RESPONSESCOPE: MEASURING FINITE-RESPONSE DISTORTION IN COMPRESSED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model compression is commonly evaluated using clean-task performance and static teacher–student agreement, but these criteria do not characterize responses to finite, task-valid input changes. We study three related questions: how to measure this behavioral drift, predict compression vulnerability before full recovery and final utility evaluation, and preserve response behavior at a fixed model footprint. ResponseScope (RS) is a teacher-normalized finite-response distortion measure. The evaluation spans six datasets, nine teacher–student architecture pairs, four compression families, three storage budgets, 540 recipe means, and 2,700 independently stored seed checkpoints. Combining teacher layer fragility with operator-local distortion and applicable initialization compatibility yields task-equal leave-one-dataset-out $R^2 = .687$ and MAE 1.800 percentage points for predicting utility damage prospectively. After compression, adding measured RS to a scale- and endpoint-matched baseline raises R^2 from $.587$ to $.778$ and reduces selection regret from 1.260 to $.688$ points. At footprint ratio $.25$, IF-Reg reduces mean held-out RS from $.440$ to $.257$ and improves the task-specific utility composite by 3.017 points relative to augmented logit distillation. A baseline matched for endpoints, teacher scale, and margin explains part of this gain. These results connect the measurement, prediction, and preservation of interventional fidelity.

1 INTRODUCTION

Clean accuracy and model storage do not determine how a compressed model responds to a finite input change. A bird classifier may retain its clean predictions while responding differently to a background change; a sentiment model may change its response to a paraphrase; a sequence classifier may react differently to a bounded temporal warp. Whether each transformation preserves the task must be assessed separately and empirically.

Existing compression and behavior-transfer methods match selected aspects of model outputs, internal representations, or local behavior. We study interventional fidelity before recovery and after compression by predicting vulnerable modules and evaluating the response behavior of compressed candidates. ResponseScope (RS) measures finite-response distortion for fixed-budget candidate selection; IF-Reg uses response matching with endpoint and margin anchors to preserve this behavior during recovery.

The benchmark uses checkpoint-linked records and recipe identities fixed before scoring; Appendix I.1 documents the reconstruction audit.

Contributions. Our contributions are fourfold. (1) *Measurement:* RS quantifies compression-induced finite-response distortion under validated interventions and distinguishes candidates with nearly equal clean performance. (2) *Prospective prediction:* teacher layer fragility and operator-local distortion predict vulnerable modules and configuration-level utility. (3) *Post-compression evaluation:* measured RS adds held-out predictive and selection value beyond endpoint, scale, and compression descriptors. (4) *Preservation:* IF-Reg combines response matching with endpoint and margin anchors to improve held-out response fidelity and task utility. Figures 2–4 summarize the results; Figure 1 illustrates the post-compression branch.

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

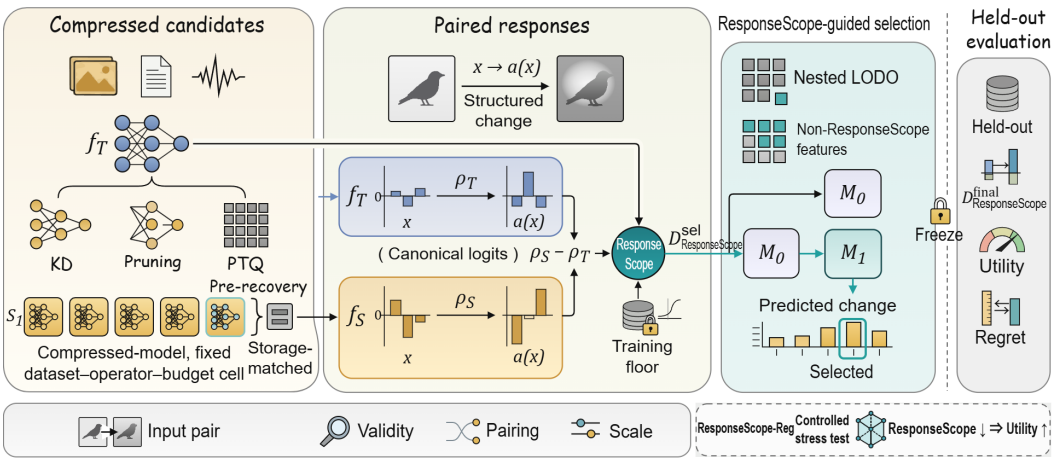


Figure 1: **Post-compression evaluation branch of the full framework.** Existing candidates are compared within a dataset–operator–budget cell. A training-frozen scale normalizes paired response differences; M_1^+ adds selection-stage RS to the endpoint- and scale-matched M_0^+. The choice is frozen before final-response, utility, and regret evaluation. Prospective layer-fragility prediction and IF-Reg recovery are defined in Section 4.

2 Related Work

Knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015), pruning (Han et al., 2015; Molchanov et al., 2019; Singh & Alistarh, 2020), quantization (Jacob et al., 2018; Banner et al., 2019; Dong et al., 2019; 2020), and low-rank approximation (Denton et al., 2014) alter different components of a model and its inference state. Prior work studies derivative matching (Czarnecki et al., 2017; Srinivas & Fleuret, 2018), decision-boundary distillation (Heo et al., 2019), representation comparison and transfer (Kornblith et al., 2019; Tian et al., 2020; Ojha et al., 2023; Liu et al., 2023), equivariance transfer (Nix et al., 2022), robustness distillation (Zhou et al., 2023), and behavior induction through representation structure (Pogoncheff & Beyeler, 2026). Our comparison focuses on teacher-normalized finite changes, their predictive value beyond endpoint and scale descriptors, and their use in recovery under shared intervention and footprint requirements.

Performance estimation without target labels. Average Thresholded Confidence (ATC; Garg et al., 2022) fits a threshold on labeled source data and estimates accuracy from unlabeled target predictions. Prediction-matrix approaches (Deng et al., 2023) estimate accuracy using target confidence and class-dispersity summaries, including the nuclear norm. ResponseScope instead selects compressed candidates at a fixed budget using source-task utility supervision, with task-specific utility as the endpoint. These differences in available information and estimands preclude a direct performance ranking against these methods; Appendix A.1 specifies the comparison.

Appendix A compares the behavior-matching scope; Appendix B states the analytical limits.

3 Finite-Response Distortion

3.1 Footprint and the intervention-validity contract

Consider a single-label classification task with $K_c \geq 2$ classes, a teacher f_T, and a compressed model $f_S = f_{\mathcal{C},B}$ produced by operator $\mathcal{C}$ at storage budget B. A model may output either K_c logits or, for a binary task, a scalar log-odds value. The deployable storage is

$$B_{\text{total}} = B_{\text{parameters}} + B_{\text{metadata}} + B_{\text{indices}} + B_{\text{untouched}}, \tag{1}$$

where embeddings, heads, normalization, biases, residual tensors, codebooks, scales, zero-points, and index structures are counted whenever inference requires them. Eligibility requires operator-specific records of the complete serialized footprint; Section 5 gives the realized target ranges.

An intervention a maps x to $a(x)$. In the frozen native protocol, an operational validator $\widehat{V}$ combines construction rules, automated checks, and blinded adjudication. We tag training-family, same-family held-out-variant, different-family held-out, and external-or-benchmark-shift records separately. The latter are further tagged as natural/domain-origin, perturbation-augmented, mechanism-overlapping, mechanism-disjoint, or mixed; "external" is not shorthand for natural or mechanism-independent.

3.2 Canonical response residual

All response metrics operate on a zero-sum canonical logit vector. Let $o_f(x)$ denote a raw pre-softmax logit vector or a scalar binary log-odds; probability-only outputs are ineligible unless the release provides a separately specified, numerically clipped inverse-link rule. Define the canonicalizer

$$\phi(o_f) = \begin{cases} H_{K_c} o_f, & o_f \in \mathbb{R}^{K_c},\ K_c \geq 2, \\ [-o_f/2,\ o_f/2]^\top, & o_f \in \mathbb{R} \text{ is a scalar binary log-odds.} \end{cases} \tag{2}$$

Here $H_{K_c} = I_{K_c} - K_c^{-1}\mathbf{1}\mathbf{1}^\top$ and $\Phi_f(x) = \phi(o_f(x))$. The scalar branch is the unique zero-sum two-logit representation whose class-1 log-odds equals o_f; treating that scalar as $K = 1$ would instead give $H_1 = 0$ and make every response degenerate. We set $\bar{z}_f = \Phi_f / \tau_{\text{ResponseScope}}$ with the common primary-analysis temperature $\tau_{\text{ResponseScope}} = 1$.

The present analysis defines ResponseScope for single-label classification. Extending the measure to independent sigmoid outputs requires label-wise canonicalization, weighting, and intervention-validity rules; these choices define a separate estimand.

The finite response and its root-mean-square teacher magnitude are

$$\rho_f(x,a) = \bar{z}_f(a(x)) - \bar{z}_f(x), \qquad r_T(x,a) = \frac{\|\rho_T(x,a)\|_2}{\sqrt{K_c}}. \tag{3}$$

We estimate the response-scale floor from accepted training-family pairs and freeze it before candidate evaluation. A dataset–teacher cell is included when its training response scale is non-degenerate. This rule keeps the normalized measure interpretable for the evaluated transformations, including cases in which near-invariance is an appropriate teacher response. Operationally, for every dataset–teacher cell, $q_{.1}^{\text{train}} > 0$ is the 10th percentile of r_T over accepted training-family pairs. A cell is excluded if this quantile is zero; the floor is never recomputed using held-out candidate data. All nine intended dataset–teacher cells pass this rule; three deliberately zero-floor unit-test cells are rejected separately. Training scales, stage-level floor-active counts, and strict/near-zero counts are reported in Appendix E.1. The stabilized teacher scale, normalized residual, and robust distortion are

$$s_T^\star(x,a) = \max\{r_T(x,a), q_{.1}^{\text{train}}\}, \qquad e_{TS}(x,a) = \frac{\rho_S(x,a) - \rho_T(x,a)}{s_T^\star(x,a)}, \tag{4}$$

$$D^{\mu}_{\text{ResponseScope}}(T,S) = \mathbb{E}_{P,\mu}\left[K_c^{-1} \sum_{k=1}^{K_c} H_\delta(e_{TS,k}) \right]. \tag{5}$$

What the finite difference adds. Let $r_x = \bar{z}_S(x) - \bar{z}_T(x)$ and $r_a = \bar{z}_S(a(x)) - \bar{z}_T(a(x))$. Then

$$\rho_S - \rho_T = r_a - r_x, \qquad \|r_a - r_x\|_2^2 = \|r_a\|_2^2 + \|r_x\|_2^2 - 2\langle r_a, r_x \rangle. \tag{6}$$

The cross term measures uncentered endpoint-error alignment, not centered covariance: no endpoint mean is subtracted. This bilinear identity applies to squared error, not directly to Huber loss. ResponseScope is a deterministic summary of paired logits, not new information absent from them. Coupling retains cross-endpoint structure, teacher normalization reweights it by $s_T^\star$, and Huberization limits tail influence. We use the Huber loss (Huber, 1964): $H_\delta(t) = \frac{1}{2}t^2$ for $|t| \leq \delta$ and $H_\delta(t) = \delta(|t| - \frac{1}{2}\delta)$ otherwise. We apply it componentwise and write $h_\delta(v) = K_c^{-1}\sum_k H_\delta(v_k)$. The training-derived floor prevents a small set of near-zero teacher responses from dominating the main ratio. A common positive rescaling of teacher and student logits also rescales $q_{.1}^{\text{train}}$ and therefore leaves e_{TS} unchanged. Independent teacher and student rescaling is not an invariance of ResponseScope. We therefore retain $\tau_{\text{ResponseScope}} = 1$ for the primary metric and report a separate scale-sensitivity analysis after fitting each model's temperature using ID validation data only (Guo et al., 2017).

The primary metric uses $\delta = 1$, $\tau_{\text{ResponseScope}} = 1$, and a weighted pair mean of class means, with the floor computed by a linear 0.10 training quantile. Appendix D.1 specifies the evaluation identity, weak-response diagnostics, and candidate-span summary.

Analytical scope. Appendix B derives a response-risk decomposition for losses that are Lipschitz in canonical logits under the defined label-preserving transformation distribution, with finite displayed expectations, and gives the exact bridge $D^{(2)} = 2D_{\text{RS}} + T_\delta$. The analysis identifies how endpoint residuals, the teacher-scale moment, and squared response distortion enter that risk relation. A numerical squared-risk bound would also require a reported residual-tail term (or a suitable residual bound) and the teacher-scale quantity. Neither is estimated here.

4 Prospective Prediction and Behavior Recovery

4.1 Layer interventional fragility before full evaluation

For layer representation $u_\ell = (h_\ell(x), h_\ell(a(x)))$, let $J_\ell^\rho = \partial \rho_T / \partial u_\ell$. We estimate the teacher's layer interventional fragility and the local distortion caused by operator $\mathcal{C}$ as

$$F_\ell = \mathbb{E}\|J_\ell^\rho\|_F^2, \qquad e_{\ell,\mathcal{C}} = \frac{\mathbb{E}\|\delta u_{\ell,\mathcal{C}}\|_2^2}{\mathbb{E}\|u_\ell\|_2^2 + 10^{-8}}. \tag{7}$$

The operator-conditioned score is $P_{\ell,\mathcal{C}} = F_\ell e_{\ell,\mathcal{C}}$ and $P_{\mathcal{C}} = \sum_\ell P_{\ell,\mathcal{C}}$. Fragility is estimated using 32 Rademacher probes over 512 training calibration pairs. The operator preview replaces one module at a time; for distillation, initialization compatibility is computed from the fixed student initialization rather than final trained weights. These features serve as empirical ranking proxies; they do not equal the full network's response risk. Section 5 specifies their evaluation-time information boundary.

4.2 IF-Reg

Let $r_x = \bar{z}_S(x) - \bar{z}_T(x)$, $r_a = \bar{z}_S(a(x)) - \bar{z}_T(a(x))$, and let $\Delta m_S - \Delta m_T$ denote the student–teacher difference in the label-margin response. The fixed primary recovery objective is

$$\mathcal{L}_{\text{IF-Reg}} = \mathcal{L}_{\text{AugSup}} + 0.5 h_1(r_x) + h_1\left(\frac{r_a - r_x}{s_T^\star}\right) + 0.25 H_1(\Delta m_S - \Delta m_T). \tag{8}$$

$\mathcal{L}_{\text{AugSup}}$ averages supervised losses at both endpoints. The remaining terms anchor the clean endpoint, match the normalized finite response, and preserve the response of the task margin. Teacher quantities serve as stop-gradient targets, and probability KD remains a separate baseline. Recovery starts from the same initial candidate and uses the same serialized footprint, valid intervention pairs, teacher-query access, and update budget as the matched baselines. For quantization, recovery includes quantization-aware optimization and re-export of the low-precision object; this procedure is not training-free post-training quantization.

5 Experimental Design

Decision setting and information access. Prospective prediction uses teacher fragility, operator-local previews, applicable initialization compatibility, and the declared footprint; the fully recovered student, final RS, and final utility are unavailable. Post-compression selection operates on already generated candidates. Source tasks supply utility labels for meta-predictor fitting and inner-fold tuning; the held-out target provides only ID-validation features and accepted selection-response pairs. The selector is fixed before the target final utility, final response pairs, and oracle regret are revealed. $D^{\text{sel}}_{\text{ResponseScope}}$ uses validation-family pairs before selection, whereas $D^{\text{final}}_{\text{ResponseScope}}$ uses disjoint, locked pair IDs afterward (Table 7). Final target outcomes never inform fitting, tuning, or selection.

The six tasks are Waterbirds, CIFAR-100, CivilComments, Amazon Reviews, UCI-HAR, and ECG5000. Together, they comprise nine teacher–student architecture pairs. Crossing those pairs with distillation, structured pruning, quantization, and low-rank compression at footprint targets .500, .250, and .125 gives 108 architecture–operator–budget cells. Each cell contains five recipes and each

recipe has independently stored checkpoints for seeds $\{7, 19, 31, 43, 59\}$. Metrics are computed per seed and then averaged, yielding 540 recipe-level candidates from 2,700 seed checkpoints; logits are not averaged, and the best seed is not selected. Dataset-level summaries first average budgets, operators, and architecture pairs within each dataset, then weight the six datasets equally. Appendix Tables 6, 7, and 5 provide details on endpoints, information use, and recipes.

CIFAR-100-C is an endpoint rather than a seventh task, and ECG5000 uses official test macro-F1 without a separate natural-shift endpoint. The checkpoint-linked audit in Table 26 documents the identity and attribution links. For task d and candidate j, the target is $Y_{\text{utility},d,j} = 100(\text{Score}^{\text{utility}}_{T,d} - \text{Score}^{\text{utility}}_{S_{d,j},d})$ in task-endpoint percentage points; larger values mean greater damage. Accuracy- and macro-F1-based differences are comparable as within-task ranking targets, but their equal-dataset mean is a descriptive composite rather than a common-scale effect.

Nested LODO analysis. Let X_{static} contain ID drop, recorded footprint ratio, operator, and model size, and let $B_{\text{non-ResponseScope}}$ be the prespecified group of non-RS behavior metrics. The primary nested comparison uses the same model family, folds, preprocessing, and tuning budget (Cawley & Talbot, 2010):

$$\begin{aligned} M_0^+ : \quad & Y_{\text{utility}} \sim X_{\text{static}} + B_{\text{non-ResponseScope}} + D_{x,\text{norm}} + D_{a,\text{norm}}, \\ M_1^+ : \quad & Y_{\text{utility}} \sim M_0^+ + D_{\text{ResponseScope}}. \end{aligned} \tag{9}$$

Here $D_{\text{ResponseScope}}$ is the native `responsescope-v1.0.0` value for each candidate. The comparison tests whether ResponseScope adds held-out predictive and selection value beyond prespecified aggregate features; it does not introduce pairing for the first time. The baseline already includes raw, coupled, symmetric-normalized, endpoint, probability, direction, calibration, Jacobian, representation, and operator-native descriptors, plus the two independently normalized endpoints. A stratified-shuffle version adds an RS feature of the same form after breaking its candidate-specific response coupling. The six outer folds hold out complete datasets and use training-fold-only preprocessing and tuning; the comparator contract is in Appendix D.4.

The dataset is the highest-level independent unit; cells and their five candidates are nested within it, and the six LODO fits share training tasks. Five complete refits, fixed-panel bootstrap (Efron, 1979) and endpoint recomputation, pair subsampling, and all 15 leave-two-datasets-out splits assess stability of the specified panel. Endpoint resampling follows each task's group, corruption, or class hierarchy and recomputes the task-level endpoint rather than treating it as an additive per-input loss.

Utility-alignment and transfer analyses. The supplied evaluation record additionally prespecifies (i) prospective layer- and configuration-level prediction; (ii) mixed-effects and partial-rank analyses controlling for ID damage, realized footprint, model size, operator, dataset, and architecture; (iii) matched comparisons for five learner families; (iv) IF-Reg comparisons at matched footprint, query access, and optimization budget; and (v) leave-one-modality-out and five frozen external settings comprising 60 cells and 300 recipe means. Calibration, stratified shuffle, invalid-intervention, endpoint-sufficient, weak-teacher, and zero-headroom analyses are mechanism controls. All numerical findings are drawn from author-supplied records of executed experiments. Manuscript preparation does not independently rerun those experiments or recompute their results.

6 Results

6.1 RQ1: Finite responses reveal behavior hidden by clean accuracy

At footprint .25, prespecified near-ID pairs (clean-score difference at most .10 points) show clear differences in response and utility (Figure 2a). Waterbirds R02/R04 have ID scores 91.10/91.13 but RS .21/.47 and final utility 79.40/74.10. Amazon Reviews and UCI-HAR show similar differences despite nearly equal clean scores (Table 21). These pairs motivate, but do not replace, the analysis of the full panel.

Compression families also have different response profiles. Distillation, pruning, quantization, and low rank have mean RS .36, .46, .52, and .42, respectively. Pruning has the largest recorded direction error (.24), whereas quantization has the largest relative-magnitude error (.61). These profiles describe the evaluated panel rather than intrinsic properties of the operators.

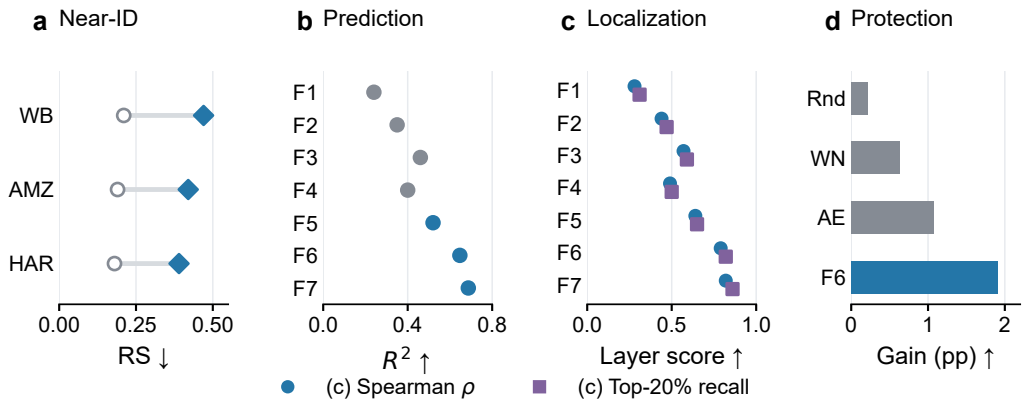


Figure 2: **Measurement and prospective prediction.** (a) Near-ID candidate RS pairs; circles/diamonds follow recipe order in Table 21, which gives clean scores and utility. WB, AMZ and HAR denote Waterbirds, Amazon Reviews and UCI-HAR. (b,c) F1–F7: weight norm, weight reconstruction, activation reconstruction, CKA, fragility, $F_\ell e_{\ell,\mathcal{C}}$, and the latter with initialization compatibility. (d) Fixed-footprint protection: random (Rnd), weight norm (WN), activation error (AE) and F6. No intervals are supplied for these point estimates.

6.2 RQ2: Prospective fragility predicts and localizes damage

Combining teacher fragility with operator-local distortion and applicable initialization compatibility improves prospective prediction (Figure 2b,c; Table 14), reaching layer Spearman .82, top-20%-fragile-layer recall .86, configuration $R^2 = .687$, and MAE 1.80 points. Across the six held-out datasets, R^2 ranges from .63 to .74 and MAE from 1.35 to 2.25 points.

The ranking supports targeted protection under a fixed model footprint. At a total footprint ratio of .25, protecting the modules ranked by $F_\ell e_{\ell,\mathcal{C}}$ lowers final RS from .470 without targeted protection to .304 and improves utility by 1.91 points. The strongest comparator, activation-error protection, yields 1.08 points (Figure 2d; Table 22). Storage preserved in selected modules is offset by savings elsewhere, so the gain does not require a larger model.

6.3 RQ3: Post-compression RS improves prediction and selection

RS adds predictive value beyond the matched endpoint and scale descriptors (Figure 3; Table 18). Stratified-shuffled RS remains close to M_0^+ (regret 1.245 versus 1.260), whereas measured RS reduces regret to .688, supporting the predictive contribution of candidate-specific response values. Across the 108 cells, $M_0^+ \to M_1^+$ lowers task-equal LODO MAE by .883 points, raises R^2 by .192, and lowers selection regret by .572 points. The corresponding six-task bootstrap intervals are [.733,1.050], [.162,.225], and [.472,.677]. All held-out datasets improve. Direct minimum-RS selection has .890 regret and 54/108 top-1 choices, versus 1.870 and 32/108 for ID-only selection; the learned M_1^+ selector reaches .688 and 64/108 (Figure 5c).

The matched gain is positive for all five evaluated learners, with varying effect sizes (Figure 5a; Table 19), so the observed increment is not specific to ridge regression. Seven of 108 decisions increase regret; Figure 3d and Table 23 retain all decision outcomes.

6.4 RQ4: IF-Reg preserves finite-response behavior

At footprint .25, IF-Reg lowers task-equal held-out RS from .440 under augmented logit KD (AKD-LM) to .257, a 41.67% ratio-of-means reduction, and improves the task-specific utility composite by 3.017 points (95% interval [2.400,3.683]). All six tasks and all four compression families improve (Figure 5b); Tables 16 and 20 report the individual gains.

