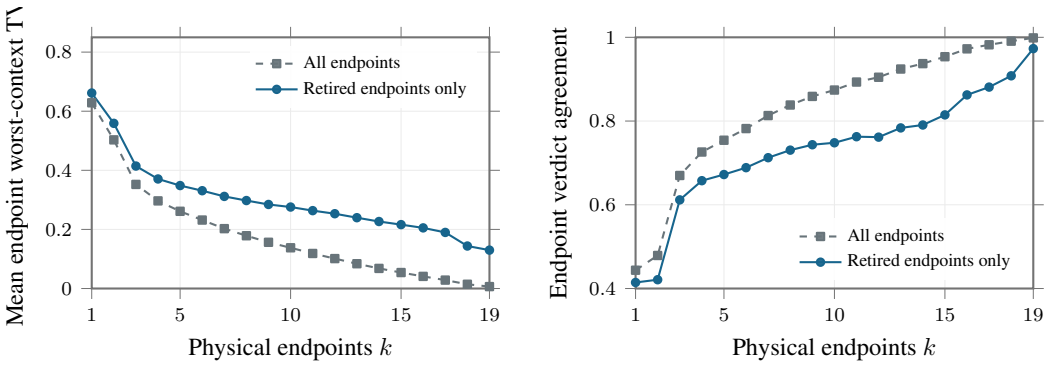
These are algebraic re-expressions of the archived summaries, not new fitted measurements. They do not recover retired-only quantiles from pooled quantiles. Figure 5 reports the corrected means and agreements for the primary split. At $k = 10$, half of all endpoint blocks are identities, making a pooled median near zero unsurprising even when retired judges incur substantial error.

C.4 Search quality and computational cost

There are two subpanels at each size 12, 13, 14, 15, and 16. The same fixed subpanels are assessed at six tolerances. Consequently the 60 instances are a systematic finite test set, not 60 independent ecosystems. A smaller neighborhood can be effective without possessing a uniform approximation guarantee. The improvement from backward to 2-swap is especially interpretable because a reduction may need to replace two retained representatives with one previously excluded model.

C.5 Reconstruction classes and sparse witnesses

The archived ablation holds the selected set fixed and compares simplex fitting with nearest-peer, uniform, nonnegative, and unconstrained ridge reconstruction. Figure 6a reports the recorded half-ℓ_1 errors. For off-simplex outputs, this quantity is not literally the total variation between two probability distributions. In particular, raw unconstrained ridge can leave the simplex; this establishes an interface



(a) Mean per-endpoint worst-context error

(b) Endpoint verdict agreement

Figure 5: **Only retired endpoints require reconstruction.** Primary-split test statistics, with identity blocks shown separately through exact mean corrections. Full-panel pooled quantiles cannot identify the error distribution of retired endpoints.

Table 6: Exact-match frequency, cardinality gap, and recorded mean runtime on the 60 small fitted-validation problems. All minimum sizes refer to those enumerated empirical instances. The timings are those of the archived executions, not a new controlled hardware comparison.

Method	Exact	Mean gap	Max gap	Seconds
Forward	12/60	1.550	4	4.28
Backward	48/60	0.200	1	8.56
Forward + trim	43/60	0.283	1	3.54
2-swap, backward start	59/60	0.017	1	50.80
2-swap, forward start	57/60	0.050	1	14.19
2-swap, priority queue	58/60	0.033	1	19.15
3-swap	60/60	0.000	0	169.87