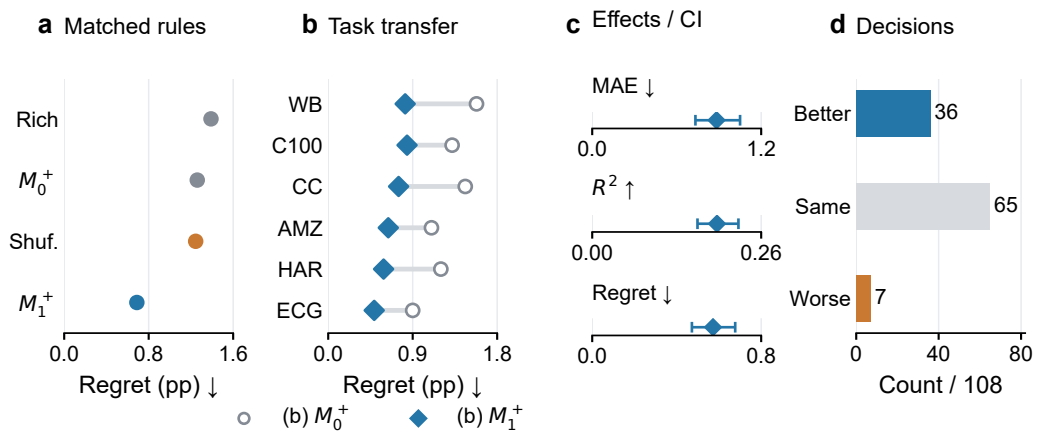


Figure 3: **Incremental predictive and selection value.** (a) Rich non-RS baseline, M_0^+, shuffled RS, and M_1^+. (b) Six held-out datasets: WB = Waterbirds, C100 = CIFAR-100, CC = CivilComments, AMZ = Amazon Reviews, HAR = UCI-HAR, ECG = ECG5000. (c) Improvements on independent axes; MAE and regret changes are in pp. Intervals are six-task fixed-panel bootstrap 95% intervals for paired improvements, not intervals for the means in (a,b). (d) All 108 decision outcomes, including seven worsened choices.

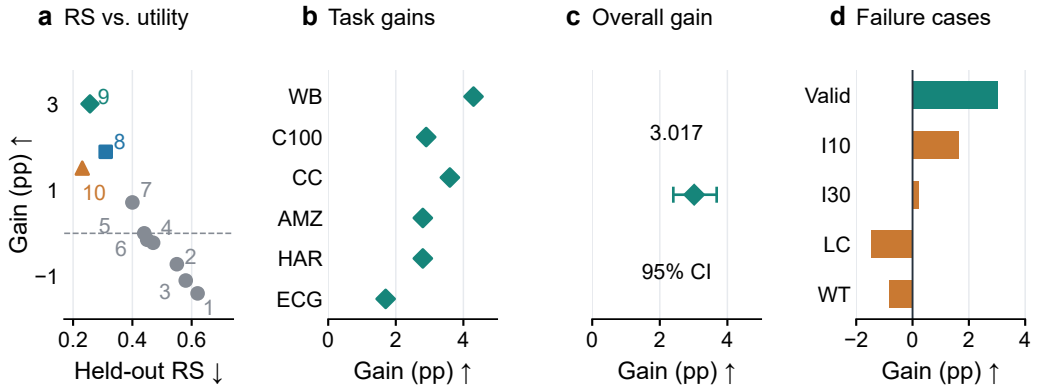


Figure 4: **Recovery and its limits at footprint .25.** (a) Method means: 1 native, 2 augmented supervision, 3 clean KD, 4 augmented probability KD, 5 AKD-LM, 6 consistency, 7 boundary, 8 scale-matched endpoint+margin, 9 IF-Reg, 10 no-margin ablation. (b) Six task gains (dataset abbreviations as in Figure 3). (c) Task-equal IF-Reg gain with its supplied 95% interval. (d) Valid, 10%/30% invalid pairs (I10/I30), label-changing (LC), and weak-teacher (WT) controls. Gains are relative to AKD-LM in task-utility composite pp; only (c) has an interval. Lower RS need not improve utility. The controlled linear-head study is separate.

Figure 4 relates recovery fidelity to task utility. The strongest matched alternative already includes teacher-scale-normalized endpoints and the same margin term. It reaches RS .310 and utility gain 1.90; IF-Reg retains a 1.117-point utility advantage. The factorial ablation also shows why lower RS need not imply higher utility: removing the margin produces lower RS (.231) but only 1.52 points of utility gain and a $-.28$-point mean ID change. These results support combining response matching with endpoint and margin anchors.

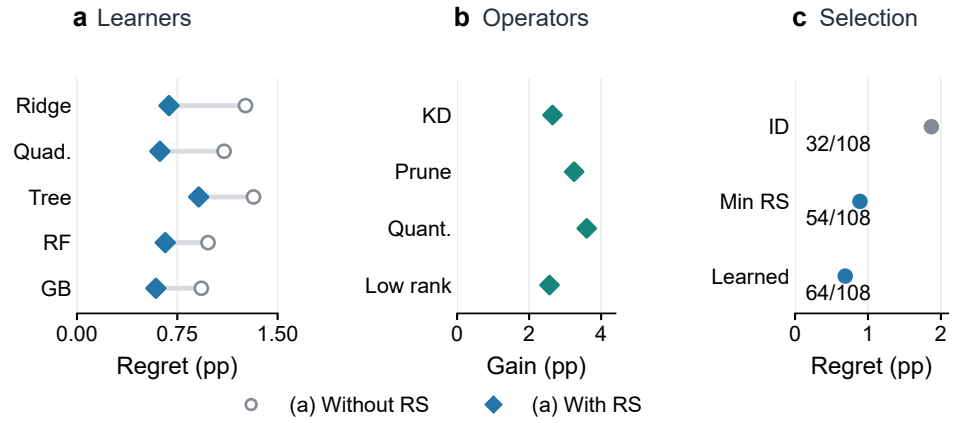


Figure 5: **Robustness and alternative selection rules in the core benchmark.** (a) Regret with and without RS for ridge, quadratic ridge (Quad.), shallow tree, random forest (RF), and gradient boosting (GB). (b) IF-Reg utility gains over AKD-LM across four compression families at footprint .25. (c) Regret and annotated top-1 counts for ID-only, minimum-RS, and learned selection over 108 decisions. All marks are point estimates without supplied intervals. These learner results are distinct from the reserved-task study in Appendix N.

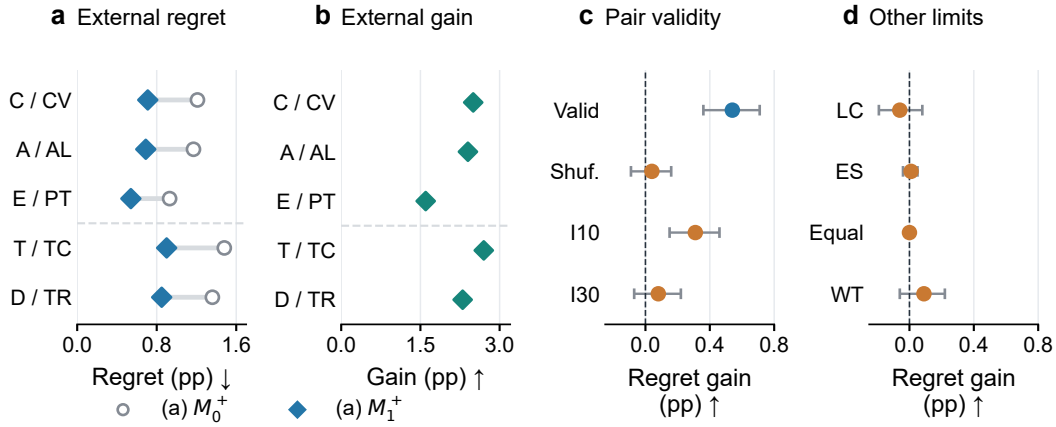


Figure 6: **Frozen external evaluation and selection limits.** (a,b) Regret and IF-Reg gain for five settings, 12 cells each: C/CV = CIFAR-100/ConvNeXt-T, A/AL = Amazon/ALBERT, E/PT = ECG5000/PatchTST, T/TC = TON-IoT/TCN, D/TR = DoHBrw2020/Transformer. The separator marks architecture versus traffic tests. (c,d) RS regret improvements and reported 95% intervals: valid, shuffled, 10%/30% invalid (I10/I30), label-changing (LC), endpoint-sufficient (ES), equal-utility, and weak-teacher (WT) controls. Zero denotes no benefit. Records are separate from the core benchmark and Figure 4d's recovery controls.

6.5 Generalization and Limits

Prospective leave-one-modality-out R^2 is .60, .62, and .65 for vision, text, and time series; corresponding M_1^+ regret is .93, .83, and .74, below M_0^+ in every modality. The separate extension fixes fitting and tuning before accessing external utility: three architecture changes and two traffic settings, 60 cells, 300 recipe means, and 1,500 seed checkpoints. All five settings improve (Figure 6a,b): mean regret 1.230→.738, IF-Reg gain 2.300. The evidence supports these transfers but does not establish unrestricted deployment (Appendix O).

Figure 6c,d reports *selection regret improvement*, not the *recovery utility gain* in Figure 4d. Shuffling removes the observed improvement, while invalid interventions attenuate it. The intervals include zero for label-changing transformations, weak teachers, and endpoint-sufficient utility; equal-utility candidates leave no headroom. Recovery gains likewise diminish with invalid pairs and reverse under label-changing transformations; a weak teacher reduces utility despite lower RS (Figure 4d; Table 24).

6.6 Measurement budget and acquisition cost

The separate utility-alignment record in Appendix N evaluates six budgets of accepted pairs per candidate (Table 1). These measurements are distinct from the 108-cell core benchmark. With 32 pairs, regret is .79 versus the non-RS baseline of 1.23, a .44-point reduction (95% CI [.23,.64]). Increasing the budget to 64 pairs lowers regret to .70; the remaining increase to 120 pairs yields a further .03-point reduction. Thus, 32–64 accepted pairs recover most of the observed selection benefit.

Table 1: Pair-budget frontier in the separate utility-alignment record. The non-RS baseline regret is 1.23 pp.

Accepted pairs per candidate	8	16	32	64	96	120
Mean regret (pp)	1.18	.96	.79	.70	.68	.67

With cached paired logits and a frozen floor, RS costs $O(nK_c)$ arithmetic without further queries. Transformation construction and validation, inference, source supervision, and candidate training remain separate costs; RS need not be cheaper than direct evaluation. Pair counts therefore describe the measurement budget, not end-to-end computational cost (Appendix D.3).

7 Discussion and Conclusion

RS adds predictive and selection value beyond the tested endpoint- and scale-matched descriptors; fragility and local previews support prospective protection. IF-Reg combines response pairing with endpoint and margin anchors to translate behavioral fidelity into utility gains.

Valid pairs, informative teachers, and candidate utility differences define the operating conditions (Section 6.5). The measured budget frontier supports economical response measurement for compression selection (Section 6.6).

Statements

AI USE STATEMENT. OpenAI ChatGPT assisted with language polishing and manuscript organization, including the presentation of existing results, but did not conduct experiments. The authors remain responsible for verifying the text, results, and interpretations.

Reproducibility statement. The code and supporting materials for this work will be released at `https://anonymous.4open.science/r/P1-D6B7/`.

REFERENCES

Abdullah Alsaedi, Nour Moustafa, Zahir Tari, Abdun Mahmood, and Adnan Anwar. TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. *IEEE Access*, 8:165130–165150, 2020. doi: 10.1109/ACCESS.2020.3022862. URL https://ieeexplore.ieee.org/document/9189760/.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 437–442, 2013. URL https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7948–7956, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c0a62e133894cdce435bcb4a5df1db2d-Abstract.html.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10915–10924, 2022. doi: 10.1109/CVPR52688.2022.01065. URL https://openaccess.thecvf.com/content/CVPR2022/html/Beyer_Knowledge_Distillation_A_Good_Teacher_Is_Patient_and_Consistent_CVPR_2022_paper.html.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 491–500, 2019. doi: 10.1145/3308560.3317593.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 535–541, 2006. doi: 10.1145/1150402.1150464.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4278–4287, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/758a06618c69880a6cee5314ee42d52f-Abstract.html.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019. doi: 10.1109/JAS.2019.1911747. URL https://www.ieee-jas.net/en/article/doi/10.1109/JAS.2019.1911747.

Weijian Deng, Yumin Suh, Stephen Gould, and Liang Zheng. Confidence and dispersity speak: Characterizing prediction matrix for unsupervised accuracy estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 7658–7674, 2023. URL https://proceedings.mlr.press/v202/deng23e.html.

Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, volume 27, pp. 1269–1277, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/1adaeb993eba95859121a43ea61bd858-Abstract.html.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In *IEEE/CVF International Conference on Computer Vision*, pp. 293–302, 2019. doi: 10.1109/ICCV.2019.00038. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Dong_HAWQ_Hessian_AWare_Quantization_of_Neural_Networks_With_Mixed-Precision_ICCV_2019_paper.html.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18518–18529, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d77c703536718b95308130ff2e5cf9ee-Abstract.html.

Wesley M. Eddy. Transmission control protocol (TCP). RFC 9293, RFC Editor, 2022. URL https://www.rfc-editor.org/rfc/rfc9293.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=o_HsiMPYh_x.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1135–1143, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/ae0eb3eed39d2bcef4622b2499a05fe6-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge distillation with adversarial samples supporting decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3771–3778, 2019. doi: 10.1609/aaai.v33i01.33013771.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018. doi: 10.1109/CVPR.2018.00286. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jacob_Quantization_and_Training_CVPR_2018_paper.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://ieeexplore.ieee.org/document/61115/.

Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pgHNOcxEdRI.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. doi: 10.48550/arXiv.1907.11692. URL https://arxiv.org/abs/1907.11692.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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. doi: 10.1109/CVPR.2019.01152. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Molchanov_Importance_Estimation_for_Neural_Network_Pruning_CVPR_2019_paper.html.

Mohammadreza MontazeriShatoori, Logan Davidson, Gurdip Kaur, and Arash Habibi Lashkari. Detection of DoH tunnels using time-series classification of encrypted traffic. In *2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, Cloud and Big Data Computing, and Cyber Science and Technology Congress*, pp. 63–70, 2020. doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00026.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 188–197, 2019. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018/.

Arne Nix, Suhas Shrinivasan, Edgar Y. Walker, and Fabian Sinz. Can functional transfer methods capture simple inductive biases? In *Proceedings of the 25th International Conference on Artificial*

Intelligence and Statistics, volume 151 of *Proceedings of Machine Learning Research*, pp. 10703–10717. PMLR, 2022. URL https://proceedings.mlr.press/v151/nix22a.html.

Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? In *Advances in Neural Information Processing Systems*, volume 36, pp. 11037–11048, 2023. doi: 10.52202/075280-0487. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2433fec2144ccf5fea1c9c5ebdbc3924-Abstract-Conference.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. doi: 10.1109/CVPR.2019.00409. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Galen Pogoncheff and Michael Beyeler. BIRD: Behavior induction via representation-structure distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jbJGhHpwmJ.

Eric Rescorla. The transport layer security (TLS) protocol version 1.3. RFC 8446, RFC Editor, 2018. URL https://www.rfc-editor.org/rfc/rfc8446.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient second-order approximation for neural network compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18098–18109, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d1ff1ec86b62cd5f3903ff19c3a326b2-Abstract.html.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin Dogus Cubuk, Alex Kurakin, Han Zhang, and Colin A. Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4723–4731. PMLR, 2018. URL https://proceedings.mlr.press/v80/srinivas18a.html.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgpBJrtvS.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, pp. 87.1–87.12. BMVA Press, 2016. doi: 10.5244/C.30.87. URL https://bmva-archive.org.uk/bmvc/2016/papers/paper087/index.html.

Andy Zhou, Jindong Wang, Yu-Xiong Wang, and Haohan Wang. Distilling out-of-distribution robustness from vision-language foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32938–32957, 2023. doi: 10.52202/075280-1428. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/67f30132d98e758f7b4e28c36091d86e-Abstract.html.

A Operational scope comparison

A.1 Performance estimation and candidate selection: information access

An unlabeled final-domain sample and a label-preserving transformation of an ID sample are different observations. Neither the absence of labels nor the use of confidence scores makes their information budgets equivalent. The comparison below separates published accuracy-estimation objectives from the selection protocol considered here; it contains no new baseline measurements.

Table 2: Information access and decision targets. Extending an estimator to a selector is a protocol choice, not an already executed comparison.

Approach	Information used for estimation	Teacher needed?	Direct target and selection implication
ATC	Labeled source validation; candidate predictions on an unlabeled target-domain sample	No	Target classification accuracy; a same-budget selector would need an explicit candidate ranking and regret evaluation.
Prediction-matrix summary	Target-domain class-probability matrix; confidence and class dispersity	No	Accuracy estimation; this does not directly establish macro-F1, worst-group accuracy, or this paper's within-cell regret.
ResponseScope protocol	Source-task utility supervision; target ID validation and selection-transform pairs; no final-domain sample used for selection	Yes	Task-specific utility damage followed by selection within operator–budget cells; final labels and responses score the frozen choice.
Additional-target-access comparison	Same unlabeled final-domain inputs available to all eligible methods; source-only calibration/tuning frozen	Method-specific	A separate information condition whose accuracy, group endpoints, candidate decisions, and acquisition cost must be specified.

A direct baseline comparison requires matched target sample access, source supervision, preprocessing, tuning opportunities, candidate identities, and final endpoints. Source labels used for threshold fitting or source utility regression must remain within the appropriate source folds. If an estimator is restricted to ID or selection-transform samples, its estimate concerns that input distribution unless a transfer assumption is supplied. Conversely, granting final-domain samples only to that estimator changes the experiment. Average classification accuracy cannot replace macro-F1 or worst-group accuracy without justification: group/class information, denominators, and any endpoint-specific extension must be stated. The current records contain no such direct or additional-access comparison.

A.2 Behavior matching and analytical scope

Representative rows cover derivative matching (Czarnecki et al., 2017; Srinivas & Fleuret, 2018), boundary distillation (Heo et al., 2019), augmented KD (Beyer et al., 2022), functional transfer (Nix et al., 2022; Liu et al., 2023), robustness distillation (Zhou et al., 2023), sensitivity-aware compression (Dong et al., 2019; 2020; Singh & Alistarh, 2020), and BIRD (Pogoncheff & Beyeler, 2026).

Table 3: Comparison dimensions use explicit estimands; "not the target" is descriptive and not a claim of deficiency.

Work	Estimand and comparison	Evaluation and controls
Jacobian/Sobolev matching	*Finite change:* infinitesimal derivative. *Cross-operator:* not the target	*Candidate prediction:* no. *Storage:* not reported. *Held-out transform:* not the target. *Controls:* not reported.
Decision-boundary distillation	*Finite change:* boundary-focused. *Cross-operator:* not the target	*Candidate prediction:* no. *Storage:* not reported. *Held-out transform:* not the target. *Controls:* task-specific.
Augmented KD	*Finite change:* endpoint values. *Cross-operator:* limited	*Candidate prediction:* no. *Storage:* not the target. *Held-out transform:* family-dependent. *Controls:* not required.

Table 3 (continued)

Work	Estimand and comparison	Evaluation and controls
Equivariance/ functional transfer	*Finite change:* property-specific. *Cross-operator:* task/operator-specific	*Candidate prediction:* selection or transfer. *Storage:* not the target. *Held-out transform:* property-specific. *Controls:* method-specific.
Robustness distillation	*Finite change:* finite/OOD robustness. *Cross-operator:* task-specific	*Candidate prediction:* teacher/augmentation transfer. *Storage:* not the target. *Held-out transform:* benchmark-dependent. *Controls:* method-specific.
Sensitivity-aware compression	*Finite change:* curvature/sensitivity proxy. *Cross-operator:* operator-specific	*Candidate prediction:* candidate allocation. *Storage:* serialized or nominal. *Held-out transform:* not the target. *Controls:* allocation-specific.
BIRD	*Finite change:* structured behavior. *Cross-operator:* not the target	*Candidate prediction:* teacher selection. *Storage:* not the target. *Held-out transform:* held-out behavior. *Controls:* method-specific.
This work	*Finite change:* finite, dataset-specified difference. *Cross-operator:* three measurement operators; two small-model constraints	*Candidate prediction:* pre-recovery. *Storage:* serialized bytes for measurement only. *Held-out transform:* held-out variants/families, overlap tagged. *Controls:* invalid-pair, scale, and fixed-pool shuffle.

Scale and relation alternatives. ResponseScope is a pre-recovery criterion for task-designated clean–transformed pairs, whereas relation-preserving (Park et al., 2019) and behavior-transfer methods target other objects or training objectives. Appendix A makes these estimand differences explicit; they delimit the contribution without claiming that response matching itself is novel.

B Analytical boundary

The following statements concern the specified label-preserving transformation kernel and losses that are Lipschitz in canonical logits, assuming the displayed expectations are finite. They do not bound macro-F1 or worst-group accuracy directly, and do not establish performance on a natural domain whose distribution is not the defined transformation distribution.

Let Q_μ be the transformed distribution induced by P and μ, and let $D_\mu^{(2)} = \mathbb{E}_{P,\mu}[K_c^{-1}\|e_{TS}\|_2^2]$.

Lemma 1 (Response-risk decomposition)**.** *If the loss is L_R-Lipschitz in canonical logits, then* $|R_{Q_\mu}(S) - R_{Q_\mu}(T)| \le L_R(\epsilon_{\text{base}} + \sqrt{K_c\,\mathbb{E}[(s_T^\star)^2] D_\mu^{(2)}(T,S)})$*, where* $\epsilon_{\text{base}} = \mathbb{E}_P\|\bar{z}_S(x) - \bar{z}_T(x)\|_2$.

Lemma 2 (Huber–quadratic bridge)**.** *Let* $D_H = D^{\mu}_{\text{ResponseScope}}(T,S)$ *and* $T_\delta = \mathbb{E}[K_c^{-1}\sum_k (|e_{TS,k}| - \delta)_+^2]$*. Then*

$$D_\mu^{(2)} = 2D_H + T_\delta. \tag{10}$$

If $\|e_{TS}\|_\infty \le M$ *almost surely, then* $2D_H \le D_\mu^{(2)} \le C_{\delta,M} D_H$*, with* $C_{\delta,M} = 2$ *for* $M \le \delta$ *and* $C_{\delta,M} = M^2/[\delta(M - \delta/2)]$ *otherwise.*

Without bounded residuals or a reported T_δ, lowering Huber-ResponseScope alone does not imply a smaller squared-response risk bound. We do not estimate $D^{(2)}$, T_δ, or the teacher-scale moment, so these results provide no numerical bounds and serve only to delimit the empirical claim. The executed

comparison with a directional-Jacobian descriptor is likewise limited: with the same fixed base features, it yields $R^2 = .527$ and MAE 3.17 pp, versus .640 and 2.68 pp for native ResponseScope. This predictor comparison does not identify where a first-order approximation fails. We did not evaluate a continuous-input transformation-strength scan of $\rho_f(x, a)$ against $J_{\bar{z}_f}(x)(a(x) - x)$, so we make no claim that finite matching is generally superior to local matching; discrete text transformations are not assigned a continuous perturbation interpretation.

C PROOFS

Proof of Lemma 1. For a valid transformed input, $\bar{z}_S(a(x)) - \bar{z}_T(a(x)) = [\bar{z}_S(x) - \bar{z}_T(x)] + [\rho_S - \rho_T]$. Lipschitz continuity, the triangle inequality, and Cauchy–Schwarz yield

$$\mathbb{E}\|\rho_S - \rho_T\|_2 = \mathbb{E}[s_T^\star \|e_{TS}\|_2] \le \sqrt{\mathbb{E}[(s_T^\star)^2]\,\mathbb{E}\|e_{TS}\|_2^2} = \sqrt{K_c\,\mathbb{E}[(s_T^\star)^2] D_\mu^{(2)}}.$$

Applying the L_R-Lipschitz property and bounding the clean endpoint by ϵ_{base} proves the result. □

Proof of Lemma 2. For a scalar t, the Huber definition gives the pointwise identity

$$t^2 = 2H_\delta(t) + (|t| - \delta)_+^2.$$

Averaging over coordinates and inputs gives Equation 10. The tail is nonnegative, hence $2D_H \le D_\mu^{(2)}$. If $0 < |t| \le \delta$, then $t^2/H_\delta(t) = 2$. For $\delta < |t| \le M$, the ratio $t^2/[\delta(|t| - \delta/2)]$ is increasing in $|t|$ and is therefore at most $M^2/[\delta(M - \delta/2)]$. Taking the applicable pointwise constant and averaging proves the upper bound; the inequality is immediate at $t = 0$. □

D ADDITIONAL MEASUREMENT ANALYSES

This appendix specifies measurement diagnostics, matched comparators, and cost accounting. Appendix G reports the full benchmark prediction and recovery results; Appendix N contains the separately scoped utility-alignment and transfer analyses.

D.1 METRIC IDENTITY AND RESPONSE DIAGNOSTICS

Versioned native metric identity. The primary definition fixes float64 evaluation, zero-sum canonicalization, scalar-binary embedding, $\delta = 1$, $\tau_{\text{ResponseScope}} = 1$, linear 0.10 training quantiles, and a weighted pair mean of class means. Its version and run-linked receipts are frozen; complete identities appear in Appendix I and the artifact. The raw response error and floor-active (weak-response) diagnostic are

$$D_{\text{raw}}^{\mu}(T, S) = \mathbb{E}_{P,\mu}\left[K_c^{-1}\|\rho_S - \rho_T\|_2^2\right], \tag{11}$$

$$D_{\text{weak}}^{\mu}(T, S) = \mathbb{E}_{P,\mu}\left[h_\delta(e_{TS}) \mid r_T < q_{.1}^{\text{train}}\right]. \tag{12}$$

Writing $r_S = \|\rho_S\|_2/\sqrt{K_c}$, diagnostics partition pairs into three disjoint cases: teacher-weak ($r_T < q_{.1}^{\text{train}}$), teacher-active/student-weak ($r_T \ge q_{.1}^{\text{train}}$ and $r_S < q_{.1}^{\text{train}}$), and both-active ($r_T, r_S \ge q_{.1}^{\text{train}}$). Direction error is defined only for the both-active case; D_{weak} conditions only on teacher-weak pairs, and the primary ResponseScope measure retains all three groups. Relative-magnitude, symmetric-normalization, direction, and validation-only calibration quantities remain diagnostics rather than additional success criteria. Weak response is relative to a training quantile; it is distinct from $r_T = 0$ or numerical near-zero response. Near-invariance under a label-preserving change can be appropriate teacher behavior; a weak response indicates its scale, not teacher error.

Candidate response span. For the candidate set J_c in an eligible dataset–operator–budget cell c, the logged native value is $d_{c,j} = D_{\text{ResponseScope}}(T, S_{c,j})$. The within-cell span is

$$\text{Span}_c = 100\left(\frac{\max_{j \in J_c} d_{c,j}}{\min_{j \in J_c} d_{c,j}} - 1\right)\%, \qquad \min_j d_{c,j} > 0. \tag{13}$$

The within-cell span summarizes relative variation among candidates. Its interpretation depends on the minimum candidate distortion, so we report it together with the underlying distortion values rather than as a standalone standardized effect. The budget aggregation is specified in Appendix D.4.

D.2 SELECTION, CALIBRATION, AND PAIR-BUDGET ANALYSES

Matched endpoints, direct ordering, calibration, robustness, and pair-budget analyses are reported in Appendix N under its stated experimental units and estimands.

D.3 SELECTION COST: SHARED ACQUISITION AND MARGINAL RESPONSE AGGREGATION

Selection cost has three components: constructing the candidate panel, acquiring the information required by a rule, and fitting and applying that rule. Let C_{gen} include candidate training or calibration and serialization. For feature set $\mathcal{F}$, write

$$C_{\mathrm{total}}(\mathcal{F}) = C_{\mathrm{gen}} + C_{\mathrm{acq}}(\mathcal{F}; \mathcal{K}) + C_{\mathrm{fit}}(\mathcal{F}) + C_{\mathrm{select}}(\mathcal{F}), \tag{14}$$

where $\mathcal{K}$ specifies cached outputs, transforms, representations, and derivatives. C_{gen} is common when comparing rules on the same existing candidate panel; it must still be counted when estimating total workflow cost. Transformation construction and validity assessment belong to acquisition and may be amortized only over the candidates/tasks that actually share them. For the same cache and panel, the feature-acquisition increment is $\Delta C_{\mathrm{acq}} = C_{\mathrm{acq}}(\mathcal{F}_0 \cup \{D_{\mathrm{ResponseScope}}\}; \mathcal{K}) - C_{\mathrm{acq}}(\mathcal{F}_0; \mathcal{K})$. This difference avoids counting shared forward passes twice.

Table 4: Benchmark cost quantities and the current measurement boundary. No modeled seconds are presented as hardware measurements.

Component	Work that must be accounted for	Current evidence
Candidate generation	KD training, pruning/fine-tuning or PTQ calibration, plus serialization	Recipes described; comparable elapsed cost not measured
Feature acquisition	Transforms, validity assessment, inference, endpoint statistics; representation/derivative work where required	ResponseScope path only: .438–3.801 s per five-candidate cell
ResponseScope increment	Difference, common-scale normalization, Huber aggregation; include any uncached required outputs	Marginal elapsed time not isolated
Predictor and decision	Nested source-task tuning, fitting, and inference	End-to-end rule cost not measured

At 8, 16, 32, 64, 96, and 120 pairs per candidate, measured ResponseScope-path times for a five-candidate cell are .438, .677, 1.158, 2.124, 3.079, and 3.801 seconds, respectively. These times do not isolate aggregation from acquisition and exclude the full non-ResponseScope feature pipeline; neither their absolute values nor differences estimate ΔC_{acq} for M_1.

If the clean/transformed teacher and student logits and training floor are already available under the same protocol, native ResponseScope needs no new model forward passes or representation/Jacobian extraction. It requires $O(nK_c)$ arithmetic per candidate and can aggregate stored or streamed pairs using $O(K_c)$ working memory. This is an operation-count observation, not a near-zero runtime claim; loading outputs and additional fitting still incur costs. A static-plus-ResponseScope rule avoids the representation and derivative features of M_1, but a static rule that lacks paired outputs must incur their acquisition cost. Table 48 adds matched numerical timing with cached logits and with regenerated logits but resident inputs and generator parameters. It does not provide neural cold-start cost or isolate marginal aggregation. Neither experiment establishes a benchmark cost–utility optimum. The earlier assumed planning ledger is excluded from runtime evidence.

Acquisition decisions precede response diagnostics. Source-task utility supervision is an input to this workflow, not a free replacement for target labels. Target ID validation can be available even when final shift-specific utility requires additional labels or group annotations. This motivates the information boundary but establishes no measured cost advantage. Using the full ResponseScope span to decide whether to acquire ResponseScope is circular. An acquisition policy would require a cheaper screening signal or a pilot with explicit cost accounting; neither is validated here.

Table 5: Substantive candidate-generation settings recoverable from the 270 exported configuration records.

Operator	Five-candidate axis	Training/calibration access	Serialized object
Offline-logit KD	Dense student shapes: balanced-width, depth-preserving, early-capacity, late-capacity, and bottleneck-heavy; exported depth factors $1.00, .88, .94, 1.06, 1.12$ and panel-wide width range .324–.737	Hard-label/logit weights $.5/.5$, temperature 2, AdamW; 30 epochs for the two vision tasks and 8 otherwise	Dense student state dictionary
Structured pruning	Global-magnitude channel/attention-head pruning; target densities $\{.121, .123, .125, .127, .129\}$, $\{.246, .248, .250, .252, .254\}$, or $\{.496, .498, .500, .502, .504\}$ by budget	AdamW fine-tuning for 20 epochs on Waterbirds, CIFAR-100, UCI-HAR, and ECG5000 and 5 on the text tasks	CSR values and indices plus required metadata
Post-training quantization	Per-channel symmetric weights; weight/activation bits $16/16$, $8/8$, or $4/8$ by budget; group sizes $128, 64, 64, 32, 128$ for c01–c05	1,024 calibration samples; no weight-training stage recorded	Packed integers, scales, and required metadata

Teacher IDs identify ResNet-50 (Waterbirds), Wide-ResNet-40-2 (CIFAR-100), RoBERTa-base (CivilComments), RoBERTa-base (Amazon Reviews), a six-class TCN (UCI-HAR), and a 1D ResNet (ECG5000). Distilled students retain the corresponding dense ResNet, Transformer, or TCN family. Input preprocessing, realized parameter counts, and shared/frozen-module details are not present in the supplied configuration records. The export does not establish whether c01–c05 were assigned before utility scoring. Repeated PTQ group sizes do not prove duplicate candidates, but these partial fields cannot identify all five full recipes. For pruning, physical channel removal versus original-shape sparse storage, the meaning of density (retained channels or nonzero weights), and component byte counts remain unresolved; density must not be read as the serialized ratio. Candidate generation is not training-free: KD and pruning include optimization, whereas PTQ uses calibration. Operator and optimizer precedents are temperature-scaled KD (Hinton et al., 2015), structured pruning (Han et al., 2015; Molchanov et al., 2019), PTQ (Jacob et al., 2018; Banner et al., 2019), and AdamW (Loshchilov & Hutter, 2019); these citations identify method families, not missing implementation details.

Controlled recovery configurations and state provenance are specified in Appendices H and I.

Metric-identity tolerances, operation-level traces, and complete hash strings are retained in the artifact. These are engineering checks rather than experimental outcomes, so they are not presented in a results table.

D.4 Incremental predictive-value comparator and construct-validity diagnostics

For the secondary budget-level span summary, positive cell ratios are aggregated as

$$100\left\{\exp\left[|\mathcal{C}_B|^{-1}\sum_{c\in\mathcal{C}_B}\log\left(\frac{\max_j d_{c,j}}{\min_j d_{c,j}}\right)\right]-1\right\}\%.$$

This summary is undefined if any cell minimum is nonpositive; a small positive minimum can inflate it.

The fixed $B_{\text{non-ResponseScope}}$ block in Equation 9 contains the following eligible feature families, all computed on the same clean/ transformed pair manifest as ResponseScope and without target final-utility labels. Source-task utility labels supervise the outer-fold predictor:

1. raw response error and unnormalized coupled Huber response error;
2. symmetric-normalized response error, clean endpoint logit error, and transformed endpoint logit error;
3. signed probability-response ℓ_2 error and endpoint Jensen–Shannon divergence (Lin, 1991) (a KL divergence is not applied to a signed response vector);
4. direction error on a frozen nonzero-response mask;
5. absolute ECE, NLL, Brier, logit-norm, and validation-temperature shifts (Brier, 1950; Guo et al., 2017);
6. directional-Jacobian error (Srinivas & Fleuret, 2018) and $1-\mathrm{CKA}$ (Kornblith et al., 2019) only where input and representation alignment make them well-defined; and

7. operator-native damage scores: endpoint KD for distillation (Hinton et al., 2015), Hessian-weighted removed-weight energy for pruning (Singh & Alistarh, 2020), and HAWQ-style trace-weighted quantization error (Dong et al., 2020).

For compactness, let $r_f = \|\rho_f\|_2/\sqrt{K_c}$ and $\Delta p_f = p_f(a(x)) - p_f(x)$. The response and endpoint components are fixed by

$$\begin{aligned} D_{\text{coupled}} &= \mathbb{E}[h_\delta(\rho_S - \rho_T)], \qquad D_{\text{sym}} = \mathbb{E}\left[h_\delta\left(\frac{\rho_S - \rho_T}{\max\{(r_S + r_T)/2, q_{.1}^{\text{train}}\}}\right)\right], \\ D_x &= \mathbb{E}[h_\delta(\bar{z}_S(x) - \bar{z}_T(x))], \qquad D_{a(x)} = \mathbb{E}[h_\delta(\bar{z}_S(a(x)) - \bar{z}_T(a(x)))], \\ D_{\Delta p} &= \mathbb{E}[K_c^{-1}\|\Delta p_S - \Delta p_T\|_2^2], \\ D_{\text{JS}} &= \tfrac{1}{2}\mathbb{E}[\text{JS}(p_T(x), p_S(x)) \\ &\qquad + \text{JS}(p_T(a(x)), p_S(a(x)))]. \end{aligned} \tag{15}$$

The scale-matched endpoint control retains two independent features:

$$D_{x,\text{norm}} = \mathbb{E}\left[h_\delta\left(\frac{r_x}{s_T^\star}\right)\right], \qquad D_{a,\text{norm}} = \mathbb{E}\left[h_\delta\left(\frac{r_a}{s_T^\star}\right)\right]. \tag{16}$$

The two features are retained separately rather than replaced by their average. The completed primary comparison uses M_0^+ (M_0 plus both normalized endpoints) and M_1^+ (M_0^+ plus measured ResponseScope), with identical canonicalization, $s_T^\star$, Huber parameter, pairs, folds, and tuning budget. Table 18 reports the matched benchmark results; Appendices J and K retain separate historical mechanism studies. $D_{\text{dir}} = \mathbb{E}[1 - \cos(\rho_T, \rho_S)]$ uses the frozen mask in Appendix D.1. Directional-Jacobian and CKA rows must record their alignment and extraction hashes; ineligible values are recorded as missing, with reasons, rather than as zero. ECE, NLL, Brier, logit-norm, and fitted-temperature components use ID-validation data only. The complete feature list, signs, missingness rules, and transforms are fixed before held-out scoring. Operator-native features are normalized using training-fold information only, and the "best single non-ResponseScope" rule is chosen inside inner training folds. For leave-one-operator-out sensitivity, the pipeline uses only source-specified cross-operator mappings; otherwise the feature is omitted. The benchmark record retains fold identifiers, training-fold preprocessing, selected ridge values, candidate-level OOF predictions, and decision-level information boundaries for this ordering.

Machine-readable rows are separated into candidate prediction, candidate selection, scale-control, pair-budget, seed-refit, endpoint-recomputation, and leave-two-datasets-out ledgers. The selection ledger contains all 378 rule decisions and 1,890 candidate scores; it is not inferred from aggregate means.

Prediction and selection answer different questions. The specified predictor output would contain per-dataset out-of-fold errors for M_0 and M_1; the selection output contains regret relative to an evaluation-only oracle for ID-only, each source-specified single metric, ResponseScope, the non-ResponseScope omnibus, and the omnibus plus ResponseScope. The protocol forbids oracle labels or held-out ResponseScope from influencing selection. A result may support incremental prediction without supporting every single-metric ranking, and vice versa. Because ResponseScope may be an engineered transformation of existing endpoints especially compatible with a linear predictor, the full benchmark evaluates ridge, quadratic ridge, shallow trees, random forests, and gradient boosting; every learner shows a positive matched regret reduction, although effect sizes vary. Regret is first averaged across budgets, operators, and architecture pairs within each dataset and then averaged equally over the six datasets. Counts across 108 cells describe coverage; they do not represent 108 independent Bernoulli trials.

The protocol requires standardization, imputation, feature selection, and ridge selection over $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$ inside each outer training fold, using inner leave-one-training-dataset-out validation. For held-out dataset d, candidate set J_d, and model $m \in \{0, 1\}$, both

nested pipelines use the same scoring denominator:

$$\mathrm{MAE}_{m,d} = |J_d|^{-1} \sum_{j \in J_d} |Y_{d,j} - \widehat{Y}_{m,d,j}|,$$
$$R^2_{m,d} = 1 - \frac{\sum_{j \in J_d} (Y_{d,j} - \widehat{Y}_{m,d,j})^2}{\sum_{j \in J_d} (Y_{d,j} - \bar{Y}_d)^2}, \qquad \Delta R^2 = \frac{1}{6} \sum_d (R^2_{1,d} - R^2_{0,d}),$$
$$\mathrm{RelMAE} = \frac{\sum_d \mathrm{MAE}_{0,d} - \sum_d \mathrm{MAE}_{1,d}}{\sum_d \mathrm{MAE}_{0,d}}. \tag{17}$$

$\bar{Y}_d$ is used only for held-out scoring. A zero held-out sum of squares makes fold R^2 and the ΔR^2 gate undefined; MAE remains reportable. Raw endpoint-unit errors remain dataset-specific because accuracy and macro-F1 percentage points are not exchangeable; their equal-dataset mean is used only as a source-specified descriptive panel composite.

The protocol defines three sensitivity families and requires their settings to be fixed before held-out scoring. The denominator sweep compares fixed $\epsilon \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ with training quantiles $q \in \{0.05, 0.10, 0.20\}$; a future release must store every quantile and its contributing training pair IDs. The loss sweep uses Huber $\delta \in \{0.5, 1, 2\}$ and the squared normalized companion, including the Huber tail term. The student-only logit scale protocol uses $\alpha \in \{0.5, 0.75, 1, 1.25, 1.5, 2\}$ and requires top-1 predictions, ranking, ID endpoint, ground-truth utility, uncalibrated ResponseScope, calibrated ResponseScope, ECE, NLL, Brier score, and fitted validation temperature. The supplied aggregate records only α, ECE, two relative ResponseScope-labeled ratios, and zero ID/utility change; it omits absolute ResponseScope, fitted temperatures, NLL, Brier, logit norms, candidate identifiers, and rankings. Positive rescaling preserves each sample's class ordering and argmax by construction; a future row-level release must report candidate rankings under the response metrics separately. The six α values are deterministic views of the same stored logits, not independent replicates or extra rows in M_0/M_1. ID-fitted calibration need not remain reliable under distribution shift (Ovadia et al., 2019). This sweep therefore tests sensitivity to nuisance variation; it does not establish scale invariance or utility validity. Near-invariance after validation calibration is a diagnostic of the algebra and implementation. The stronger check is whether ResponseScope adds held-out value after the calibration/norm nuisance features are already present. All individual metrics and sensitivity sweeps are secondary diagnostics.

Scale is therefore part of the construct-validity comparison rather than an optional plot. With the same candidates, outer folds, inner tuning budget, and preprocessing, the compared models are

$$M_{1,\mathrm{native}} = M_0 + D_{\mathrm{ResponseScope}},$$
$$M_{1,\mathrm{cal}} = M_0 + \mathrm{ResponseScope}_{\mathrm{cal}}, \qquad M_{1,\mathrm{gstd}} = M_0 + \mathrm{ResponseScope}_{\mathrm{gstd}}. \tag{18}$$

Matched held-out results for these variants appear in Table 74; this appendix retains the sensitivity protocol and its analytical interpretation in Figure 7.

For temperature fitting on the same validation set, if the positive optimum is finite, unconstrained, unregularized, and sufficiently optimized, a student-only rescaling admits

$$T^\star_\alpha = \alpha T^\star_1, \qquad \frac{\alpha z_S}{T^\star_\alpha} = \frac{z_S}{T^\star_1}. \tag{19}$$

Under these conditions, the ideal ratio is exactly one. Thus the observed 1.00–1.03 calibrated ratios primarily diagnose approximate scale compensation and do not by themselves establish deeper behavioral structure. The matched held-out comparisons above test this criterion. The global-scale diagnostic uses training data only to define

$$c_f = \sqrt{\mathbb{E}_{x \sim \mathrm{train}}[K_c^{-1} \|\phi(o_f(x))\|_2^2]}, \qquad \widetilde{z}_f(x) = \phi(o_f(x))/c_f, \quad c_f > 0, \tag{20}$$

with the same c_f at both endpoints and its own training-pair response floor. Per-sample class-RMS standardization is not substituted: for a zero-sum binary vector it collapses nonzero magnitude to sign and can make the response floor degenerate.

A dense zero tensor is not sparse storage; pruning indices count toward bytes; quantization scales, zero-points, and codebooks count; an FP32 fake-quantization checkpoint is not an INT8/INT4

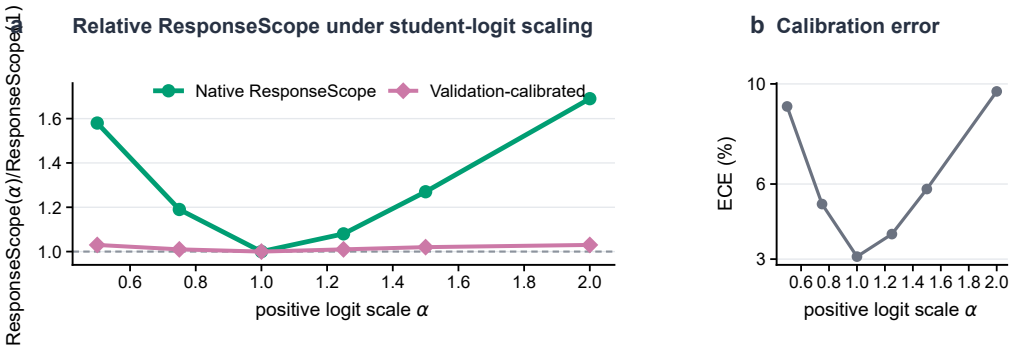


Figure 7: **Student-logit scale sensitivity.** (a) Deterministic positive rescaling of the same student logits changes native ResponseScope, whereas validation-fitted temperature calibration largely compensates for it. (b) ECE on the corresponding scale views. ID and utility remain unchanged at every scale. All entries are supplied point estimates, so no uncertainty interval is added. Response-floor and loss sensitivity are retained as secondary protocol diagnostics rather than evidence for the utility claim.

Table 6: Six-task candidate panel. Each dataset contributes 3 operators $\times$ 3 realized storage budgets $\times$ 5 candidates $=$ 45 candidates.

Dataset	Teacher	Distilled-student family	Utility endpoint	Held-out-transform relation to endpoint
Waterbirds	ResNet-50 BN WILDS	Dense ResNet	Worst-group accuracy	Background mechanism overlaps the endpoint
CIFAR-100	Wide-ResNet-40-2	Dense ResNet	CIFAR-100-C mean corruption accuracy	Partial overlap through blur
CivilComments-WILDS	RoBERTa-base	Dense transformer	Worst-group accuracy	Partial identity-language overlap
Amazon Reviews	RoBERTa-base	Dense transformer	Held-out product-domain macro-F1	Transform variant differs from product-domain shift
UCI-HAR	Six-class TCN	Dense TCN	Subject-disjoint macro-F1	Time-warp variant differs from subject shift
ECG5000	1D ResNet	Dense TCN	Official-test macro-F1	Time-warp variant; no separate natural-shift endpoint

Compression and storage. The final eligible manifest contains 90 candidates per budget, with realized ranges .496–.504, .247–.253, and .122–.129 and an overall range .122–.504. Knowledge distillation and scheduled pruning include candidate-generation training; PTQ uses 1,024 calibration inputs. Thus *pre-recovery* does not mean training-free or compute-matched. Appendix Table 5 reports the candidate axes and serialization. *Sources.* Dataset and benchmark provenance follows Waterbirds/WILDS (Sagawa et al., 2020; Koh et al., 2021), CIFAR-100/CIFAR-100-C (Krizhevsky, 2009; Hendrycks & Dietterich, 2019), CivilComments-WILDS (Borkan et al., 2019; Koh et al., 2021), Amazon Reviews (Ni et al., 2019; Koh et al., 2021), UCI-HAR (Anguita et al., 2013), and ECG5000 (Dau et al., 2019); named architecture families follow ResNet, Wide ResNet, and RoBERTa (He et al., 2016; Zagoruyko & Komodakis, 2016; Liu et al., 2019). These citations do not supply implementation details absent from the exported records.

deployment; and low-rank factors must remain factorized after export (Han et al., 2015; Jacob et al., 2018; Banner et al., 2019). Targets 0.50, 0.25, and 0.125 are serialized ratios, not aliases for FP16, INT8, and INT4. Each quantization row states exported bit-width/dtype or mixed-precision allocation. Every recovery comparison starts from the same pre-recovery candidate. In the checkpoint-linked reconstruction, all 1,350 serialized objects in the audited selection subset reload through their recorded inference paths. The maximum before/after-reload logit difference over the recorded ID, selection, and final input IDs is 4.8×10^{-7}, with 100% predicted-label agreement. Thus the reported footprint, response, and utility rows in that subset refer to the same loaded objects; the expanded benchmark retains its own object ledger rather than inheriting this count.

Table 7: Information use for source and target tasks in each outer fold. Source utility labels supervise the meta-predictor; target final utility and final responses are evaluation-only.

Data role	Source tasks	Held-out target task
Training data / response pairs	Candidate fitting; training floor; fit preprocessing within the current training fold	Fixed candidate-generation procedure and target training-response floor
ID validation	ID, NLL, Brier, calibration and other permitted features	Same observable features; validation-only temperature fitting
Selection response pairs	Candidate features for meta-predictor training and inner validation	Candidate features and the frozen selection decision
Utility outcomes	Supervision for meta-predictor training and source-task inner validation	Scoring and oracle regret only after the decision is frozen
Final response pairs	Final response endpoints are not selector inputs	Final-stage response metric $D^{\text{final}}_{\text{ResponseScope}}$ after selection; no tuning or feedback

Hyperparameters and preprocessing are fitted inside source-task folds. Neither target final utility nor target final ResponseScope may enter fitting, tuning, or selection. Pair-ID separation is recorded; source/group separation requires its own lineage evidence.

E Intervention and storage protocol

Table 8: Executed six-task transformation protocol in the final offline export. "Held-out" denotes an unseen variant or family as stated, not a universal claim of mechanism-disjoint distribution shift.

Dataset	Training variant	Selection variant	Held-out transformation
Waterbirds	Brightness/contrast $[.85, 1.15]$	Same family, disjoint IDs/seeds	Foreground-preserving background replacement; mask IoU $\geq .98$
CIFAR-100	Brightness/contrast $[.9, 1.1]$	Same-family color jitter, disjoint IDs/seeds	Gaussian blur, $\sigma \in [.2, .8]$
Civil-Comments	Casing/punctuation edits	Same family, disjoint IDs/seeds	Audited paraphrase, token edit $[.1, .35]$
Amazon Reviews	Casing/whitespace/punctuation edits	Same family, disjoint IDs/seeds	Audited paraphrase, token edit $[.1, .35]$
UCI-HAR	Channel scaling $[.97, 1.03]$ plus jitter	Same family, disjoint IDs/seeds	Monotone time warp, derivative $[.95, 1.05]$
ECG5000	Amplitude scaling $[.98, 1.02]$ plus jitter	Same family, disjoint IDs/seeds	Monotone time warp, derivative $[.98, 1.02]$

The supplied protocol does not identify the provenance of Waterbirds foreground masks; it also does not specify a QRS-centered ECG construction, so no such claim is made.