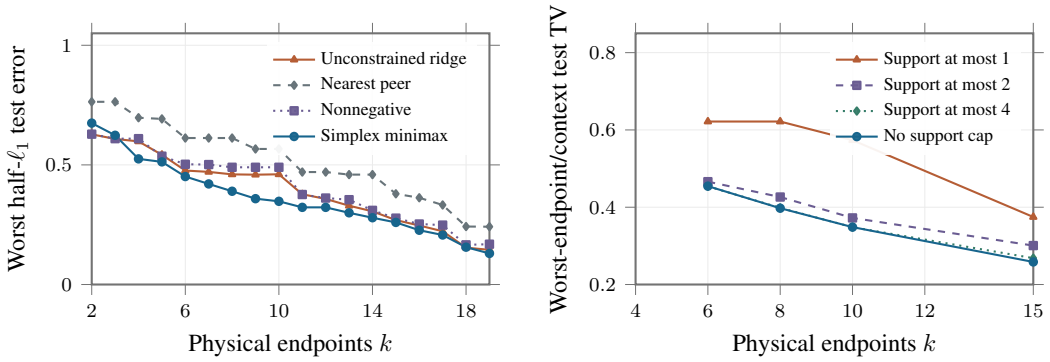
mismatch for that tested decoder, not the impossibility of a projected, renormalized, nonlinear, or otherwise better-tuned alternative.

The archived PCA and NMF fits optimize a low-rank reconstruction criterion different from worst-endpoint/worst-context TV and are evaluated after fitting. Their held-out scores are diagnostic references, not mathematical lower bounds on all physical subsets. We retain their source tables but do not label their curves as compression floors. Only the completed exact fitting-objective searches support the empirical infeasibility statements in the main text.

Sparse reconstruction limits the number of retained endpoints used by a particular virtual endpoint; physical-set size instead limits how many models must be available to support the whole ecosystem. These are different costs. The archived support-cap results make this distinction measurable, but a support-limited heuristic result should not be presented as an exact sparse optimum. Similarly, cost-aware selection is relevant when inference times differ, but the main cardinality objective intentionally uses equal call counts to isolate behavioral coverage.

C.6 Controlled constructions

The archived controlled checks include duplicated simplex extremes and a nine-point curved arc. The duplicated-extreme construction requires at least one representative of each extreme while every duplicate is individually redundant. Deleting all duplicates leaves no retained set, which is infeasible rather than a probability-valued reconstruction with infinite TV. In the curved-arc construction, at tolerance 0.005, independent removal leaves two endpoints with reported error 0.055, while an exact feasible five-endpoint set exists. The same geometric mismatch appears without duplicate models. These checks validate specific behaviors of the implementation; they do not eliminate every possible measurement or model-class explanation for the real-panel results.



(a) Tested reconstruction rules

(b) Restricting witness support

Figure 6: **The declared reconstruction class matters.** (a) Same retained sets, different fitted rules; the axis is half-ℓ_1 error because unconstrained outputs need not be probabilities. (b) Recorded support-cap comparisons. These ablations do not establish optimality over every probability-preserving decoder or every sparse support.

C.7 Risk-bound diagnostics are not population certification

The recorded risk-control study compares empirical thresholds and confidence-bound rules, including empirical Bernstein bounds (Maurer & Pontil, 2009), over 120 repartitions of the same scored data. It is useful as a conservativeness diagnostic. It is not 120 independent replications, and a zero count of held-out exceedances is not a distribution-free certification result. In addition, the implementation's certification summary averages at the source-task level whereas a test summary averages pairs. Those estimands must agree before a population risk statement can be attached to the empirical procedure.

For clarity, a valid upper-bound argument would freeze S, W independently of the certification sample and define the same bounded per-source-task loss throughout. For a fixed candidate, let r_ℓ be its true error in endpoint–context coordinate ℓ, and let U_ℓ be a marginal upper bound satisfying $\Pr(r_\ell > U_\ell) \le \delta$. A bound on the maximum can then use

$$\Pr\left(\max_\ell r_\ell > \max_\ell U_\ell\right) \le \delta, \tag{26}$$

since the event implies failure for a fixed coordinate attaining the maximum true risk. This is not the same as constructing a simultaneously valid box for every coordinate. If many candidate sets are selected using the same certification data, candidate-selection multiplicity must still be controlled. Learning candidates separately from testing their risk, with an appropriate multiple-testing rule, is the principle developed by Angelopoulos et al. (2025). A theoretical risk statement concerns $\Pr(\text{release and unsafe})$, not automatically $\Pr(\text{unsafe} \mid \text{release})$.

We do not use the archived diagnostic to claim population-safe retirement, and it does not strengthen the exact empirical cardinality statements. The empirical optimization, fitted held-out risks, and statistical risk-control problem remain distinct.