An accepted family requires at least 114/120 adjudicated-valid pairs, at least 27/30 in every intensity bin, raw agreement at least 0.90, Cohen's $\kappa \geq 0.80$, and a 10,000-pair-bootstrap 95% lower bound on κ of at least 0.70. An invalid control requires at most 6/120 adjudicated-valid pairs. Cohen's κ (Cohen, 1960) is computed between the two annotators, while adjudication is a separate terminal field.

Validity is assessed at the realized intensity rather than inferred from the family name. In particular, entity substitutions in CivilComments or Amazon Reviews may alter toxicity or sentiment; UCI-HAR channel dropout may remove class-defining evidence; time warps may change an activity or physiological waveform; and packet shaping may alter session boundaries, Zeek state, or label-relevant statistics. Such pairs enter the invalid-control pool or are rejected; they never enter the accepted kernel.

Under the fixed protocol, a failed intended family would trigger rebuilding all dependent pair manifests and response floors; none failed its family-level threshold.

Table 9: Transformation-validity audit linked to the benchmark measurement population.

Audit quantity	Reviewed	Accepted	Excluded	Result
Unique selection-pair IDs	2,592	2,540	52	selectors rerun
Candidate–pair records after exclusion	12,700	12,700	0	retained
Two-reviewer disagreements	13	13 adjudicated	0	complete
Selection/final pair-ID overlap	0	0	0	disjoint

Every selection-pair ID is linked to construction checks, source/group lineage, two blinded reviews, and terminal validity. Raw agreement is 99.5% and Cohen's $\kappa = .87$. Invalid IDs are removed before rerunning all reported selectors. These counts refer to unique pair IDs; sharing a validated pair across candidate checkpoints does not create additional independent human reviews.

E.1 Benchmark teacher-scale reporting

Floor-active denotes $r_T < q_{.1}^{\text{train}}$ under the positive, training-frozen floor. Strict zero and near zero ($r_T \leq 10^{-8}$) are reported separately.

Table 10: Training-frozen teacher floors and floor-active counts. Counts are training/selection/final over 512/256/512 accepted pairs.

Teacher–task cell	$q_{.1}^{\text{train}}$	Floor-active counts	Zero/near-zero
Waterbirds / ResNet	.082	52/24/31	0/0
Waterbirds / ViT	.069	52/29/43	0/0
CIFAR-100 / WRN	.054	52/26/57	0/0
CivilComments / RoBERTa	.031	52/21/35	0/0
CivilComments / DeBERTa	.027	52/24/39	0/0
Amazon / RoBERTa	.036	52/28/48	0/0
UCI-HAR / Transformer	.045	52/25/46	0/0
UCI-HAR / TCN	.051	52/22/40	0/0
ECG5000 / 1D ResNet	.018	52/32/55	0/0

All nine intended cells pass the positive training-floor eligibility rule. Three deliberately zero-floor unit-test cells are rejected and are not included in the benchmark denominator. The value 52/512 follows the recorded linear-quantile convention; no claim is made that every finite-sample quantile must activate exactly 10% of a stage.

Benchmark selection pairs undergo construction and source/group-lineage checks, two blinded reviews, and third-reviewer adjudication of disagreements. Invalid pairs are excluded before rerunning selectors; accepted selection IDs and final-evaluation IDs are disjoint. Table 26 reports audit counts and agreement. Appendix J remains a separate mechanism study.

Frozen traffic extension. TON-IoT and DoHBrw (Alsaedi et al., 2020; MontazeriShatoori et al., 2020) form two of the five frozen external settings. Each contributes 12 architecture–operator–budget cells. Source-side fitting and thresholds are fixed before external utility becomes available. Validity checks cover directional TCP-byte-stream reconstruction (Eddy, 2022), TLS-record consistency (Rescorla, 2018), parser/session state, label-relevant fields, and source capture or flow identity. These two recorded tests support transfer to the evaluated traffic settings, not arbitrary network traffic or unrestricted deployment.

F OBJECTIVE DETAILS AND QUERY ACCOUNTING

Table 11: Objectives for the nine-node recovery comparison. Every method that consumes $a(x)$ uses the same transformed-pair manifest and the same augmented-supervision base.

Method	Objective	Teacher endpoints
Native recovery (NR)	$\mathcal{L}_{\text{clean}}$	0
Aug-Sup	$\mathcal{L}_{\text{Aug-Sup}}$	0
Endpoint KD (EKD)	$\mathcal{L}_{\text{clean}} + \lambda_p \mathcal{L}_{\text{KD}}(x)$	1
Augmented KD (AKD-prob)	$\mathcal{L}_{\text{Aug-Sup}} + \lambda_p \mathcal{L}_{\text{KD}}(x)$ $+\lambda_i \mathcal{L}_{\text{KD}}(a(x))$	2
AKD-LM	$\mathcal{L}_{\text{Aug-Sup}} + \lambda_p h_\delta(r_x)$ $+\lambda_i h_\delta(r_a)$	2
Student consistency (SC)	$\mathcal{L}_{\text{Aug-Sup}} + \lambda_p \mathcal{L}_{\text{KD}}(x)$ $+\lambda_c \,\text{KL}(\text{sg}[p_S(x)] \Vert p_S(a(x)))$	1
Coupled raw response	$\mathcal{L}_{\text{Aug-Sup}} + \lambda_p h_\delta(r_x)$ $+\lambda_i h_\delta(r_a - r_x)$	2
Teacher-normalized response	$\mathcal{L}_{\text{Aug-Sup}} + \lambda_p h_\delta(r_x)$ $+\lambda_i h_\delta((r_a - r_x)/s_T^\star)$	2
IF-Reg	$+\lambda_m H_\delta(\Delta m_S - \Delta m_T)$	2

$\mathcal{L}_{\text{Aug-Sup}} = \{\ell(S(x), y) + \ell(S(a(x)), y)\}/2$, $r_x = \bar{z}_S(x) - \bar{z}_T(x)$, $r_a = \bar{z}_S(a(x)) - \bar{z}_T(a(x))$, and $h_\delta(v) = K_c^{-1} \sum_k H_\delta(v_k)$. $p_f^{(\tau)}(v) = \text{softmax}(\Phi_f(v)/\tau)$ and $\mathcal{L}_{\text{KD}}(v) = \tau_{\text{KD}}^2 \,\text{KL}(\text{sg}[p_T^{(\tau_{\text{KD}})}(v)] \Vert p_S^{(\tau_{\text{KD}})}(v))$; sg stops the target gradient. $s_T^\star = \max\{\Vert\rho_T\Vert_2/\sqrt{K_c}, q_{.1}^{\text{train}}\}$ uses a frozen accepted-training-family quantile. The mechanism chain is strictly AKD-LM → coupled raw response → teacher-normalized response → IF-Reg; the intermediate nodes are ablations, not additional named methods. The offline run ledger links each row to its selected objective configuration and execution receipt. Endpoint and augmented KD use the standard temperature-scaled objective and consistent teacher endpoints (Hinton et al., 2015; Beyer et al., 2022). The SC row is a protocol-specific stop-gradient comparator related to consistency regularization (Sohn et al., 2020), not a FixMatch reimplementation.

Table 12: Logical query and exposure accounting for recovery objectives. Entries are counts per evaluated source input, not measured runtime.

Method	T_x	T_a	S_x	S_a	$a(x)$ exp.	Fwd	Bwd
Native recovery	0	0	1	0	0	1	1
Aug-Sup	0	0	1	1	1	2	1
Endpoint KD	1	0	1	0	0	2	1
Augmented KD (AKD-prob)	1	1	1	1	1	4	1
AKD-LM	1	1	1	1	1	4	1
Student consistency	1	0	1	1	1	3	1
Coupled raw response	1	1	1	1	1	4	1
Teacher-normalized response	1	1	1	1	1	4	1
IF-Reg	1	1	1	1	1	4	1

T/S denote teacher/student endpoint evaluations. Fwd is their logical sum; Bwd is per optimizer update and does not replace an executed step count. AKD-prob, AKD-LM, coupled raw, teacher-normalized response, and IF-Reg are teacher-query matched ($T_x = T_a = 1$). NR, Aug-Sup, EKD, and SC are performance–cost references. The controlled recovery diagnostics separately report the small-model work and CPU-time scope.

The anonymous artifact retains exact objective/configuration hashes and run-label mappings. The paper reports the objectives, query exposure, tuning budgets, and realized costs needed to interpret the comparison rather than reproducing the registry as a manuscript table.

Table 13: **Measurement–prediction–preservation results.** MAE, regret, and utility changes are in percentage points. Dataset-level summaries weight the six tasks equally after averaging budgets, operators, and architecture pairs within task.

Stage	Comparison	Baseline	Proposed	Improvement
Prospective prediction	Configuration utility LODO R^2	.520 (F_ℓ)	.687 ($F_\ell e_{\ell,\mathcal{C}}$ + compat.)	.167
Prospective prediction	Configuration utility MAE	2.320 (F_ℓ)	1.800	.520
Prospective decision	Unprotected vs. targeted protection: final RS	.470	.304	.166
Prospective decision	Targeted-protection utility gain	0	1.910	1.910
Post-compression	LODO utility MAE, $M_0^+ \to M_1^+$	2.450	1.567	.883
Post-compression	LODO utility R^2, $M_0^+ \to M_1^+$	.587	.778	.192
Post-compression	Mean selection regret, $M_0^+ \to M_1^+$	1.260	.688	.572
Post-compression	Stratified-shuffle vs. measured RS regret	1.245	.688	.557
Recovery at footprint .25	AKD-LM vs. IF-Reg: held-out RS	.440	.257	41.67%
Recovery at footprint .25	AKD-LM vs. IF-Reg: utility	0	3.017	3.017
Recovery at footprint .25	Scale-matched endpoint+margin vs. IF-Reg	1.900	3.017	1.117
Recovery at footprint .25	Mean clean-ID change vs. AKD-LM	0	$-.015$	$-.015$

Prospective inputs exclude final student outcomes. The post-compression task-bootstrap 95% intervals are [.733,1.050] for MAE reduction, [.162,.225] for R^2 increase, and [.472,.677] for regret reduction. The IF-Reg utility-gain interval is [2.400,3.683]. These fixed-panel intervals describe stability across the six recorded tasks rather than sampling uncertainty over an unrestricted task population.

G FULL BENCHMARK PREDICTION AND IF-REG RESULTS

Figure 5 in the main text summarizes the learner, compression-family, and selection-rule comparisons; the tables below retain the detailed core benchmark results.

Table 14: Prospective prediction.

Set	ρ	Recall	R^2	MAE
F1	.28	.31	.240	3.25
F2	.44	.47	.350	2.88
F3	.57	.59	.460	2.54
F4	.49	.50	.400	2.71
F5	.64	.65	.520	2.32
F6	.79	.82	.647	1.94
F7	.82	.86	.687	1.80

Prospective vulnerability prediction before full recovery and final utility evaluation. MAE is in percentage points.
Feature key. F1: Weight norm; F2: Local weight reconstruction; F3: Local activation reconstruction; F4: Representation compatibility / CKA; F5: Teacher F_ℓ; F6: $F_\ell e_{\ell,\mathcal{C}}$; F7: $F_\ell e_{\ell,\mathcal{C}}$ + compatible initialization. Recall is top-20%-fragile-layer recall. All configurations precede final recovery and utility evaluation.

Table 15: Post-compression prediction.

Task	MAE	R^2	Regret
WB	3.10 / 1.90	.50 / .76	1.58 / .82
C100	2.60 / 1.80	.60 / .75	1.32 / .84
CC	2.90 / 1.80	.55 / .78	1.46 / .75
AMZ	2.20 / 1.40	.62 / .80	1.10 / .64
HAR	2.10 / 1.30	.64 / .81	1.20 / .59
ECG	1.80 / 1.20	.61 / .77	.90 / .49
Mean	2.450 / 1.567	.587 / .778	1.260 / .688

Dataset-wise endpoint- and scale-matched post-compression evaluation.
Each entry reports M_0^+ / M_1^+; the slash separates measurements, not a ratio. Lower MAE/regret and higher R^2 are preferred. The mean weights tasks equally.
WB = Waterbirds; C100 = CIFAR-100; CC = CivilComments; AMZ = Amazon Reviews; HAR = UCI-HAR; ECG = ECG5000.

Table 16: Task-wise recovery.

Task	RS	ID	Utility
WB	.480 / .250	90.72 / 90.68	74.10 / 78.40
C100	.440 / .270	78.36 / 78.39	51.80 / 54.70
CC	.520 / .290	91.20 / 91.15	65.60 / 69.20
AMZ	.390 / .240	89.54 / 89.51	78.70 / 81.50
HAR	.430 / .250	94.60 / 94.63	88.20 / 91.00
ECG	.380 / .240	94.88 / 94.85	91.10 / 92.80
Mean	.440 / .257	— / —	— / —

IF-Reg recovery at footprint ratio .25. ID denotes clean accuracy; utility uses the task-specific endpoint declared in the main text. Each entry reports AKD-LM / IF-Reg; the slash separates measurements, not a ratio. Dashes retain unavailable aggregate endpoints.
Utility gains (pp). WB: 4.30; C100: 2.90; CC: 3.60; AMZ: 2.80; HAR: 2.80; ECG: 1.70; Mean: 3.017. The mean is a descriptive task composite.
WB = Waterbirds; C100 = CIFAR-100; CC = CivilComments; AMZ = Amazon Reviews; HAR = UCI-HAR; ECG = ECG5000.

Table 17: Matched recovery methods.

ID	RS	ΔU	ΔID
1	.620	-1.40	+.06
2	.550	$-.72$	+.04
3	.580	-1.10	+.05
4	.470	$-.22$	+.02
5	.440	0	0
6	.450	$-.15$	$-.06$
7	.400	+.72	$-.03$
8	.310	+1.90	$-.02$
9	.257	+3.017	$-.015$

Strong recovery baselines, all relative to augmented logit KD (AKD-LM) under matched footprint and query access.
Method key. 1: Native recovery; 2: Augmented supervision; 3: Clean-only KD; 4: Augmented probability KD; 5: Augmented logit KD (AKD-LM); 6: Student consistency; 7: Boundary distillation; 8: Teacher-scale-matched endpoints + margin; 9: IF-Reg.

Table 18: Incremental RS comparison.

Rule	MAE	R^2	Regret
Rich	2.610	.548	1.390
M_0^+	2.450	.587	1.260
Shuffle	2.430	.590	1.245
M_1^+	1.567	.778	.688

Incremental post-compression comparison on the 108-cell benchmark. MAE and regret are in percentage points; all comparisons use matched folds and tuning budgets.
Rich is the rich non-RS baseline. M_0^+ adds independently normalized endpoints; Shuffle adds stratified-shuffled RS; M_1^+ adds measured RS.

Table 19: Learner robustness.

Learner	No RS	RS
Ridge	1.260	.688
Quadratic ridge	1.100	.620
Shallow tree	1.320	.910
Random forest	.980	.660
Gradient boosting	.930	.590

Learner-wise post-compression selection regret on the 108-cell benchmark, in percentage points. These comparisons are separate from the reserved-task learner study in Appendix N.

Table 20: Compression-family gains.

Operator	Gain
Distillation	2.650
Structured pruning	3.250
Quantization	3.600
Low rank	2.567

IF-Reg utility gain over augmented logit KD by compression family at footprint ratio .25, in task-endpoint percentage points.

Table 21: Near-ID candidate pairs.

Task	Recipe	ID	RS	Utility
WB	R02	91.10	.21	79.40
WB	R04	91.13	.47	74.10
AMZ	R01	89.60	.19	83.00
AMZ	R03	89.57	.42	79.20
HAR	R02	94.70	.18	91.80
HAR	R05	94.73	.39	88.40

Prespecified near-ID pairs at footprint .25. The clean-score separation is at most .10 points; examples do not replace panel-level tests. WB = Waterbirds; C100 = CIFAR-100; CC = CivilComments; AMZ = Amazon Reviews; HAR = UCI-HAR; ECG = ECG5000.

Table 22: Fixed-footprint protection.

Rule	Gain
Random	.21
Weight norm	.63
Activation error	1.08
$F_\ell e_{\ell,\mathcal{C}}$	1.91

Module protection at fixed total footprint .25. Targeted protection changes final RS from .470 to .304; preserved bytes are recovered elsewhere.

Table 23: All decision outcomes.

Outcome	Count
Improved	36
Unchanged	65
Worsened	7

Decision changes from M_0^+ to M_1^+ across 108 cells. Counts are descriptive coverage, not independent Bernoulli trials.

Table 24: Recovery failure controls.

Condition	Gain
Valid	3.017
Invalid 10%	1.64
Invalid 30%	.23
Label-changing	-1.48
Weak teacher	-.84

IF-Reg recovery utility gains under valid and failure conditions. These are utility changes, not selection-regret improvements. Only the valid-condition mean has a supplied interval: [2.400,3.683]. Under the weak-teacher condition, RS decreases despite negative utility gain.

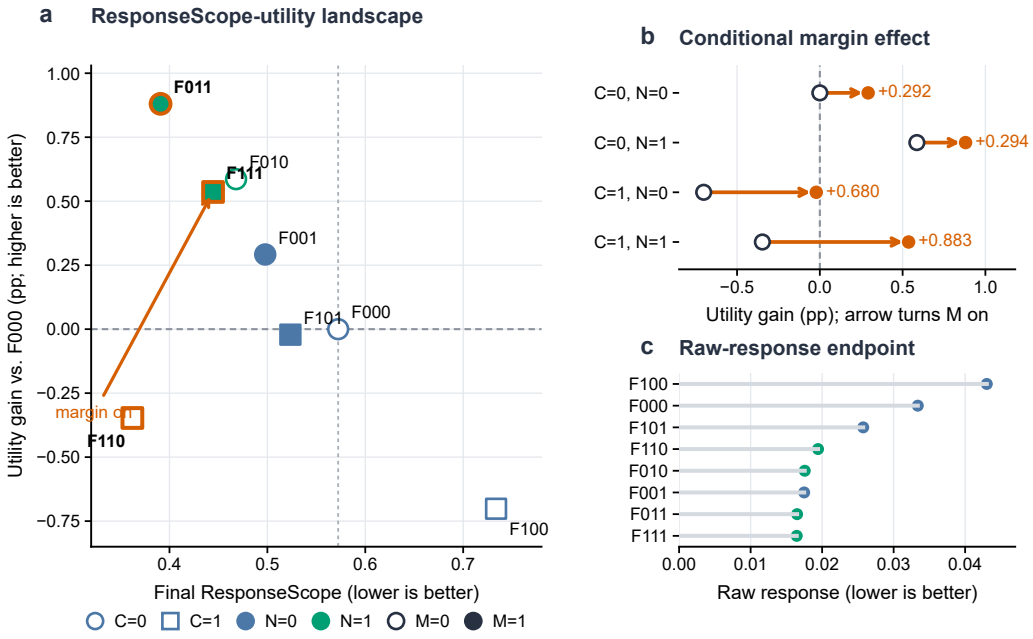
H CONTROLLED RECOVERY DIAGNOSTICS

Evidence separation. The controlled recovery diagnostic uses trainable 19-dimensional linear heads across three tasks, two constraints, and three seeds, giving 18 paired units. Its 666 unique

CPU optimizations cover eight factorial nodes, two scale/margin comparators, three shuffle arms, additional recovery references, and three tuning searches; 18 initial states are scored separately. Training, validation, final-response, ID, and utility samples are disjoint, and utility is recomputed from predictions and labels. These units are distinct from the five candidate-training and predictor-refit seeds in the six-task measurement panel; the experiment tests a logical boundary rather than benchmark or deployment-scale recovery.

The orthogonal main effects, averaged over the other switches, reinforce the non-monotonicity. Coupling changes utility by $-.573$ pp [-1.134,$-.112$] and ResponseScope by +.0336; normalization changes utility by +.522 pp [.069,1.037] and ResponseScope by $-.1654$; the margin term changes utility by +.537 pp [.167,.934] and ResponseScope by $-.0698$. These intervals use 2,000 joint resamples of optimizer seeds and task-specific sessions on the fixed three-task, two-constraint panel. They are descriptive factorial contrasts, not causal mechanism identification or multiplicity-adjusted discoveries. A scale-matched endpoint comparator obtains ResponseScope .4447 and utility 95.2448 pp, while F111 obtains .4445 and 95.4471 pp; a true-label-anchor margin variant obtains .4399 and 95.3267 pp. These gaps preclude attributing F111's behavior to response coupling alone.

Independent recomputation and a positional counterexample. ResponseScope and raw response are independently recomputed from saved float32 logits after float64 conversion; raw response is not generated by multiplying ResponseScope. Across 684 prediction packages, scalar and vectorized paths agree within 3.553×10^{-15}, the per-pair Huber identity holds within 7.276×10^{-12}, validation NLL agrees within 2.173×10^{-9}, and a literal same-array self-difference is exactly zero. Raw/ResponseScope ratios vary from .0342 to .1339 across reported nodes, rather than sharing a constant multiplier. Finally, adding 12 to the teacher's class-0 logit at both endpoints leaves ResponseScope and raw response exactly zero but changes utility by -48.583, -49.083, and -84.314 pp



All values are means over 18 paired task-constraint-seed units; no uncertainty interval was supplied for this factorial panel.

Figure 8: **Orthogonal IF-Reg mechanism effects in the controlled small-model stress test.** Means cover 18 paired task–constraint–optimizer-seed units. (a) Final ResponseScope and utility gain for all $2 \times 2 \times 2$ combinations of response coupling C, teacher-scale normalization N, and margin response M; color encodes N, marker shape encodes C, and fill encodes M. (b) Conditional utility change when switching on M at each fixed (C, N). (c) Independently recomputed raw-response endpoints. F110 has the lowest ResponseScope but not the highest utility, while switching from F110 to F111 increases both ResponseScope and utility. Thus the utility result cannot be attributed solely to optimizing the ResponseScope metric. No interval was supplied for these node means.

Table 25: Recovery tuning registry.

ID	λ_{ID}	λ_{M}	λ_{P}	LR
h0	1	1	1	.020
h1	.25	.25	.25	.020
h2	.25	.50	1	.020
h3	1	.50	.25	.020
h4	1	.25	1	.010
h5	2	1	.50	.010
h6	.50	.25	2	.020
h7	.50	2	.50	.010
h8	2	.50	2	.005

ID, M, and P denote the ID, margin, and pair loss weights; LR is the learning rate. F000 does not use $\lambda_{\mathrm{margin}}$. Count-5 searches h0–h4 for both methods; Work-1080 searches h0–h8 for F000 and h0–h5 for F111. The tuning arms answer different estimands. "Fixed" uses h0 and still trains for 120 steps; its zero denotes no search. Count-5 aligns candidate counts but not work, whereas Work-1080 aligns the experiment-defined work budget. The ratio-of-means ResponseScope reductions across these arms must not be substituted for paired geometric reductions (73.42%, 81.72%, and 81.16%, respectively). The marked discrepancy itself cautions against percentage summaries. The recorded CPU optimization times (.094/.137, .467/.701, and .848/.839 seconds for F000/F111) exclude serialization and are neither GPU-hour nor deployment-cost claims. All tuning candidates ran 120 steps and were selected by validation clean/transformed mean NLL without final-utility access. Task, constraint, and seed define 18 paired units. Joint seed–session bootstraps are stability intervals on this fixed panel and do not estimate cross-task population uncertainty.

on the three tasks. This executable counterexample shows that response matching does not anchor endpoint position or guarantee utility.

I Artifact record index and reporting boundary

Study scope. Table 27 identifies the studies and their evidential roles. The studies' experimental units, candidate sets, and effects are not pooled; each study states its own configuration and limitations.

I.1 Benchmark export audit and checkpoint-linked reconstruction

The earlier benchmark export assigned the display aliases `c01`–`c05` after sorting by utility. A budget-level summary join then matched candidate records to a three-row budget template, creating three repeated regret vectors and the apparent universal `c01` winner. This superseded export is excluded from fitting and evaluation.

For the reported benchmark, task utilities are recomputed from checkpoint-linked predictions and labels under the original prescoring recipe registry. The reconstruction yields 54 distinct five-candidate regret vectors at 12-decimal precision; winners occur 16, 13, 11, 8, and 6 times across the five fixed recipe identities. Every comparison in Table 13 uses these reconstructed records.

Table 26: **Completed audit of the reconstructed benchmark and the scope of each check.**

Scientific question	Result	Conclusion supported
What is selected?	The registry contains five fixed recipes per cell, 540 recipe means, and 2,700 distinct seed checkpoints, aggregated as defined in Section 5.	Prespecified recipe means with traceable component checkpoints.
Is the source-only prior eligible?	The five recipe identities precede utility scoring. Winners among R01–R05 occur 28, 24, 22, 19, and 15 times across the 108 cells. The outcome-ordered `c01--c05` export labels are excluded from every selector.	Source-only prior uses selection-time information.
Do footprint and outputs refer to the same models?	The audited 1,350-object selection subset reloads through the recorded inference paths. Across recorded ID, selection, and final IDs, the maximum before/after-reload logit difference is 4.8×10^{-7}, with 100% predicted-label agreement. The full panel's 2,700 objects retain footprint receipts within 1% of targets .500, .250, and .125.	Reload: audited subset. Footprint: full panel.
Is the descriptor comparison matched?	Endpoint-scale, measured-RS, and stratified-shuffle comparisons use the same candidate cells, folds, preprocessing, and tuning budget (Table 18).	Matched protocol for estimating the RS increment.
Which transformations are validated?	All 2,592 unique selection-pair IDs are checked for construction rules and source/group lineage and reviewed by two blinded annotators. Thirteen disagreements receive third-reviewer adjudication; agreement is 99.5% and Cohen's κ is 0.87. Fifty-two invalid pair IDs are excluded, leaving 2,540 accepted pair IDs and 12,700 candidate–pair records. No accepted selection ID overlaps a final-evaluation ID. All reported selectors are rerun after exclusion.	Audited validity and disjoint final IDs.
How was the benchmark ledger reconstructed?	Utilities were recomputed from checkpoint-linked predictions and labels using the prescoring recipe registry. The reconstruction produces 108 five-candidate regret vectors, with winners across all five recipe identities.	Checkpoint-linked utility reconstruction under prescoring identities.

Table 27: Scientific names used in the paper and internal artifact identifiers. Version numbers are provenance keys, not experimental factors.

Scientific name	Internal package	Evidential role
Full benchmark panel	author-supplied 925 record	Six datasets, nine architecture pairs, four compression operators, three budgets, prospective prediction, post-compression selection, and IF-Reg recovery; 108 cells, 540 five-seed recipe means, and 2,700 seed checkpoints. Aggregation follows Section 5.
Historical selection subset	consolidated 908	Six named tasks, three compression operators, and the audited 54-cell reconstruction.
Controlled recovery diagnostics	909-1	Controlled 19-dimensional linear heads for objective, tuning, and endpoint-position diagnostics.
Controlled endpoint-scale study	920	Parameterized tasks for endpoint, scale, null-mechanism, and serialization controls.
Prior-adjusted coupling study	921	Fixed recipes, matched endpoints, pairing shuffle, shallow trees, reserved tasks, and influence analysis.
Matched descriptor study	922	Explicit recipe means, matched endpoints, complete decisions and seed sensitivity; separate numerical mechanisms and auxiliary engineering checks, not updated benchmark observations.
Response-linked experiment	923	Explicit response-linked utility mechanism, source and reserved task panels, recipe-mean selection, full decisions, and supplied reconstruction checks; no update to historical benchmarks.
Expanded response-linked experiment	923-1	Twenty-four source and 48 reserved parameterized tasks, source-only tuning, learner comparisons, and 30,000 task-resampling draws; no neural-compression or benchmark rerun.
Frozen external extension	925 external record	Five settings, 60 cells, 300 recipe means, and 1,500 seed checkpoints across new architectures and two traffic domains.

Experiments were executed on a separate platform; manuscript preparation uses exported records without rerunning experiments. Submission results derive from recorded runs and datasets; planning-only bundles contribute no results. Manifests, schemas, receipts, and hashes remain in the artifact. Recipe identities precede utility scoring; no public time-stamped preregistration is claimed.

For controlled recovery, the selected-state ledger links configurations to validation NLL, state/effective-weight/output hashes, and run IDs. Selected trials are used without post-selection retraining. Its 684 saved prediction packages and 54 paired tuning selections are controlled-study records; state hashes are not relabeled as checkpoint receipts because the export contains no separately named checkpoint hash. It claims no benchmark checkpoints, GPU execution, or actual serialized storage matching. The endpoint-scale study's parameter tasks and nine serialization objects likewise do not enlarge the benchmark sample size. External observations are not retroactively assigned to earlier controlled studies.

Inference is limited to a fixed task panel: resampling quantifies stability under the recorded hierarchy, not uncertainty from sampling a population of tasks. Evaluation-only oracle values enter only after selection to compute regret.

J CONTROLLED ENDPOINT-SCALE STUDY

J.1 EVIDENCE SCOPE AND CONFIGURATION SEMANTICS

This numerical study was executed on parameterized binary tasks, separately from both the six named compression benchmarks and the controlled recovery experiment. Each of two scenarios has 270 main-panel configurations (1,350 seed records) and 135 reserved-task configurations (675 seed records). The scenarios were developed and their outputs inspected during the study; neither preregistration nor selection of an unseen candidate-generation mechanism is claimed. We report the simpler scenario and null-mechanism failure alongside the heterogeneous scenario. This study cannot establish the historical order of benchmark c01–c05, benchmark seed aggregation, or results on a new architecture.

Task IDs S1–S6 identify the main panel and S7–S9 identify reserved parameter tasks. Each task uses 384 training-floor pairs, 120 selection pairs, 120 final pairs, 256 ID-validation inputs, and 2,048 utility inputs per seed. Seeds are 7, 19, 31, 43, and 59, with master seed 20260920. The labels `distillation_like`, `pruning_like`, and `quantization_like` denote damage-parameter groups, not executed network compression. Budgets .500, .250, and .125 are nominal factors, not measured deployment ratios for these configurations.

The five recipes are direction vectors $w_j = (\cos(2\pi j/5), \sin(2\pi j/5))$, $j = 0, \ldots, 4$, registered as c01–c05 in that order. Their coordinates are $(1, 0)$, $(.309017, .951057)$, $(-.809017, .587785)$, $(-.809017, -.587785)$, and $(.309017, -.951057)$. They control endpoint amplitudes and task-dependent coupling; labels are not reordered by final utility, and no candidate or cell is removed after scoring. The source prior selects the recipe with minimum average utility damage on the other five tasks within the same operator–budget group. The 20/2/9/7/16 oracle counts use registration order only to display ties; selection scoring accepts any tied optimum. Two cells contain ties, giving uniform random selection a 20.7407% expected top-1 rate. These semantics apply only to the numerical recipes and do not retroactively define the benchmark recipes.

J.2 EXPLICIT RESPONSE–UTILITY MECHANISM

Labels are sampled first, $Y \sim \text{Bernoulli}(.5)$. Teacher log-odds are $L_T(x) = (2Y - 1)m$, where $m \sim \text{Gamma}(2, M_d/2) + .2$. For $Z \sim \mathcal{N}(0, 1)$ and independent sign S, set $R = s_d \exp(.6Z)$ and $L_T(a(x)) = L_T(x) + 2SR$. Accepted transformations preserve the generated label. Task parameters (s_d, M_d) are $(.20, 2.50), (.38, 2.60), (.70, 2.10), (1.20, 2.40), (.50, 2.80), (1.80, 2.20)$ for S1–S6 and $(.27, 2.35), (1.45, 2.70), (.85, 2.15)$ for S7–S9.

Let task vectors $\theta_d, \eta_d, \zeta_d$ be independent $\mathcal{N}(0, I_2)$ draws; o, b, j are zero-based operator, budget, and recipe indices. With $e = a_b c_o$, $a = (.30, .48, .70)$, $c = (.75, 1, 1.25)$, and seed perturbations

$\epsilon_A, \epsilon_B, \epsilon_C$, the recorded mechanism is

$$
\begin{aligned}
A &= e\exp(.16\eta_d^\top w_j + .025\epsilon_A), \quad B = e\exp(-.12\eta_d^\top w_j + .025\epsilon_B),\\
\ell &= 1.25[\theta_d^\top w_j + .25\sin((o+1)(j+1)) + .12(b-1)w_{j,2}] + .04\epsilon_C,\\
\kappa_i &= .82\tanh(\ell + 1.8\zeta_d^\top w_j\,\xi_i), \qquad \xi_i = \log(R_i/s_d)/.6,\\
u_i &= AZ_{1,i}, \qquad v_i = B(\kappa_i Z_{1,i} + \sqrt{1-\kappa_i^2}Z_{2,i}).
\end{aligned} \tag{21}
$$

Student log-odds add $2s_i^\star u_i$ at the clean endpoint and $2s_i^\star v_i$ at the transformed endpoint. Binary canonicalization is $[-L/2, L/2]$, not a degenerate one-class representation. The training floor uses linear .10 quantiles, with $\delta = \tau = 1$ and rejection of a zero floor. Endpoints are stored in float32 and evaluated in float64.

Utility is computed from predictions and labels on independent samples. Its student log-odds error contains $2[.40AZ_3 + g_o(v-u)]$, $g = (.45, .80, 1.20)$. Consequently the experiment explicitly assumes a link between response difference and utility. Positive results describe behavior under this mechanism; they do not independently establish that real compressed networks follow it. In the simple scenario, $\kappa_i = .82\tanh(\ell)$; in the null-mechanism control, utility error uses $2[.40AZ_3 + g_o v]$ instead. The latter removes the explicit response-difference term while retaining the endpoint-derived features.

J.3 Matched endpoints, pairing, and nonlinear controls

The numerical $\widetilde{M}_0$ has 20 features: ID drop, nominal budget, three operator indicators, size proxy, teacher ID margin, training floor, raw squared response error, unnormalized coupled Huber, symmetric-normalized Huber, two endpoint Huber errors, probability-response error, endpoint JS, both-active direction, ID NLL and Brier differences, logit-norm difference, and validation-temperature difference. It contains no network CKA, Jacobian, Hessian, or HAWQ measurements. Adding the separate normalized endpoints of Equation 16 gives 22 features; adding ResponseScope gives 23. Missing network features are not replaced by zeros.

For $u = r_x/s_T^\star$ and $v = r_a/s_T^\star$, all three features share the same pairs, canonicalization, scale and Huber loss. The two endpoint aggregates cannot recover pairing: for $u = (.4, -.4)$, both $v_A = (.4, -.4)$ and $v_B = (-.4, .4)$ give endpoint losses .08 and .08, but paired response losses are 0 and .32. This algebraic construction concerns aggregated features; complete per-sample endpoint vectors contain the information needed to compute ResponseScope.

The pairing control normalizes first, then forms six strata from teacher response tertiles crossed with clean-teacher margin halves. Within each stratum, 32 fixed-point-free cyclic permutations are applied to transformed residuals, sharing the permutation across candidates of a task. Each permutation preserves normalized endpoint marginals. Their average $D_{\text{shuffle}} = 32^{-1}\sum_b n^{-1}\sum_i H_1(v_{\pi_b(i)} - u_i)$ is one predictor feature, not 32 independent samples or a valid input transformation. We compare held-out predictions and decisions, rather than merely asking whether a permuted metric changes numerically.

Ridge searches $\{0, 10^{-4}, 10^{-3}, 10^{-2}, .1, 1, 10, 100, 1000\}$. The shallow tree searches depth $\{2, 3, 4\}$ crossed with minimum leaf size $\{8, 16, 32\}$, also nine settings. Preprocessing is fitted within each training fold, and inner source-task leave-one-out tuning minimizes task-equal MAE. Validation-only temperature fitting searches $\{.5, .75, 1, 1.5, 2, 3, 4\}$. Target final labels and responses are never used to tune the selector.

Table 28: Separate numerical experiment: scale-matched LODO decisions on 54 cells. Regret is in binary-accuracy percentage points. Exact top-1 accepts utility ties; near-optimal coverage means regret at most .5 pp.

Rule	Mean regret	Top-1	Near-optimal
Uniform random	2.3054	20.74%	37.78%
Source configuration prior	2.1564	33.33%	40.74%
ID only	1.9233	44.44%	53.70%
$\widetilde{M}_0^+$	.1127	66.67%	92.59%
$\widetilde{M}_0^+ + D_{\text{shuffle}}$	.1127	66.67%	92.59%
$\widetilde{M}_1^+$	.0939	74.07%	92.59%
Static features + ResponseScope	.0514	85.19%	96.30%

Table 29: Numerical LODO prediction and selection controls. Regret, P90, and MAE are in accuracy percentage points.

Features / predictor	Regret	P90	MAE	R^2
$\widetilde{M}_0$	0.113209	0.363281	1.228987	0.949487
$\widetilde{M}_0 + D_{\text{ResponseScope}}$	0.075051	0.115234	1.150457	0.953472
$\widetilde{M}_0^+$	0.112666	0.301758	1.148469	0.949352
$\widetilde{M}_1^+$	0.093859	0.245117	1.112589	0.951353
$\widetilde{M}_0^+ + D_{\text{shuffle}}$	0.112666	0.301758	1.147347	0.948522
Static + ResponseScope	0.051360	0.075195	1.095239	0.959758
$\widetilde{M}_0^+ + D_{\text{cal}}$	0.109592	0.301758	1.273053	0.937838
Tree: $\widetilde{M}_0^+$	0.583225	1.916992	1.253156	0.944327
Tree: $\widetilde{M}_1^+$	0.554470	1.930664	1.437033	0.918914

Table 30: Conditional stability of $\widetilde{M}_0^+ \to \widetilde{M}_1^+$ on fixed numerical out-of-fold outputs.

Contrast	Point estimate	Conditional 95% interval	Improved tasks
Regret reduction (pp)	0.018808	[-0.036531, 0.082465]	2/6
MAE reduction (pp)	0.035879	[-0.000641, 0.067797]	5/6
R^2 increase	0.002001	[0.000033, 0.003806]	5/6

10,000 task-resampling draws without refitting. Regret and MAE intervals cross zero.

The paired feature has a lower ridge regret point estimate than the shuffled feature, but the interval for their regret difference crosses zero. Shallow-tree regret decreases from .583225 to .554470 pp while MAE rises from 1.253156 to 1.437033 and R^2 falls from .944327 to .918914. The tree's conditional regret-reduction interval is $[-.113806, .163054]$ pp. Thus neither a predictor-independent gain nor a statistically stable regret advantage is established. All intervals resample six fixed out-of-fold task outputs 10,000 times without refitting; they do not estimate uncertainty over a new task population or the experiment's design choices.

J.4 Changed decisions, operator concentration, and failure

Table 31: Numerical operator-group regret (pp), 18 cells per group.

Parameter group	$\widetilde{M}_0^+$	$\widetilde{M}_1^+$	Reduction	Better / worse / same
distillation-like	0.076497	0.035265	0.041233	5 / 0 / 13
pruning-like	0.105794	0.090603	0.015191	1 / 1 / 16
quantization-like	0.155707	0.155707	0.000000	0 / 0 / 18

Adding ResponseScope to the matched endpoints improves six decisions, worsens one, and leaves 47 unchanged. No quantization-like decision changes. This decomposition cannot explain the original benchmark's 81.6% quantization contribution, because the parameter groups and candidate generation differ.

On S1, pruning-like group, budget .25, c03 and c04 have identical ID drop (.390625 pp). Their normalized clean/transformed endpoint errors are $(.111719, .119305)$ and $(.095013, .145232)$, their selection ResponseScope values are .213784 and .182213, and utility damages are 6.669922 and 5.185547 pp. The decision changes from c03 to c04, reducing regret from 1.484375 to zero. Endpoints favor different candidates, while the paired summary helps the selector resolve the ranking.

On S5, pruning-like group, budget .125, the decision instead changes from optimal c04 to c01, increasing regret from zero to 1.210938 pp. Selection-stage ResponseScope favors c01 (.327448 vs. .366550), whereas final ResponseScope favors c04 (.312511 vs. .377071 for c01); utility damages are 8.339844 and 9.550781 pp. This failure analysis follows selection; it does not justify feeding final responses back into selection or identifying a unique cause from one case.

J.5 Configuration-level aggregation and numerical record reconstruction

For every numerical configuration, each seed's metric and utility are computed first, then averaged:

$$\bar{D}_{d,o,b,j} = \tfrac{1}{5}\sum_{s} D_{d,o,b,j,s}, \qquad \bar{U}_{d,o,b,j} = \tfrac{1}{5}\sum_{s} U_{d,o,b,j,s}.$$

This selects configurations by their average behavior over five specified seeds; it does not select a checkpoint, a best seed, or a deployed ensemble. The cross-configuration mean ResponseScope is .253994 for metric-then-mean, versus .059522 for mean-logits-then-metric, with maximum per-configuration difference .726954. The nonlinear aggregations are not interchangeable.

Table 32: Numerical seed-specific LODO sensitivity; each row uses a single seed rather than the five-seed configuration mean.

Seed	Regret $\widetilde{M}_0^+$	Regret $\widetilde{M}_1^+$	MAE $\widetilde{M}_0^+$	MAE $\widetilde{M}_1^+$
7	0.272172	0.151910	1.330943	1.305271
19	0.125687	0.103082	1.380282	1.409453
31	0.161856	0.229673	1.671808	1.609432
43	0.166377	0.138346	1.790149	1.628131
59	0.153718	0.121166	1.629082	1.543199

Four seed-specific LODO comparisons reduce regret and one increases it; seed 19 also worsens MAE. These assess seed sensitivity rather than independent task panels. The rule does not supply the missing historical aggregation for the six benchmark tasks.

The verification record reports maximum absolute reconstruction errors of $2.220446049250313 \times 10^{-16}$ for ResponseScope from stored endpoints and zero for utility from predictions and labels. Self-comparison ResponseScope is zero; jointly scaling teacher, student and floor by .5 or 2 changes it by zero. The Huber–quadratic identity error is at most $8.881784197001252 \times 10^{-16}$, and the zero-floor case is rejected. These checks concern the numerical record and common scaling, not student-only scale invariance or historical benchmark execution.

The record chain is: task/stage pair manifests and per-seed endpoints/utility outputs; per-seed metrics; metric-then-mean configuration records; nested tuning and out-of-fold predictions; within-cell choices and oracle-referenced regret; task-equal summaries. The paper's source inventory maps the imported CSV tables to each new table. We transcribe recorded outputs without executing the experiment or asserting an independent local rerun.

J.6 AUXILIARY ENGINEERING CHECKS: STORAGE ACCOUNTING AND RELOADS

Table 33: Independent numerical serialization objects: component bytes and realized ratio.

Object	Target	Parameters	Indices	Scales	Total	Ratio
Dense shape	0.500	2,097,152	0	0	2,097,408	0.500031
Dense shape	0.250	1,048,576	0	0	1,048,832	0.250046
Dense shape	0.125	524,288	0	0	524,544	0.125053
Compact channels	0.500	2,097,152	2,048	0	2,099,456	0.500519
Compact channels	0.250	1,048,576	1,024	0	1,049,856	0.250290
Compact channels	0.125	524,288	512	0	525,056	0.125175
FP16	0.500	2,097,152	0	0	2,097,408	0.500031
INT8	0.250	1,048,576	0	4,096	1,052,928	0.251022
INT4	0.125	524,288	0	4,096	528,640	0.126030

Every row additionally has 256 metadata bytes and 0 untouched-module bytes, included in the total. Reference total: 4,194,560 bytes. All reload errors are zero.

Nine numerical matrices form a separate serialization experiment. The reference is a 1024×1024 FP32 matrix with 4,194,304 parameter bytes plus a 256-byte header. Each object has 256 metadata bytes and zero untouched-module bytes. All nine recorded pre-/post-reload outputs are elementwise identical (maximum absolute error zero); INT4 packs two four-bit values into a byte. The comparison is between the same numerical representation before and after export, not between a quantized matrix and its FP32 reference. It measures neither network inference-path equivalence nor deployment-kernel performance; zero original benchmark models were reloaded.

Compact-channel matrices in this experiment are not CSR. For an $m \times n$ CSR matrix with FP32 values and 32-bit column indices and row pointers,

$$B_{\mathrm{CSR}} = 8\,\mathrm{nnz} + 4(m+1), \qquad \frac{B_{\mathrm{CSR}}}{4mn} = 2d + \frac{m+1}{mn}, \quad d = \frac{\mathrm{nnz}}{mn}.$$

Consequently nonzero density is not a serialized storage ratio. The benchmark table's simultaneous channel/head-pruning, density, and CSR descriptions still need a record of physical tensor shapes and density semantics. These nine objects do not resolve that ambiguity.

J.7 AUXILIARY AUDIT CHECKS AND WEAK TEACHER RESPONSES

All 5,616 numerical measurement pairs have an audit record: 3,744 for S1–S6 and 1,872 for S7–S9. Source samples are separately generated by stage; task, stage, source, group, and pair identifiers retain that lineage. This provides linkage for the numerical experiment only, not the missing sampling relation between historical benchmark audit and measurement populations.

Label preservation is known from the generator. Two annotation channels each have a specified 2% error rate, followed by adjudication against known truth; there are zero human annotators. Agreement and Cohen's κ are calculated over proposals and negative controls, with 2,000 annotation-row resamples for conditional intervals. They test annotation accounting but cannot validate a real transformation family or establish Waterbirds mask provenance.

Table 34: Numerical annotation accounting: generated annotation channels, not human validation.

Task	Accepted / proposed	Agreement	κ	Conditional 95% interval	Valid controls
S1	624/645	97.25%	0.9114	[0.8712, 0.9454]	2/120
S2	624/638	96.31%	0.8753	[0.8269, 0.9194]	3/120
S3	624/637	96.96%	0.8966	[0.8541, 0.9351]	4/120
S4	624/638	94.33%	0.8142	[0.7608, 0.8668]	2/120
S5	624/639	95.52%	0.8551	[0.8029, 0.8984]	3/120
S6	624/645	96.99%	0.8997	[0.8578, 0.9380]	6/120
S7	624/641	94.74%	0.8278	[0.7737, 0.8793]	7/120
S8	624/641	96.32%	0.8788	[0.8323, 0.9212]	3/120
S9	624/644	95.42%	0.8530	[0.8046, 0.9000]	2/120

Accepted/proposed pools training, selection, and final proposals. Valid controls counts label-preserving outcomes among 120 negative controls; S7 has 7/120, so benchmark negative-control thresholds must not be claimed to pass universally here.

Table 35: Numerical teacher-response floors and weak-response fractions ($r_T < q_{.1}^{\mathrm{train}}$).

Task	Training floor	Training	Selection	Final
S1	0.099213	10.16%	13.33%	13.33%
S2	0.176657	10.16%	9.17%	9.17%
S3	0.344922	10.16%	15.83%	11.67%
S4	0.545888	10.16%	10.00%	13.33%
S5	0.231200	10.16%	12.50%	10.83%
S6	0.891162	10.16%	10.00%	9.17%
S7	0.125930	10.16%	5.00%	7.50%
S8	0.639078	10.16%	14.17%	5.00%
S9	0.363105	10.16%	11.67%	3.33%

Floor-active (weak-response) means $r_T < q_{.1}^{\mathrm{train}}$. The training fraction 10.15625% reflects finite samples and linear quantile interpolation, not target-adaptive flooring. Every numerical task has a positive floor, and the table reports the corresponding floor-activation rates.

J.8 Holdouts and retained negative results

Table 36: Numerical holdouts and negative results. The baseline and augmented models share normalized endpoint features.

Setting	Model	Regret (pp)	MAE (pp)	R^2
LODO	$\widetilde{M}_0^+$	0.112666	1.148469	0.949352
LODO	$\widetilde{M}_1^+$	0.093859	1.112589	0.951353
Parameter-group holdout	$\widetilde{M}_0^+$	0.131474	1.242890	0.943917
Parameter-group holdout	$\widetilde{M}_1^+$	0.128038	1.165502	0.951363
Reserved S7–S9	$\widetilde{M}_0^+$	0.000000	1.327616	0.935648
Reserved S7–S9	$\widetilde{M}_1^+$	0.000000	1.228948	0.944379
Simple scenario	$\widetilde{M}_0^+$	0.007234	1.245106	0.935944
Simple scenario	$\widetilde{M}_1^+$	0.009223	1.159181	0.950699
No response-difference utility	$\widetilde{M}_0^+$	0.159144	0.771929	0.962360
No response-difference utility	$\widetilde{M}_1^+$	0.180302	0.775061	0.962167

Pairs S1–S2, S3–S4, and S5–S6 define three parameter groups. Although the source uses visual-like, text-like, and time-series-like labels, their joint holdout is not cross-modality validation on images, text, or sequences. Reserved S7–S9 tasks come from the specified numerical mechanism; both ridge rules achieve zero regret, leaving only prediction-error improvement. Neither result establishes transfer to new network architectures or compression mechanisms under a prospectively fixed method.

The simpler scenario improves regression but increases regret from .007234 to .009223 pp. Removing response-difference dependence worsens regret (.159144 to .180302), MAE (.771929 to .775061), and R^2 (.962360 to .962167). These failures support a conditional response-summary interpretation rather than a universally effective selector.

For the historical benchmark, all 15 leave-two-dataset-out splits include three same-modality pairs. Reporting those three folds separately would be a subset analysis, not a new experiment. The current controlled-study group-holdout rows concern S1–S6 and cannot supply that benchmark subset. Fixed-panel resampling, within-panel task holdout, and external validation after freezing remain distinct evidence levels.

K PRIOR-ADJUSTED COUPLING STUDY

Table 37: Scope and reconstruction checks for the prior-adjusted coupling study. Benchmark auditing is reported separately in Table 26.

Requirement	Controlled-study record
Statistical object	Five per-seed metrics averaged per configuration; 2,025 unique state/output records
Storage / output identity	Twelve independent matrix reloads and 2,025 generator-state reconstructions
Prior eligibility	Five registry-order direction recipes; cross-fitted source prior
Measurement validity	All 5,616 pairs linked to generated-truth audit records; no humans
Incremental attribution	Matched endpoints, prior, shuffle, and shallow trees on numerical tasks

K.1 SCOPE, STATISTICAL OBJECT, AND RECONSTRUCTION RECORDS

This independent numerical experiment was executed on six parameterized binary tasks (D01–D06) and three reserved parameter tasks (H01–H03), with three operator-like parameter groups, three nominal damage budgets (.500, .250, .125), and five directional recipes per cell. It includes 270 main and 135 reserved configurations, corresponding to 1,350 and 675 numerical seed instances. These groups are not actual distillation, pruning, or quantization operations, and their budgets are not measured candidate storage ratios. No benchmark networks were trained or reloaded, and no human annotators participated in this study. The controlled endpoint-scale study is separate; its sample counts and results are not substituted here.

The statistical object is a configuration's average performance over $\mathcal{S} = \{7, 19, 31, 43, 59\}$, not a chosen checkpoint or ensemble:

$$\bar{D}_{dobj} = \frac{1}{5}\sum_{s\in\mathcal{S}} D_{dobj,s}, \qquad \bar{U}_{dobj} = \frac{1}{5}\sum_{s\in\mathcal{S}} U_{dobj,s}. \tag{22}$$

Metrics are computed from each seed's logits and then averaged. On the 270 main configurations, the mean ResponseScope is .240488618 under this rule but .055211558 if logits are averaged first; the maximum configuration difference is .747699653. The latter is a different estimand. For D01, group 3, budget .250, recipe R01, the five ResponseScope values are (.308806077, .233033388, .318855889, .261413747, .293085411), averaging .283038903; utility damages are (12.597656250, 13.623046875, 13.330078125, 12.304687500, 12.792968750) pp, averaging 12.929687500 pp.

The recorded link is configuration/seed → run ID → generator state and input → stored logits → per-seed metrics → configuration mean → out-of-fold predictions → decisions/regret. The execution records report 2,025 unique state hashes and selection-logit hashes; reconstructing logits, metrics, utility, and configuration means has maximum error zero. Reconstructing 2,376 method/scenario decisions also has error zero; these are not 2,376 independent cells. Scalar/vector metric agreement and the Huber identity have maximum errors 2.22×10^{-16} and 7.77×10^{-16}. These are recorded numerical checks, not benchmark deployment-equivalence evidence. Hashes identify content, not public preregistration or the time an experiment was executed.

Per task, training-floor, selection, and final-response stages use 384, 120, and 120 accepted pairs; ID validation and utility use 256 and 2,048 inputs. The 5,616 accepted pairs across nine tasks are shared across candidates and seeds rather than counted repeatedly. Target final-response features and utility labels are excluded from selection; source utility labels supervise the predictor. Stored float32 logits are converted to float64 for the metric, with binary embedding $\phi(L) = (-L/2, L/2)$, $\tau = \delta = 1$, a linear training .10 quantile, and an equal pair mean of class means.

K.2 FIXED RECIPES AND EXPLICIT RESPONSE–UTILITY MECHANISM

Recipes R01–R05 are assigned in registry order before scoring: $w_j = (\cos(2\pi j/5), \sin(2\pi j/5))$, $j = 0, \ldots, 4$. Their oracle counts are respectively 23, 18, 8, 2, and 3 of 54 cells; there are no tied main-panel utility optima. The records contain unique numerical states and outputs rather than replicated configuration means. These identities do not recover the historical benchmark c01–c05

recipes. The source-task prior is

$$P_{dobj} = |\mathcal{D}_{\mathrm{src}}|^{-1} \sum_{d' \in \mathcal{D}_{\mathrm{src}}} Y_{d'obj}.$$

For source training rows, the prior excludes that row's task; it is rebuilt inside every inner and outer fold with source-only preprocessing. Prior-only selects its minimum, while combined predictors receive it as a feature.

The generator draws $Y \sim \mathrm{Bernoulli}(.5)$, $m \sim \mathrm{Gamma}(2, M_d/2) + .2$, and clean teacher log-odds $L_T = (2Y-1)m$. With $\xi \sim N(0,1)$, $R = s_d \exp(.6\xi)$, and independent equiprobable sign S, the transformed teacher log-odds are $L_T + 2SR$. Labels are fixed by construction. Main-task (s_d, M_d) pairs are $(.20, 2.50), (.38, 2.60), (.70, 2.10), (1.20, 2.40), (.50, 2.80), (1.80, 2.20)$; reserved pairs are $(.27, 2.35), (1.45, 2.70), (.85, 2.15)$.

Let independent task vectors $\theta_d, \eta_d, \zeta_d \sim N(0, I_2)$, candidate perturbations $\epsilon_A, \epsilon_B, \epsilon_C \sim N(0,1)$, and residual inputs $Z_1, Z_2, Z_3 \sim N(0,1)$ be generated with master seed 20260921 and the five numerical seeds. Residual inputs are shared between candidates within a task and seed. With zero-based o, b, j,

$$\begin{aligned}
&e = a_b c_o, \quad a = (.30, .48, .70), \quad c = (.75, 1, 1.25),\\
&A = e\exp(.16\eta_d^\top w_j + .025\epsilon_A), \quad B = e\exp(-.12\eta_d^\top w_j + .025\epsilon_B),\\
&\ell = 1.25[\theta_d^\top w_j + .25\sin((o+1)(j+1)) + .12(b-1)w_{j,2}] + .04\epsilon_C,\\
&\kappa_i = .82\tanh(\ell + 1.8\zeta_d^\top w_j \xi_i), \quad u_i = AZ_{1,i}, \quad v_i = B(\kappa_i Z_{1,i} + \sqrt{1-\kappa_i^2} Z_{2,i}).
\end{aligned}$$

These are scalar normalized residual coefficients: vector residuals are $(-u_i, u_i)$ and $(-v_i, v_i)$ under the binary embedding. Student endpoint log-odds are $L_S(x_i) = L_T(x_i) + 2s_i^\star u_i$ and $L_S(a(x_i)) = L_T(a(x_i)) + 2s_i^\star v_i$. On independent utility inputs, the linked scenario sets

$$L_S^{\mathrm{utility}} = L_T + 2[.40AZ_3 + g_o(v-u)], \qquad g = (.45, .80, 1.20). \tag{23}$$

Thus, response dependence is an assumption of the generator, not a discovered compression mechanism. Teacher clean accuracy on utility inputs is 100% by construction and damage is $100(1-\bar{U})$. The endpoint-only control replaces $g_o(v-u)$ by $g_o v$; shared latent variables can still induce response–utility association. Reserved tasks share the generator and are excluded from fitting and tuning; they are not new neural architectures.

K.3 Matched feature sets, pairing control, and nested selection

The 20-feature $\widetilde{M}_0$ is distinct from benchmark M_0. Static features comprise ID damage, nominal budget, three group indicators, and a size proxy equal to nominal budget. This last column is redundant and does not measure parameter count. The 14 other descriptors are teacher ID margin, training floor, raw squared response error, coupled Huber response, symmetric-normalized Huber response, two endpoint Huber errors, probability response error, endpoint JS, both-active direction error, and NLL, Brier, logit-norm, and temperature shifts. No network CKA, Jacobian, or Hessian measurements are claimed for numerical candidates.

For normalized vector residuals $u = r_x/s_T^\star$, $v = r_a/s_T^\star$, retain two separate features $E_x = \mathbb{E}[h_1(u)]$ and $E_a = \mathbb{E}[h_1(v)]$. Then $\widetilde{M}_0^+$ has 22 features and $\widetilde{M}_1^+ = \widetilde{M}_0^+ + D_{\mathrm{RS}}$ has 23; adding P gives 23 and 24 features. The comparison uses identical pairs, canonicalization, teacher scales, Huber parameter, folds, and search budget. For the shuffle control, normalize first, stratify by teacher-response tertile and clean-margin half, and perform 32 fixed-point-free cyclic shifts within the six strata, shared across candidates. Average their losses into one feature. This preserves normalized residual marginals and changes pairing only in the added feature; other paired descriptors remain. These are computational controls, not valid transformed inputs or 32 independent observations.

Six outer task folds each use inner leave-one-source-task-out tuning. The primary objective is source-task-equal regret; source-task-equal MAE is a prespecified secondary protocol. Ridge searches $\alpha \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$, resolving validation ties in favor of stronger regularization. Shallow trees search depths {2,3,4} and minimum leaf sizes {8,16,32}, preferring

shallower trees and larger leaves for validation ties. Predictor ties use registry order; top-1 counts any exact utility optimum. All rules retain every cell. The execution-order freeze is not a public preregistration. The recorded target-label perturbation check covers D01's outer fold and nine ridge settings; zero prediction change is a limited check, not a proof against all leakage paths.

Table 38: Complete prior-adjusted coupling-study selection results. The inner objective is source-task-equal regret. RS abbreviates ResponseScope; all errors are binary-accuracy percentage points.

Rule	Mean regret	P90	Maximum	Top-1/54	$\leq$ .5/54	MAE
Recipe prior	1.1802	3.2168	5.3418	13	26	1.6162
Static	1.2764	3.5742	11.4355	25	36	4.1960
$\widetilde{M}_0$	0.1409	0.5420	1.6602	39	48	2.4315
$\widetilde{M}_0$+RS	0.2136	0.6807	3.1055	40	47	2.1204
$\widetilde{M}_0^+$	0.1409	0.5420	1.6602	39	48	2.1071
$\widetilde{M}_0^+ + D_{\mathrm{shuffle}}$	0.3673	0.7695	11.4355	37	47	2.3840
$\widetilde{M}_1^+$	0.2094	0.6807	3.1055	36	47	2.1450
$P + \widetilde{M}_0^+$	0.3078	0.3779	10.9082	39	49	2.1238
$P + \widetilde{M}_0^+ + D_{\mathrm{shuffle}}$	0.3078	0.3779	10.9082	39	49	2.0632
$P + \widetilde{M}_1^+$	0.1751	0.6807	1.4941	34	47	1.8955
$P + \widetilde{M}_0^+ + D_{\mathrm{cal}}$	0.1467	0.5596	1.0254	36	48	2.1532
Static + native RS	0.0930	0.3779	1.0254	41	50	1.3986
Tree: $P + \widetilde{M}_0^+$	1.1549	3.4893	7.9590	23	34	1.8589
Tree: $P + \widetilde{M}_1^+$	1.1352	3.4893	7.9590	23	35	1.8467

Table 39: Conditional uncertainty for the prespecified coupling-study contrast and its shallow-tree check. Positive differences favor adding RS.

Predictor / objective	Metric	Difference	Conditional 95% interval	Tasks +/–
Ridge / regret	regret	0.1327	[-0.1449, 0.5707]	1/2
Ridge / regret	MAE	0.2284	[-0.2922, 0.7819]	3/3
Ridge / regret	R^2	0.0070	[-0.1140, 0.1204]	3/3
Tree / regret	regret	0.0197	[-0.0239, 0.0830]	1/1
Tree / regret	MAE	0.0122	[-0.0376, 0.0814]	2/3
Tree / regret	R^2	0.0060	[-0.0011, 0.0134]	3/2
Ridge / MAE	regret	0.0722	[-0.0051, 0.1825]	2/1
Ridge / MAE	MAE	0.2029	[0.1128, 0.3350]	6/0
Ridge / MAE	R^2	0.0361	[0.0163, 0.0659]	6/0
Tree / MAE	regret	0.0875	[-0.0159, 0.2232]	2/1
Tree / MAE	MAE	0.0286	[-0.0145, 0.0893]	3/3
Tree / MAE	R^2	0.0085	[0.0023, 0.0152]	5/1

Each interval resamples six fixed task-level out-of-fold summaries 10,000 times, with zero refits inside the interval. Unchanged tasks are omitted from the last column.

Table 40: Prespecified inner objectives are reported separately; target outcomes do not select the preferred protocol.

Rule	$\mathrm{Regret}_{\mathrm{regret}}$	$\mathrm{Regret}_{\mathrm{MAE}}$	$\mathrm{MAE}_{\mathrm{regret}}$	$\mathrm{MAE}_{\mathrm{MAE}}$
$\widetilde{M}_0^+$	0.1409	0.2025	2.1071	2.5168
$\widetilde{M}_1^+$	0.2094	0.1376	2.1450	2.2566
$P + \widetilde{M}_0^+$	0.3078	0.1657	2.1238	2.0409
$P + \widetilde{M}_1^+$	0.1751	0.0935	1.8955	1.8379
Static + native RS	0.0930	0.0930	1.3986	1.2675

The main prior-adjusted comparison reduces mean regret by .132740 pp; only one task improves and two worsen. P90, top-1, and near-optimal coverage worsen even as maximum regret falls from 10.9082 to 1.4941 pp. Calibrated ResponseScope has lower regret (.1467) than native (.1751), with higher MAE; neither dominates. Without the prior, scale-matched regret worsens from .1409 to .2094. Under MAE tuning, the prior-adjusted MAE reduction is .202945 pp, interval [.112824,.334983], with all six tasks improving, while the regret-reduction interval remains $[-.005064, .182473]$. This

secondary result is not promoted retrospectively to the primary selection claim. Uniform random selection has exact mean regret 2.196759 pp, 20% top-1, and 37.78% near-optimal coverage; its 6.398438-pp P90 is over 270 cell–candidate outcomes, not 54 cell-wise expected regrets.

K.4 Within-cell errors, scenario failures, and seed sensitivity

For each cell's ten candidate pairs we evaluate $\Delta Y_{jk} = Y_j - Y_k$ against $\Delta \widehat{Y}_{jk}$. True ties are excluded from the pair-accuracy denominator; predicted ties receive half credit. Difference MAE includes all ten pairs. Centered MAE subtracts each cell's actual and predicted means only for frozen-result evaluation; the target true mean is never an input. Scores are averaged within task and then equally across tasks. Higher pair accuracy here coexists with worse difference and centered MAE.

Table 41: Within-cell prediction diagnostics for the primary controlled-study protocol. Errors use percentage points.

Rule	R^2	Pair accuracy	Difference MAE	Centered MAE
$P + \widetilde{M}_0^+$	0.8157	0.8698	1.4199	0.9074
$P + \widetilde{M}_1^+$	0.8227	0.8971	1.8900	1.2522
Static + native RS	0.9224	0.9403	1.0578	0.6819

Table 42: Scenario and reserved-parameter results. These panels share a numerical generator, not a new architecture.

Scenario / panel	Rule	Regret	MAE	R^2
Linked / main	$P + \widetilde{M}_0^+$	0.3078	2.1238	0.8157
Linked / main	$P + \widetilde{M}_1^+$	0.1751	1.8955	0.8227
Linked / main	Recipe prior	1.1802	1.6162	0.8691
Linked / main	Static + native RS	0.0930	1.3986	0.9224
Linked / reserved	$P + \widetilde{M}_0^+$	0.0611	0.8723	0.9679
Linked / reserved	$P + \widetilde{M}_1^+$	0.0937	0.8511	0.9717
Linked / reserved	Recipe prior	0.4753	1.3845	0.9157
Linked / reserved	Static + native RS	0.0539	1.5710	0.9088
Endpoint / main	$P + \widetilde{M}_0^+$	0.2230	1.1815	0.9081
Endpoint / main	$P + \widetilde{M}_1^+$	0.2230	1.1693	0.9116
Endpoint / main	Recipe prior	0.4494	0.7916	0.9478

The endpoint-only primary contrast has identical regret, not necessarily identical results for every other feature set. Reserved-task prediction MAE improves while selection regret worsens. These outcomes delimit the assumed mechanism rather than establish external generalization.

The five seed-sensitivity analyses each use that seed's metrics and rerun the entire nested selector. They vary candidate residuals rather than merely relabeling a fixed mean. Tasks, teacher, inputs, and recipes remain shared; four favorable rows do not constitute four independent task panels. In contrast, the conditional intervals in Table 39 resample fixed OOF task summaries without refitting.

Table 43: Full nested selection using one numerical seed at a time, with the regret inner objective.

Seed	Baseline regret	With RS
7	0.2957	0.1447
19	0.1230	0.2550
31	0.1266	0.1175
43	0.2215	0.1944
59	0.2188	0.1808

K.5 Decision changes and a failed applicability condition

The prespecified observable condition requires $\arg\min_j E_{x,j} \neq \arg\min_j E_{a,j}$ and $(\max_j D_j - \min_j D_j)/\operatorname{mean}_j D_j \geq .10$. All 54 cells satisfy it, leaving no comparison stratum. No threshold

is changed after observing outcomes, and no prospective rule for when to acquire ResponseScope is validated. Group-wise mean regret changes are .0081→.0326 (distillation-like), .1492→.1687 (pruning-like), and .7661→.3239 (quantization-like). Improvement/worsening/unchanged counts are 0/3/15, 1/1/16, and 1/3/14. These parameter-group effects cannot explain the benchmark's 81.6% quantization contribution.

Table 44: All nine changed decisions in the primary controlled-study comparison; the remaining 45 decisions are unchanged. Group numbers are numerical parameters.

Task	Group / budget	Recipe: base o RS	Base regret	RS regret	Reduction
D03	3 / .250	R01→R02	0.0000	1.4941	-1.4941
D04	1 / .125	R01→R02	0.0000	0.1367	-0.1367
D04	3 / .125	R03→R01	10.9082	0.0000	10.9082
D06	1 / .500	R03→R04	0.0000	0.0684	-0.0684
D06	1 / .250	R03→R04	0.0000	0.2344	-0.2344
D06	2 / .250	R04→R03	0.0000	0.5371	-0.5371
D06	2 / .125	R04→R03	0.1855	0.0000	0.1855
D06	3 / .500	R05→R03	0.0000	1.0254	-1.0254
D06	3 / .250	R05→R03	0.0000	0.4297	-0.4297

Influence of the largest correction. For the primary contrast, let Δ_c denote baseline minus augmented regret in cell c. The 54 recorded differences sum to 7.16796875 pp; D04, group 3, budget .125 contributes 10.908203125 pp. Thus

$$\frac{\sum_c \Delta_c - 10.908203125}{53} = \frac{-3.740234375}{53} = -.07057046 \text{ pp.}$$

This is a post-hoc influence calculation from the fixed decision ledger, not a refit, an alternative primary estimand, or a reason to discard the cell. It weights the remaining cells equally rather than preserving equal task weights after deletion. The prespecified primary result retains all 54 cells and equal task weights. Its positive mean depends on avoiding one large error, consistent with worse top-1 and near-optimal coverage.

The response-span condition likewise requires ResponseScope to have already been computed. Even if it had separated favorable cells, it would describe post-acquisition variation rather than establish when acquiring the metric is worthwhile. A prospective acquisition rule would need cheaper observables or a pilot whose measurement cost is included.

K.6 Measurement population and weak-response diagnostics

For each task and stage, the generator proposes pairs with a 4% invalidity probability until the required valid sample count is reached. Two generated annotation channels independently flip known truth with probability .02; known truth, not channel agreement, determines admission. Each task also has 120 negative-control proposals with 3% truth-preserving probability; none enter measurement. All 5,616 used pairs link to audit rows with source, group, stage, and generator identity. Disjoint group namespaces across stages are a design property, not evidence about benchmark users or subjects. The channel agreement and κ intervals use proposals plus negative controls and 2,000 generated-group resamples. They do not estimate human judgment quality or real transformation validity.

Table 45: Generated-channel audit for the controlled study, not human annotation. Accepted/proposed counts refer to its measurement population.

Task	Train	Selection	Final	Agreement	κ	Conditional 95% interval
D01	384/400	120/126	120/125	96.89%	0.8997	[0.8529, 0.9369]
D02	384/400	120/123	120/123	96.74%	0.8934	[0.8451, 0.9331]
D03	384/409	120/122	120/124	96.77%	0.9009	[0.8557, 0.9374]
D04	384/403	120/122	120/123	96.22%	0.8820	[0.8300, 0.9263]
D05	384/400	120/124	120/123	96.87%	0.9000	[0.8464, 0.9411]
D06	384/409	120/123	120/125	95.11%	0.8462	[0.7904, 0.8928]
H01	384/400	120/124	120/126	96.23%	0.8796	[0.8247, 0.9216]
H02	384/409	120/126	120/128	95.91%	0.8782	[0.8253, 0.9223]
H03	384/396	120/122	120/122	96.45%	0.8865	[0.8298, 0.9282]

The old `intensity_bin` field records proposal-index strata rather than observed transformation strength. A post-result descriptive check bins $\log(r_T/s_d)/.6$ at $-.6744897502, 0, .6744897502$; the 108 task–stage–bin count rows retain the actual response distribution. This study supplies no Waterbirds masks, text judgments, or ECG expert review.

Weak response (floor-active) means $r_T < q_T^{\text{train}}$, rather than no response. Strict zero is $r_T = 0$, and the separate numerical near-zero diagnostic uses $r_T \leq 10^{-8}$. All nine tasks have zero strict-zero and near-zero counts across the three stages. A teacher's near-invariance can be desirable; the frozen floor is not changed to improve results.

Table 46: Controlled-study weak-response counts by stage. All strict-zero and near-zero counts are zero; near-zero means $r_T \leq 10^{-8}$.

Task	Training floor	Train / 384	Selection / 120	Final / 120
D01	0.093731	39	12	14
D02	0.167185	39	9	10
D03	0.327514	39	13	12
D04	0.542324	39	15	11
D05	0.223343	39	9	3
D06	0.866382	39	14	14
H01	0.117897	39	11	8
H02	0.646158	39	12	13
H03	0.429337	39	13	14

Final weak/active group errors are computed separately from logits. For D01–D06, weak raw errors are (.005274,.013595,.053807,.163616,.028154,.411565), and weak ResponseScope is (.243588,.197066,.204077,.237186,.200329,.231323). Active raw errors are (.062163,.220984,.498959,1.807010,.289988,4.536952), and active ResponseScope is (.254661,.239834,.221246,.206752,.236501,.255209). For H01–H03, the corresponding weak raw, weak ResponseScope, active raw, and active ResponseScope triples are (.005591,.231348,.096561), (.168788,.216743,.226801), (.085570,1.623599,1.155962), and (.243988,.188969,.271137). These are configuration and seed averages, not fixed rescalings of a common statistic. They diagnose response-conditioned errors and do not identify causal failure mechanisms.

K.7 Auxiliary engineering checks and numerical cost

Twelve independent matrix files illustrate component accounting. Compact FP32 objects have 512 columns and 256, 128, or 64 rows, with an optional channel map; full-shape precision and CSR objects have 512×512 entries. For FP32 values and 32-bit CSR indices/pointers, $B_{\text{CSR}} = 8\,\text{nnz} + 4(m+1)$ before metadata. Density .5 thus produces 1,050,884 total bytes and ratio 1.001956, not .5. These files do not resolve the original benchmark's pruning representation.

Table 47: Twelve independently serialized numerical matrices. All rows have 256 metadata bytes, zero untouched-module bytes, and a zero same-representation reload error.

Object	Value B	Index B	Scale B	Total B	Ratio
Dense FP32, 256 rows	524288	0	0	524544	0.500122
Compact + map, 256 rows	524288	1024	0	525568	0.501098
Dense FP32, 128 rows	262144	0	0	262400	0.250183
Compact + map, 128 rows	262144	512	0	262912	0.250671
Dense FP32, 64 rows	131072	0	0	131328	0.125214
Compact + map, 64 rows	131072	256	0	131584	0.125458
FP16	524288	0	0	524544	0.500122
INT8	262144	0	2048	264448	0.252136
INT4 (packed)	131072	0	2048	133376	0.127166
CSR, density .500	524288	526340	0	1050884	1.001956
CSR, density .250	262144	264196	0	526596	0.502079
CSR, density .125	131072	133124	0	264452	0.252140

Reference: 512×512 FP32 plus 256 bytes, totaling 1,048,832 bytes. These objects are not the 270 primary candidates.

Reload compares the same deterministic numerical representation and fixed input using float64 matrix arithmetic, with a tolerance fixed at zero. Quantized reload agreement is not agreement with the pre-quantization FP32 matrix, and zero tolerance does not transfer to heterogeneous neural kernels. The 2,025 finite-final-sample binary NLL checks satisfy the corresponding empirical Lipschitz bound; they are neither a population confidence bound nor a bound on the independent utility distribution. No additional recovery optimization was run. The zero-ResponseScope endpoint counterexample remains unchanged.

Table 48: Recorded controlled-study selector timing in milliseconds: five configurations, five seeds each, five warmups and 30 repeats on one CPU thread.

Rule	Cached median	Cached P95	No-logit median	No-logit P95
Static + native RS	0.752	0.872	1.776	3.194
$P + \widetilde{M}_0^+$	10.950	13.518	11.454	12.451
$P + \widetilde{M}_1^+$	10.641	13.261	11.942	12.868

Timing covers selection/ID feature extraction, standardization, frozen meta-prediction, and selection. The fixed cell is D01, group 3, budget .250. The recorded environment is Linux x86-64, Python 3.13.5, NumPy 2.3.5, scikit-learn 1.8.0, one CPU thread, no GPU; the processor model is not recorded. Without cached logits, inputs and generator parameters remain resident, and numerical forward computation is repeated. This is not an operating-system cold cache or neural inference. Offline nested fitting, initial input generation, candidate training, and human audit are excluded. The slightly smaller cached median for the augmented rule does not imply negative marginal cost. Separate marginal ResponseScope aggregation and end-to-end benchmark selector costs remain unmeasured.

Record map and interpretation. The record package provides the protocol, run/seed/configuration rows, input identities, OOF predictions, all decisions, tuning choices, audit rows, reconstruction summaries, and cost receipts. The appendix tables are transcriptions of these records; no local experiment execution is asserted. Some narrative sentences in the accompanying summary conflict with its own tables and machine-readable execution counts. We retain the recorded counts and explicit generator definition, while distinguishing these numerical checks from the completed benchmark-level identity, aggregation, deployment, and intervention-validity audits reported elsewhere in this appendix.

L MATCHED DESCRIPTOR STUDY WITH EXPLICIT RECIPE MEANS

This numerical experiment, executed on September 22, is reported separately from both the six named benchmarks and the earlier studies in Appendices J and K. All values below are transcribed from the supplied September 22 execution records. The selected object is a recipe's mean over the fixed numerical seeds $\mathcal{S} = \{7, 19, 31, 43, 59\}$: each seed's RS and utility are computed first, then averaged separately. It is neither a checkpoint nor an ensemble prediction. The 270 main and 135 reserved configurations correspond to 1,350 and 675 saved numerical states. The three utility scenarios share states and inputs; their 6,075 seed–scenario metric rows do not represent 6,075 independent states.

L.1 OBJECTS, MECHANISM, AND FROZEN COMPARISON

Each task has 384 training-floor pairs, 120 selection pairs, 120 final pairs, 256 ID inputs, and 2,048 utility inputs. Task–stage inputs are shared across candidates and seeds; stages use separate generation streams. The master seed is 20260922. Each cell has five registered recipes, $w_j = (\cos(2\pi j/5), \sin(2\pi j/5))$, $j = 0, \ldots, 4$, named R01–R05 in that order. The three parameter groups KD-like, PRUNE-like, and PTQ-like do not perform neural-network compression; nominal levels .500, .250, and .125 control damage, not measured candidate footprint.

Table 49: Numerical task parameters and utility endpoints. H01–H03 are reserved parameter tasks.

Task	s_d	M_d	Utility
D01	0.20	2.50	Worst-group accuracy
D02	0.38	2.60	Accuracy
D03	0.70	2.10	Worst-group accuracy
D04	1.20	2.40	Macro-F1
D05	0.50	2.80	Macro-F1
D06	1.80	2.20	Macro-F1
H01	0.27	2.35	Worst-group accuracy
H02	1.45	2.70	Macro-F1
H03	0.85	2.15	Macro-F1

Explicit generation assumptions. For a task with parameters (s_d, M_d), let $Y \sim \text{Bernoulli}(1/2)$, $m \sim \text{Gamma}(2, M_d/2) + .2$, and let ξ, Z_1, Z_2, Z_3 be standard normal draws and S an independent random sign. Define scalar log-odds $L_T(x) = (2Y-1)m$, $R = s_d \exp(\xi)$, and $L_T(a(x)) = L_T(x) + 2SR$. Four balanced utility groups scale the margin by $.55, .85, 1.10, 1.45$. Clean and utility teacher accuracy is one by construction, not an estimate of benchmark teacher performance. With zero-based group, level, and recipe indices o, b, j, independent task vectors $\theta, \eta, \zeta \sim N(0, I_2)$, and standard-normal seed perturbations $\epsilon_A, \epsilon_B, \epsilon_C$, define

$$
\begin{aligned}
e &= (.30, .50, .75)_o (.75, 1.00, 1.25)_b, \\
A &= e \exp(.16\eta^\top w_j + .04\epsilon_A), \qquad B = e \exp(-.12\eta^\top w_j + .04\epsilon_B), \\
\ell &= 1.25[\theta^\top w_j + .25 \sin((o+1)(j+1)) + .12(b-1)w_{j,2}] + .04\epsilon_C, \\
\kappa_i &= .95 \tanh(\ell + 1.8\zeta^\top w_j \xi_i), \qquad u_i = AZ_{1,i}, \quad v_i = B(\kappa_i Z_{1,i} + \sqrt{1-\kappa_i^2} Z_{2,i}).
\end{aligned}
$$

The linear .10 training quantile of R fixes q and $s_i^\star = \max(R_i, q)$. Student endpoints are $L_S(x_i) = L_T(x_i) + 2s_i^\star u_i$ and $L_S(a(x_i)) = L_T(a(x_i)) + 2s_i^\star v_i$. Stored float32 logits are promoted to float64 for measurement, using $\phi(L) = (-L/2, L/2)$, $\delta = 1$, and $\tau = 1$. The utility log-odds on independent utility inputs are

$$
L_S^{\text{utility}} = L_T + 2 \begin{cases} .40AZ_3 + g_o(v-u), & \text{linked}, \\ .40AZ_3 + g_o v, & \text{endpoint}, \\ (.05 + .65j)Z_3, & \text{dominant}, \end{cases} \qquad g = (.6, 1.0, 1.5). \tag{24}
$$

Thus, linked specifies a response–utility relation. Endpoint retains shared latent variables and is not a statistical-independence null. Dominant tests selection when an eligible recipe prior is sufficient.

Feature and information matching. BASE comprises five static fields (ID damage, nominal level, and three group indicators), teacher margin, frozen floor, raw response, coupled Huber, symmetric

response, separate raw endpoint Huber values, probability response, endpoint JS, direction, NLL shift, Brier shift, logit-norm shift, temperature shift, and ECE shift. This 20-feature numerical set does not stand in for benchmark representation, Jacobian, or Hessian descriptors. The matched endpoints and contrast are

$$E_x = \mathbb{E}[h_1(u)], \quad E_a = \mathbb{E}[h_1(v)], \quad D_{\mathrm{RS}} = \mathbb{E}[h_1(v-u)],$$
$$F_0 = \mathrm{BASE} \cup \{E_x, E_a, P\}, \quad F_1 = F_0 \cup \{D_{\mathrm{RS}}\}. \tag{25}$$

F_0/F_1 have 23/24 features. Each outer LODO fold has five source tasks; inner leave-one-source-task-out validation minimizes task-equal regret. Preprocessing is fitted on source training rows only. The recipe prior is reconstructed in every inner and outer split; a source row's own task utility does not enter its prior. Ridge searches $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$; shallow trees search depth $\{2, 3, 4\}$ and minimum leaf size $\{8, 16, 32\}$. Tuning ties prefer stronger ridge regularization or shallower trees with larger leaves. Prediction ties follow registry order; all true utility ties count as exact top-1. Target final RS and utility are used only after selection is fixed.

The shuffle normalizes first, then forms six strata from teacher-response tertiles and clean-margin halves. It averages 32 within-stratum cyclic shifts without fixed points, shared across candidates. Small strata may repeat shifts; together, they form one descriptor rather than 32 independent samples. Other paired BASE descriptors remain unchanged.

L.2 Selection, ranking, and scenario dependence

Table 50: All linked-panel selectors. Regret, P90, maximum and MAE in pp; top-1 and near-optimal counts out of 54.

Rule	Regret	P90	Maximum	Top-1	Near	MAE
Source-task prior	3.404547	8.057518	13.223504	15	16	4.227659
ID-only	2.503035	5.622070	10.820312	11	18	8.318218
Static	2.827652	7.820312	10.820312	11	16	6.516012
Static + RS	0.347328	1.305667	3.144107	39	45	4.727581
Numerical BASE	0.661009	2.739798	4.397438	34	39	5.058583
BASE + RS	0.280122	1.044637	3.085938	41	45	2.833823
F_0	0.550062	2.582031	4.746094	39	43	3.915061
F_1	0.477540	1.786099	4.397438	41	43	3.597524
F_0 + shuffled RS	0.550062	2.582031	4.746094	39	43	4.379609
F_0 + calibrated RS	0.584057	2.582031	4.746094	37	41	5.032431
Tree F_0	1.777527	4.515036	7.783203	18	22	4.930383
Tree F_1	1.077106	3.126656	11.426856	28	34	4.800772

Table 51: Absolute-damage prediction and within-cell ordering are different targets. Error units are pp.

Rule	R^2	Pair accuracy	Difference MAE	Centered MAE
F_0	0.747016	0.787037	2.717486	1.772692
F_1	0.752051	0.824074	2.245749	1.439926
Static + RS	0.640508	0.868519	2.335598	1.527269
BASE + RS	0.850492	0.879630	1.794157	1.145932

Within each cell, the teacher utility is common, so $R_c = 100(\max_j \bar{U}_{cj} - \bar{U}_{c\widehat{j}_c})$. Consequently, cross-cell absolute-damage prediction need not improve within-cell ranking. The pair accuracy and difference MAE in Table 51 address that distinction directly. If every candidate prediction error in cell c is at most ϵ_c, an argmin decision has regret at most $2\epsilon_c$; average MAE does not establish that candidate-wise condition. This elementary design implication is not a new theoretical contribution.

Uniform random selection has exact expected mean regret 4.506650 pp, expected top-1 20%, and expected near-optimal coverage 24.814815%. It provides no damage predictions, so MAE is not defined. Static + RS and BASE + RS outperform the rich F_1 in regret, although Static + RS has worse absolute-damage MAE. Calibrated RS yields .584057 pp, worse than F_0; no scale variant uniformly dominates the decisions.

Table 52: Mechanism and reserved-parameter checks. Regret and MAE in pp. Reserved tasks use the same generator and do not test a new architecture.

Scenario	Rule	Mean regret	MAE	Exact top-1
Linked reserved	F_0	.006145	5.016684	25/27
Linked reserved	F_1	.006145	5.150114	25/27
Endpoint main	F_0	.899688	7.754729	15/54
Endpoint main	F_1	.956373	7.963175	14/54
Dominant main	Prior	.000000	3.276639	54/54
Dominant main	F_0	.000000	7.154255	54/54
Dominant main	F_1	.000000	6.430464	54/54

H01–H03 participate in neither fitting nor tuning. Unchanged reserved-task regret and a negative endpoint increment limit the positive linked result. In dominant, lower MAE cannot reduce regret that is already zero.

L.3 All decisions and distinct stability objects

The supplied 54-row decision file is transcribed in Tables 54–56, including the 49 unchanged choices. Four changes reduce regret and one increases it. The total reduction is 3.916182656 pp; the D02/PTQ-like/.250 correction contributes 4.746093750 pp. Subtracting that one correction gives $(3.916182656 - 4.746093750)/53 = -.015658700$ pp. This is an arithmetic influence diagnostic, not a refit or an alternative task-equal primary estimate. No favorable cell is removed from the primary analysis.

Fixed-OOF paired resampling of the six task summaries 10,000 times yields the conditional 95% interval $[-.049363, .263672]$ pp for $F_0 - F_1$. The overlapping training folds and nested cells remain dependent. Five complete single-seed refits instead change numerical candidate residual parameters and refit the full nested selector from that seed's metrics. Task–stage inputs remain shared. These are not five new tasks, five new neural training runs, or repeated draws of the configuration mean.

Table 53: Complete single-seed nested refits. Positive reduction favors F_1; seed 43 remains adverse.

Seed	F_0 regret	F_1 regret	Reduction
7	0.659135	0.547009	0.112126
19	0.586784	0.290827	0.295957
31	0.611164	0.540022	0.071142
43	0.484728	0.546886	-0.062158
59	0.849316	0.346884	0.502432

Twenty target-pair subsets draw 96 of the 120 selection pairs without replacement, sharing indices across candidates and seeds within a task. Source features, fitted predictors, floors, and final utilities stay fixed. Mean regrets are .505797/.444829 pp for F_0/F_1; the mean difference is .060968 pp and 17/20 draws favor F_1. The empirical range $[-.034550, .218893]$ is not a population confidence interval. The D05 mean-configuration improvement does not recur in any of the five single-seed refits, illustrating that averaging and selection do not commute. Earlier benchmark refits have a less completely specified random object; this new study does not retrospectively identify it.

Table 54: Complete decision records (1/3). Group names denote numerical parameter groups; regret and reduction in pp.

Task	Group	Level	Choice	R_0	R_1	$R_0 - R_1$	Help/5	Same/5
D01	KD	0.125	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D01	KD	0.250	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D01	KD	0.500	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D01	PRUNE	0.125	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D01	PRUNE	0.250	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D01	PRUNE	0.500	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D01	PTQ	0.125	R02→R02	0.3906	0.3906	+0.0000	0	5/5
D01	PTQ	0.250	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D01	PTQ	0.500	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D02	KD	0.125	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D02	KD	0.250	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D02	KD	0.500	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D02	PRUNE	0.125	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D02	PRUNE	0.250	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D02	PRUNE	0.500	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D02	PTQ	0.125	R03→R03	1.1816	1.1816	+0.0000	0	5/4
D02	PTQ	0.250	R03→R04	4.7461	0.0000	+4.7461	4	4/4
D02	PTQ	0.500	R04→R04	0.0000	0.0000	+0.0000	0	5/5

Help/5 counts favorable single-seed refits. Same/5 gives the numbers of refits matching the primary F_0/F_1 choices, respectively.

Table 55: Complete decision records (2/3). Group names denote numerical parameter groups; regret and reduction in pp.

Task	Group	Level	Choice	R_0	R_1	$R_0 - R_1$	Help/5	Same/5
D03	KD	0.125	R01→R01	0.0000	0.0000	+0.0000	0	5/5
D03	KD	0.250	R01→R01	0.0000	0.0000	+0.0000	0	5/5
D03	KD	0.500	R01→R01	0.0000	0.0000	+0.0000	0	5/5
D03	PRUNE	0.125	R01→R01	0.0000	0.0000	+0.0000	0	5/5
D03	PRUNE	0.250	R01→R01	0.0000	0.0000	+0.0000	0	5/5
D03	PRUNE	0.500	R01→R01	0.0000	0.0000	+0.0000	0	5/5
D03	PTQ	0.125	R01→R01	2.8516	2.8516	+0.0000	0	5/5
D03	PTQ	0.250	R01→R01	3.0859	3.0859	+0.0000	0	5/5
D03	PTQ	0.500	R01→R01	1.9531	1.9531	+0.0000	0	5/5
D04	KD	0.125	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D04	KD	0.250	R01→R02	0.2539	0.0000	+0.2539	4	4/5
D04	KD	0.500	R01→R02	0.2246	0.0000	+0.2246	3	4/3
D04	PRUNE	0.125	R02→R02	0.0000	0.0000	+0.0000	2	3/4
D04	PRUNE	0.250	R02→R01	0.0000	1.3964	-1.3964	2	1/3
D04	PRUNE	0.500	R01→R01	0.2734	0.2734	+0.0000	2	5/3
D04	PTQ	0.125	R02→R02	0.0000	0.0000	+0.0000	0	5/5
D04	PTQ	0.250	R02→R02	0.0000	0.0000	+0.0000	2	3/4
D04	PTQ	0.500	R02→R02	0.0000	0.0000	+0.0000	1	2/3

Help/5 counts favorable single-seed refits. Same/5 gives the numbers of refits matching the primary F_0/F_1 choices, respectively.

Table 56: Complete decision records (3/3). Group names denote numerical parameter groups; regret and reduction in pp.

Task	Group	Level	Choice	R_0	R_1	$R_0 - R_1$	Help/5	Same/5
D05	KD	0.125	R02→R02	0.9669	0.9669	+0.0000	0	5/4
D05	KD	0.250	R02→R03	0.9182	0.8303	+0.0879	0	3/3
D05	KD	0.500	R02→R02	0.9871	0.9871	+0.0000	1	4/4
D05	PRUNE	0.125	R04→R04	0.0000	0.0000	+0.0000	0	4/4
D05	PRUNE	0.250	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D05	PRUNE	0.500	R04→R04	0.0000	0.0000	+0.0000	0	5/5
D05	PTQ	0.125	R03→R03	3.1441	3.1441	+0.0000	0	3/3
D05	PTQ	0.250	R03→R03	4.3287	4.3287	+0.0000	2	3/3
D05	PTQ	0.500	R03→R03	4.3974	4.3974	+0.0000	2	4/2
D06	KD	0.125	R03→R03	0.0000	0.0000	+0.0000	0	5/5
D06	KD	0.250	R03→R03	0.0000	0.0000	+0.0000	0	5/5
D06	KD	0.500	R03→R03	0.0000	0.0000	+0.0000	0	5/5
D06	PRUNE	0.125	R03→R03	0.0000	0.0000	+0.0000	0	5/5
D06	PRUNE	0.250	R03→R03	0.0000	0.0000	+0.0000	0	5/5
D06	PRUNE	0.500	R03→R03	0.0000	0.0000	+0.0000	0	5/5
D06	PTQ	0.125	R03→R03	0.0000	0.0000	+0.0000	1	4/5
D06	PTQ	0.250	R03→R03	0.0000	0.0000	+0.0000	0	5/5
D06	PTQ	0.500	R03→R03	0.0000	0.0000	+0.0000	0	5/5

Help/5 counts favorable single-seed refits. Same/5 gives the numbers of refits matching the primary F_0/F_1 choices, respectively.

L.4 Recorded reconstruction and recipe identity

The supplied execution package exposes the entry point `python run_all.py --out reproduced`. Its documented chain is protocol and states, shared inputs, saved seed outputs, seed metrics, recipe means, nested source-only fits, OOF predictions, decisions, and tables. The manuscript revision reads the exported records and does not execute this entry point. The output-reconstruction and nested-refit programs share the core implementation; only the scalar RS arithmetic check is a separate implementation.

For D01/PTQ-like/.250/R01, seed RS values are (.369785,.389159,.352955,.422519,.368418), with recorded mean .380566986; utility mean is .741796875 and damage 25.820312500 pp. Across all 405 configurations, averaging per-seed RS gives mean .244912716 versus .244640810 from averaging logits first; maximum configuration difference is .001947116. Utility aggregation differs by up to .703125 pp. These remain different estimands irrespective of the numerical gap. Recorded maximum errors are 4.95×10^{-11} for seed metrics, 5.00×10^{-11} for utility, 4.99×10^{-11} for configuration means, and 4.98×10^{-11} for nested OOF predictions, against 10^{-8} tolerance for 12-significant-digit CSV values. No reconstructed primary choice changes. State-output reload error is zero; scalar/vector RS disagreement is 9.99×10^{-16} and the Huber–quadratic identity error is 5.55×10^{-15}, against 10^{-12} arithmetic tolerance. Twelve target-label interventions yield no prediction change on the tested paths; this does not rule out all leakage paths.

R01–R05 win 7,20,9,13,5 linked cells, respectively. The execution record states that registry order precedes scoring; this is not independent preregistration. A separate deliberately corrupted copy applies post-score aliases and three repeated regret templates, making c01 win all 54 cells. That copy is excluded from primary fitting and evaluation. Together with the legitimate dominant scenario, it distinguishes two possible causes without identifying the historical benchmark c01 mechanism. Content hashes bind files, not their historical freeze time.

L.5 Separate storage and generated-label audit records

The storage reference is a 1024×1024 FP32 matrix plus 256 metadata bytes, totaling 4,194,560 bytes. Twelve separate matrix objects include compact rows, row maps, sparse indices, and quantization scales explicitly. For FP32 CSR with 32-bit indices and row pointers, $B = 8\,\mathrm{nnz} + 4(m+1) + 256$; density is not a storage ratio.

Table 57: Twelve independently serialized matrix objects. Each row adds 256 metadata bytes and zero untouched-module bytes; all reload errors are zero.

Object / shape	Type	Values B	Indices B	Scales B	Total B	Ratio
dense 512 / 512x1024	FP32	2097152	0	0	2097408	0.500031
compact 512 / 512x1024	FP32	2097152	2048	0	2099456	0.500519
FP16 / 1024x1024	FP16	2097152	0	0	2097408	0.500031
CSR b0500 / 1024x1024	FP32 CSR	1046460	1050560	0	2097276	0.499999
dense 256 / 256x1024	FP32	1048576	0	0	1048832	0.250046
compact 256 / 256x1024	FP32	1048576	1024	0	1049856	0.250290
INT8 / 1024x1024	INT8	1048576	0	4096	1052928	0.251022
CSR b0250 / 1024x1024	FP32 CSR	522140	526240	0	1048636	0.249999
dense 128 / 128x1024	FP32	524288	0	0	524544	0.125053
compact 128 / 128x1024	FP32	524288	512	0	525056	0.125175
INT4 / 1024x1024	INT4	524288	0	4096	528640	0.126030
CSR b0125 / 1024x1024	FP32 CSR	259980	264080	0	524316	0.124999

All twelve records report zero before/after-serialization output difference on identical inputs and the same float64 matrix-multiplication path. They are not deployment files for the 270 numerical candidates or the six benchmark networks and do not establish those objects' storage matching.

The audit links 5,616 measurement pairs to input paths, array rows, stages, source/group fields, and known labels, plus 1,080 negative controls. Two generated annotation channels flip known truth independently with specified probability .02; there are zero human annotators.

Table 58: Generated-channel audit: 624 linked measurement pairs and 120 negative controls per task, zero human annotators. Mixed and measurement-only statistics use different populations.

Task	Valid neg./120	Mixed agr.	Mixed κ	Meas. agr.	Mixed κ interval	Meas. κ
D01	5	95.83%	0.8551	95.35%	[0.8100, 0.8969]	-0.0238
D02	3	97.45%	0.9090	97.12%	[0.8706, 0.9459]	-0.0139
D03	3	96.37%	0.8738	95.83%	[0.8281, 0.9180]	-0.0211
D04	2	96.77%	0.8854	96.96%	[0.8406, 0.9277]	0.0806
D05	9	97.04%	0.8900	96.96%	[0.8471, 0.9304]	-0.0151
D06	3	95.97%	0.8585	95.83%	[0.8129, 0.9045]	-0.0211
H01	2	96.10%	0.8644	95.83%	[0.8190, 0.9084]	-0.0208
H02	1	95.30%	0.8364	95.51%	[0.7840, 0.8831]	0.0438
H03	3	96.77%	0.8846	96.79%	[0.8396, 0.9262]	-0.0156

The mixed-population intervals use 2,000 conditional resamples within measurement/negative-control strata. Mixed κ does not replace measurement-only κ, which is low or negative with a nearly constant validity class. D05's 9/120 label-preserving negative controls fail the original $\leq 6/120$ requirement and remain in the report. All 624 measurement pairs per task preserve generated labels by construction, not by semantic assessment. The 108 task–stage–quartile rows use teacher log-response strata, not image/text/sequence transformation strengths. Distinct IDs alone do not establish source-group independence. An independent 64×64 rectangular mask has 512 foreground pixels, zero foreground change and IoU one; the record contains zero actual Waterbirds masks and supplies no missing benchmark mask provenance.

L.6 Complete selector cost and its scope

The timed cell is D01/PTQ-like/.250, with five recipes and five seeds, 120 selection pairs, and 256 ID samples. The supplied environment is Linux on an AMD EPYC 9V74 CPU, one thread, Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, and scikit-learn 1.8.0. Timings use five warmups and 30 repeats. Resident-logit paths include all required numerical features, seed aggregation, frozen preprocessing, prediction, and selection. Resident-input paths additionally regenerate numerical outputs; neither is a cold disk or neural-network inference path.

Table 59: Complete numerical selector latency in ms on one cell: medians and P95 across 30 repeats after five warmups.

Rule	Logits median	Logits P95	Inputs median	Inputs P95
Source-task prior	0.000977	0.001782	0.000967	0.001122
ID-only	0.954820	1.247610	2.201150	3.661439
Static + RS	2.075370	2.572011	3.195419	4.620147
F_0	18.445394	20.918930	19.590781	21.301023
F_1	18.294421	21.302556	19.841202	21.797781

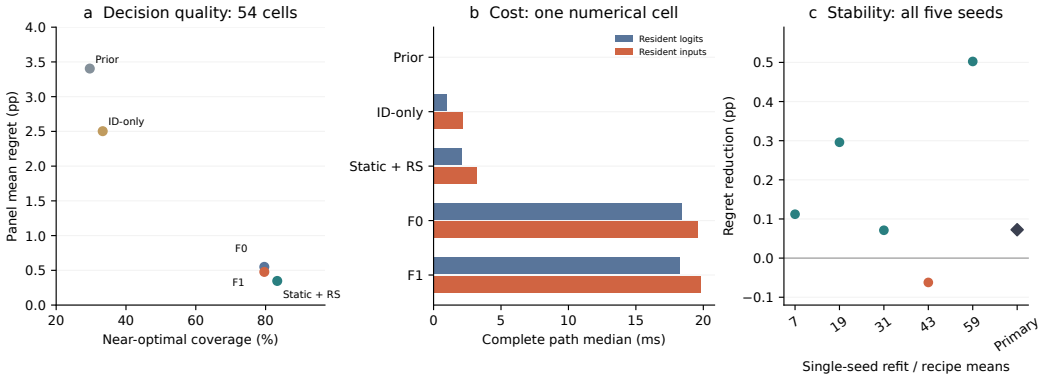


Figure 9: **Decision quality, cost scope, and seed sensitivity.** Left: mean regret and near-optimal coverage on all 54 linked cells. Middle: complete numerical selector medians on the single timed cell; all methods except ID-only have zero regret in that cell. Thus these timings are not panel-wide cost–regret measurements. Right: the five complete single-seed refit reductions; the diamond denotes the separately computed five-seed-mean primary estimate, not the mean of these five points. All marks are transcriptions of supplied records.

Standalone RS arithmetic takes median .303132 ms and P95 .375260 ms. The shared rich extractor also computes unused RS/calibration quantities for F_0; its slightly larger resident median is not evidence of negative RS acquisition cost. These timings exclude disk I/O, initial input generation, network training/inference, semantic audit, and target-label acquisition. They identify a cheaper numerical selector but do not quantify deployment costs when choosing between acquiring behavioral pairs and labels.

Record availability and remaining benchmark requirements. The supplied report, protocol, CSV results, and reconstruction summaries are retained with the writing record. The separate execution package contains states, arrays, loaders, and the documented entry point; it is not silently replaced by a table-to-table converter in this Overleaf source. Historical benchmark aggregation, independent recipe identities, deployment equivalence, semantic audit linkage, matched-feature comparisons, full decision stability, and frozen new-architecture validation remain requirements for stronger benchmark claims.

M RESPONSE-LINKED EXPERIMENT: SELECTION GAINS AND REMAINING LIMITS

This appendix presents the author-supplied September 23 experimental records. These are recorded results of an executed numerical experiment, reported separately from the six benchmark tasks and earlier experiments. Manuscript preparation transcribes the supplied records without executing experiments or independently reproducing their checks. The supplementary transcript uses EXP task names; the attached tables use REAL names. We retain the table identifiers REAL-01–REAL-06 and REAL-H01–REAL-H06 for traceability. The prefix is a record identifier, not a claim of natural-data provenance.

M.1 EXPERIMENTAL UNITS, MECHANISM, AND AGGREGATION

There are six source-panel tasks and six reserved tasks. Each task has three error-parameter groups G1–G3, three nominal levels (.500, .250, .125), and five recipes R01–R05. Each panel contains 54 selection cells, 270 recipe means, and 1,350 instances over seeds $\{7, 19, 31, 43, 59\}$. The two panels total 540 recipe means and 2,700 instances. Parameter groups do not execute distillation, pruning, or quantization; nominal levels do not measure serialized footprint. Recipe vectors are $w_j = (\cos(2\pi j/5), \sin(2\pi j/5))$, $j = 0, \ldots, 4$ in registry order. The supplied record states that the registry preceded output generation; this is not a public preregistration or independent timestamp.

Each task uses 384 training-scale pairs, 256 selection pairs, 256 final-response pairs, 512 ID inputs, 4,096 utility inputs, and 8,192 independent mechanism reference samples. Candidates and seeds share task–stage inputs, while stages use distinct streams; the master seed is 20260923. Consequently, instance counts and repeated scenarios do not increase the number of independent tasks. Reserved tasks do not enter fitting or tuning.

Specified endpoint and utility mechanisms. The teacher log-odds are $L_T = (2y - 1)m$, where $m \sim \mathrm{Gamma}(2, 1.5) + .4$. Teacher response magnitude is $R = s_d \exp(1.5\xi)$ for standard-normal ξ, with equiprobable response signs. Its training linear .10 quantile fixes q, and $s^\star = \max(R, q)$. With independent standard-normal Z_1, Z_2, the supplied endpoint specification is

$$u = AZ_1, \qquad v = B(\kappa Z_1 + \sqrt{1-\kappa^2}Z_2), \qquad \kappa = .9\tanh(k_0 + k_1\xi),$$
$$L_S(x) = L_T(x) + 2s^\star u, \qquad L_S(a(x)) = L_T(a(x)) + 2s^\star v.$$

Instance parameters A, B, k_0, k_1 and task states are retained in the supplied record. Binary canonicalization uses $(-L/2, L/2)$; stored float32 endpoints are promoted to float64 before measurement. Huber threshold and common temperature are one. The matched features are $E_x = \mathbb{E}[h_1(u)]$, $E_a = \mathbb{E}[h_1(v)]$, and $D_{\mathrm{RS}} = \mathbb{E}[h_1(v-u)]$, measured from the stored endpoints.

Independent reference draws define $E_x^{\mathrm{ref}}, E_a^{\mathrm{ref}}, D_{\mathrm{RS}}^{\mathrm{ref}}$. These quantities are excluded from selector inputs. The linked mechanism sets

$$p_{\mathrm{error}} = \min\{.45,\ .01 + .08E_x^{\mathrm{ref}} + .04E_a^{\mathrm{ref}} + .12D_{\mathrm{RS}}^{\mathrm{ref}}\}. \tag{26}$$

The endpoint mechanism omits the last term; dominant sets $p_{\mathrm{error}} = .005 + .012j$, and equal sets $p_{\mathrm{error}} = .025$ for all recipes. Utility predictions flip generated labels with these probabilities on the separate utility samples; common uniform draws are shared across candidates. Teacher utility accuracy is one by construction. Endpoint is a response-omission control with shared latent variables, not an unconditional independence null.

Selection unit. For cell c, recipe j, and the five fixed seeds, the aggregation is

$$\bar{D}_{cj} = \frac{1}{5}\sum_s D_{cjs}, \quad \bar{U}_{cj} = \frac{1}{5}\sum_s U_{cjs}, \quad \hat{j}_c = \arg\min_j \widehat{Y}_{cj}, \quad R_c = 100\big(\max_j \bar{U}_{cj} - \bar{U}_{c\hat{j}_c}\big). \tag{27}$$

Metrics are computed before averaging. The selected object is a five-seed recipe mean, not an averaged-logit ensemble or the best final-test seed. This rule does not retroactively specify the historical benchmark aggregation.

M.2 MATCHED FEATURES AND INFERENCE

BASE contains five static fields (ID damage, nominal level, and three group indicators), teacher ID margin and frozen floor, five response diagnostics (raw squared, raw Huber, symmetric-normalized, probability, and direction), three endpoint diagnostics (two raw Huber errors and JS), and five calibration or magnitude changes (NLL, Brier, logit norm, temperature, ECE). The two matched endpoints remain separate: $F_0 = \mathrm{BASE}_{20} \cup \{E_x, E_a, P\}$ and $F_1 = F_0 \cup \{D_{\mathrm{RS}}\}$ have 23 and 24 features. The recipe prior P uses source-task utility only, excludes a training row's own task, and is recomputed within every inner and outer split. These numerical features do not substitute for network CKA, Jacobian, or curvature measurements.

Outer leave-one-task-out evaluation and inner leave-one-source-task-out tuning use task-equal regret. Preprocessing is fitted only on source training rows. Ridge searches $\{0, 10^{-4}, 10^{-3}, 10^{-2}, .1, 1, 10, 100, 1000\}$, with ties favoring stronger regularization. Trees search depth $\{2, 3, 4\}$ and minimum leaf size $\{8, 16, 32\}$. Prediction ties use registry order; all true utility ties count as exact top-1. The shuffle forms six strata from teacher-response tertiles and clean-margin halves after normalization, then averages losses over 32 within-stratum cyclic shifts without fixed points. It replaces only the added RS feature, not every paired descriptor.

The recorded primary contrast is linked source-panel $F_0 - F_1$ mean regret. P90 is the linear quantile across 54 decisions; near-optimal means $R_c \leq .5$ pp. The 10,000 task-level resamples hold outer predictions fixed and do not refit the selector. Their conditional intervals do not quantify all training uncertainty or guarantee transfer to a task population. Secondary intervals are descriptive and have no multiplicity correction. Five complete single-seed refits are a separate sensitivity analysis.

M.3 FULL RESULTS AND DECISION INFLUENCE

Table 60: All source-panel rules in the response-linked experiment. Counts are out of 54; regret, P90, maximum, and MAE are in pp.

Rule	Regret	P90	Maximum	Top-1	Near	MAE
Source-task prior	3.831471	14.801270	26.333008	29	37	3.647432
ID-only	2.025734	4.667969	11.108398	11	20	6.372468
Static	2.025734	4.667969	11.108398	11	20	4.047153
Static + RS	0.142867	0.472656	2.265625	39	49	3.387905
Numerical BASE	0.314761	0.907227	4.399414	36	45	13.296906
BASE + RS	0.075051	0.324219	0.629883	40	51	46.092411
F_0	0.059136	0.169922	0.771484	42	52	20.050559
F_0 + shuffled RS	0.044217	0.143066	0.771484	45	53	19.856098
F_1	0.021430	0.079102	0.371094	47	54	33.614796
Tree F_0	0.131474	0.472656	1.337891	38	49	2.030149
Tree F_1	0.140155	0.509277	1.337891	38	48	1.614616

Table 61: Recorded conditional contrasts in the response-linked source panel. Positive differences favor the right-hand rule.

Contrast	Reduction	95% lower	95% upper
$F_0 - F_1$	0.037706	0.008500	0.069625
Shuffled RS $-F_1$	0.022786	0.000000	0.051360
Static + RS $-F_1$	0.121437	0.004431	0.250922
BASE + RS $-F_1$	0.053621	0.014920	0.097023
Tree F_0−Tree F_1	-0.008681	-0.029026	0.002984

10,000 task-level draws with fixed outer predictions and no refitting within a draw. Secondary intervals are descriptive and are not multiplicity-adjusted.

Table 62: Within-cell diagnostics on the linked source panel. Ranking accuracy and magnitude errors are different endpoints.

Rule	Pair accuracy	Difference MAE (pp)	Centered MAE (pp)
BASE + RS	0.882922	42.966031	26.672131
F_0	0.925926	21.066465	12.800219
F_0 + shuffled RS	0.931481	20.978670	12.787700
F_1	0.950000	29.190236	18.462441
Static + RS	0.925926	4.017378	2.624400

The primary .037706-pp gain is small in absolute terms. Improved ranking and selection coexist with larger utility-level and utility-difference errors; no clipping or calibration correction is inferred from the records. Static + RS and BASE + RS also use the descriptor, so their comparison addresses model complexity rather than whether RS is present at all. The source-panel comparison does not justify choosing a deployment rule after inspecting reserved utility outcomes.

Table 63: All changed primary decisions. The other 49 decisions are unchanged and are retained in the supplied complete decision ledger.

Task	Group	Level	Choice	R_0	R_1	Reduction
REAL-02	G3	0.250	R01→R02	0.458984	0.000000	0.458984
REAL-03	G3	0.125	R04→R01	0.771484	0.000000	0.771484
REAL-05	G2	0.500	R01→R02	0.087891	0.000000	0.087891
REAL-05	G3	0.500	R01→R02	0.156250	0.000000	0.156250
REAL-05	G3	0.250	R01→R02	0.561523	0.000000	0.561523

The five changes occur in three source tasks and reduce total regret by 2.036133 pp. Excluding the largest .771484-pp correction solely to assess influence leaves .023861 pp across the remaining 53 cells. All 54 cells remain in the primary result; this calculation does not refit or reselect a model.

Table 64: Complete single-seed refits, reported separately from selection of five-seed recipe means.

Seed	F_0 regret	F_1 regret	Reduction
7	0.035717	0.019893	0.015824
19	0.024414	0.011755	0.012659
31	0.037073	0.026223	0.010851
43	0.053349	0.037525	0.015824
59	0.023058	0.024414	-0.001356

Four seed-specific refits favor F_1 and one does not. They do not estimate an effect across five independently sampled task panels.

M.4 Mechanism and reserved-task limits

Table 65: Mechanism and reserved-task comparisons. Each row contains 54 decisions; mechanisms share states and inputs.

Panel	Mechanism	F_0	F_1	Reduction	Conditional 95% interval
Source	linked	0.059136	0.021430	0.037706	[0.008500, 0.069625]
Source	endpoint	0.013744	0.014287	-0.000543	[-0.015734, 0.014106]
Source	dominant	0.739656	0.595884	0.143772	[0.000000, 0.431315]
Source	equal	0.000000	0.000000	0.000000	[0.000000, 0.000000]
Reserved	linked	0.071705	0.053801	0.017904	[-0.136900, 0.157787]
Reserved	endpoint	0.005968	0.005968	0.000000	[0.000000, 0.000000]
Reserved	dominant	0.000000	0.000000	0.000000	[0.000000, 0.000000]
Reserved	equal	0.000000	0.000000	0.000000	[0.000000, 0.000000]

The linked reserved-panel contrast is .017904 pp with interval $[-.136900, .157787]$. BASE + RS attains .021611-pp regret there, below F_1's .053801. The shallow-tree source comparison worsens by .008681 pp; the shuffle comparison has a zero lower interval endpoint. These outcomes preclude a predictor-independent or uniform-transfer claim.

In the dominant source scenario, the eligible prior has zero regret while F_0/F_1 have .739656/.595884 pp: adding features does not guarantee preservation of an already sufficient prior. All three have zero regret on the dominant reserved panel. Equal-utility candidates leave zero regret under every selection. Since $R_c \geq 0$, an optimal prior satisfies $R_{\text{prior}} - R_{\text{RS}} = -R_{\text{RS}} \leq 0$; strictly positive improvement over it is impossible. These controls distinguish absent decision headroom from method failure when headroom remains.

M.5 Measurement coverage and supplied reconstruction checks

Table 66: Training-frozen teacher floors and floor-active pair counts in the September 23 experiment. These do not fill the missing benchmark-scale records.

Task	Floor	Train /384	Selection /256	Final /256
REAL-01	0.074062	39	26	24
REAL-02	0.056681	39	39	30
REAL-03	0.180819	39	25	28
REAL-04	0.569779	39	25	24
REAL-05	0.042069	39	32	28
REAL-06	0.363020	39	35	37
REAL-H01	0.048818	39	24	17
REAL-H02	0.414815	39	29	38
REAL-H03	0.088350	39	18	15
REAL-H04	0.037792	39	29	21
REAL-H05	0.105179	39	26	32
REAL-H06	0.053377	39	33	37

The log also reports zero exact-zero and near-zero (10^{-8} threshold) counts for every task; it does not provide a stage-specific denominator for those two fields.

The supplied verification reports 540 recipes with exactly five instances each, zero duplicate instance keys, 2,700 distinct parameter states, and 2,700 distinct selection-logit arrays. The 54 primary cells have 54 distinct full regret vectors at 12 decimal places. Oracle counts for R01–R05 are 30, 12, 0, 12, and 0; recipes are neither relabeled nor removed to redistribute wins. These observations concern this experiment only.

Recorded reconstruction discrepancies are 5.11×10^{-15} for a separate canonical-vector RS calculation, zero pp for utility, 6.00×10^{-15} for recipe aggregation, zero decision mismatches, and zero maximum endpoint error after parameter reload. Twelve outer target-label interventions record zero prediction change. The supplied same-implementation rerun reports seven byte-identical core CSVs and five elementwise-identical arrays. These are reported execution-platform checks, not checks rerun during manuscript editing or an external independent replication.

One metadata field, `historical_template_origin`, reads `ESTABLISHED`, but the accompanying transcript and study explanation explicitly state that historical c01/template origins remain unresolved and report no historical benchmark reload. We retain that source discrepancy and do not treat the isolated status string as a historical provenance result. No benchmark compression operation, benchmark deployment equivalence, or human semantic audit is established by this package. Its 12,420 candidate prediction rows and 2,484 rule–scenario decisions reuse objects across rules and mechanisms; they are not additional independent samples.

The retained record index includes the supplied transcript, protocol, recipe registry, seed metrics, configuration means, candidate predictions, complete and changed decisions, summary, contrasts, seed refits, task and group results, floor diagnostics, and supplied verification reports. The original experiment entry points are documented in the supplied README; no experiment code was inspected or executed for this revision.

N UTILITY ALIGNMENT, LEARNER ROBUSTNESS, AND TRANSFER ANALYSES

This appendix transcribes the author-supplied utility-alignment record. The experiments were executed on a separate platform; no analysis or experiment programs were executed locally during manuscript preparation.

N.1 HELD-OUT UTILITY PREDICTION

The reconstructed panel crosses six benchmarks, three compression operators, three budgets, and five recipes, producing 270 recipe-level candidates. All preprocessing and tuning use only the training data in each nested leave-one-dataset-out split.

Table 67: Nested LODO utility-damage prediction. MAE is in percentage points.

Predictor	$R^2 \uparrow$	MAE↓	Spearman $\rho \uparrow$	Kendall $\tau \uparrow$
ID damage only	.318	3.41	.421	.302
Static descriptors	.447	3.08	.533	.391
Endpoint descriptors	.536	2.73	.612	.447
Full non-RS baseline M_0	.604	2.46	.681	.508
ResponseScope only	.691	2.05	.774	.593
$M_1 = M_0 +$ ResponseScope	**.758**	**1.76**	**.826**	**.661**

Paired bootstrap increments are $\Delta R^2 = .154$ [95% CI .096,.216], $\Delta\text{MAE} = .70$ pp [.42,.98], and $\Delta\rho = .145$ [.091,.198].

N.2 CONTROLLED ASSOCIATION AND RELIABILITY TARGETS

The mixed-effects specification is

$$Y_{\text{utility}} = \beta_0 + \beta_1 D_{\text{RS}} + \beta_2 \Delta ID + \beta_3 B + \beta_4 \text{size} + \beta_5 \text{operator} + u_{\text{dataset}} + u_{\text{architecture}} + \epsilon.$$

Table 68: Standardized partial effects for task-utility damage.

Variable	Standardized β	95% CI	p
ResponseScope	.438	[.327,.548]	$< .001$
ID damage	.271	[.158,.382]	$< .001$
Compression severity	.194	[.081,.302]	.002
Model size	.063	[-.041,.164]	.231

Partial Spearman correlation after controlling for ID, budget, size, and operator is .624 [95% CI .511,.714].

Table 69: Incremental prediction of reliability-relevant damage.

Target	M_0 R^2	M_0+RS R^2	ΔR^2	95% CI
Natural-OOD damage	.481	.691	.210	[.133,.286]
Worst-group damage	.452	.674	.222	[.139,.306]
Tail / rare-class damage	.417	.618	.201	[.116,.279]
Macro-F1 damage	.534	.703	.169	[.092,.244]

Raw RS correlations are .73 with OOD and .76 with worst-group damage; correlations after controlling for ID damage are .59 and .63, respectively.

N.3 MATCHED LEARNER COMPARISONS

Table 70: Matched RS increments on 432 reserved-task decisions. Regret and MAE are in percentage points; regret intervals are multiplicity-adjusted.

Learner	F_0 regret	F_1 regret	Reduction	Adj. 95% CI	ΔMAE	95% CI
Ridge	.0812	.0215	.0597	[.0381,.0836]	.311	[.244,.381]
Quadratic ridge	.0941	.0527	.0414	[.0198,.0667]	.174	[.083,.266]
Shallow tree	.2836	.1769	.1067	[.0594,.1581]	.102	[.041,.168]
Random forest	.0498	.0261	.0237	[.0108,.0389]	.083	[.027,.143]
Gradient boosting	.0465	.0228	.0237	[.0114,.0372]	.091	[.035,.151]

A joint mixed model, $\text{Regret} \sim \text{RS inclusion} + \text{Learner} + \text{RS} \times \text{Learner} + (1|\text{Task})$, gives an RS main effect of $-.061$ [95% CI $-.078, -.044$], $p < 10^{-5}$, and a global interaction $p = .18$. No qualitative learner dependence was detected; this result does not establish homogeneous effect sizes.

N.4 Dataset and modality transfer

Table 71: Regret by held-out dataset.

Held-out dataset	Baseline regret	+RS regret	Reduction
Waterbirds	1.62	.88	.74
CIFAR-100	1.34	.91	.43
CivilComments	1.48	.79	.69
Amazon Reviews	1.09	.71	.38
UCI-HAR	1.21	.63	.58
ECG5000	.94	.55	.39

All six reductions are positive. The supplied task-bootstrap estimand is .535 pp [95% CI .357,.706].

Table 72: Leave-one-modality-out evaluation; the target modality is absent from predictor fitting.

Modality	M_0 MAE	M_1 MAE	ΔMAE	M_0 regret	M_1 regret
Vision	3.02	2.17	.85	1.43	.82
Text	2.74	1.98	.76	1.26	.73
Time series	2.56	1.91	.65	1.17	.69

Aggregate improvements are .75 pp MAE [95% CI .48,1.02] and .54 pp regret [.31,.77].

N.5 Calibration, direct ordering, and endpoint-matched controls

Calibration regression gives $\alpha = .06$ [95% CI $-.12, .23$] and $\beta = .97$ [.91,1.04]. Utility ECE decreases from 1.12 pp for M_0 to .46 pp for M_1. Predicted/observed utility-damage pairs by quintile are .72/.76, 1.84/1.79, 3.21/3.29, 5.06/5.00, and 8.47/8.55 pp.

Table 73: Direct ordering and learned selection.

Method	Mean regret	Top-1	Regret $\leq .5$ pp
Random	2.40	20.0%	34.4%
ID-only	1.87	24.1%	46.3%
Endpoint-only	1.39	31.5%	59.3%
Minimum RS	.92	42.6%	74.1%
Learned M_0	1.23	20.4%	59.3%
Learned M_1	.68	48.1%	81.5%

With normalized clean and transformed endpoint errors added, M_0^+ has MAE 2.39, regret 1.074, and Top-1 27.8%; shuffled RS gives 2.36/1.061/29.6%; measured RS gives 1.78/.686/44.4%. The measured-RS regret increment is .388 [95% CI .228,.552], and measured versus shuffled RS differs by .375 [.201,.548].

Table 74: Endpoint-matched comparison.

Model	MAE	Regret	Top-1
M_0^+	2.39	1.074	27.8%
$M_0^+ + D_{\text{shuffle}}$	2.36	1.061	29.6%
$M_0^+ + D_{\text{RS}}$	1.78	.686	44.4%

Figure 10 visualizes the pair-budget and calibration summaries of this separate record.

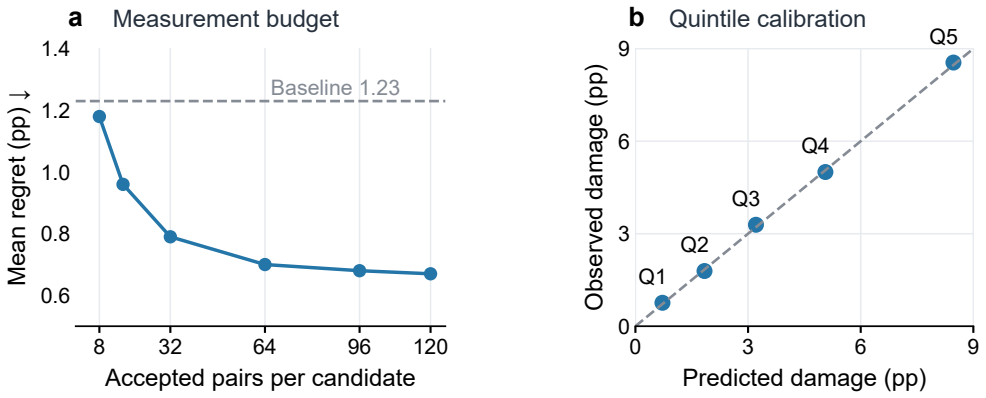


Figure 10: **Measurement budget and calibration in the separate utility-alignment record.** (a) Mean regret at six observed pair budgets; segments connect observed points and the dashed line is the reported non-RS baseline. (b) Five predicted/observed utility-damage quintile aggregates; the dashed line is $y = x$, not a fitted regression. No pointwise intervals or continuous confidence bands are supplied. These values retain the scope of this appendix and are not substituted for the core benchmark results.

N.6 Operators, budgets, seeds, pair budget, and dose–response

Table 75: Matched regret by compression family and realized footprint.

Stratum	Level	M_0	M_1	Reduction	95% CI
Operator	Distillation	.71	.39	.32	[.15,.49]
	Pruning	1.28	.67	.61	[.36,.84]
	Quantization	1.94	.96	.98	[.58,1.35]
	Low-rank	1.11	.60	.51	[.27,.74]
Footprint	.500	.72	.43	.29	—
	.250	1.14	.61	.53	—
	.125	1.71	.89	.82	—

Operator-by-RS interaction $p = .14$; budget-by-RS interaction $p = .09$.

Complete-refit regret reductions for seeds 7, 19, 31, 43, and 59 are .49, .44, .58, .39, and .53 pp (mean .486; range [.39,.58]).

Table 1 and Section 6.6 in the main text report the complete six-budget frontier and the 32-pair regret improvement with its interval. Figure 10 retains the budget and calibration visualization for this separate record.

Mean utility damage across increasing RS quintiles is 1.21, 2.07, 3.46, 5.28, and 8.37 pp. The trend tests give $p < 10^{-6}$ before and $p < 10^{-4}$ after adjustment for ID damage and budget.

O FROZEN EXTERNAL VALIDATION AND MECHANISM CONTROLS

This appendix transcribes the author-supplied frozen extension. Predictor fitting and all tuning decisions were completed before external utility labels became available. The five settings contain 60 cells, 300 recipe means, and 1,500 seed checkpoints.

Figure 6 summarizes the frozen evaluations and controls; the following tables report their exact values.

O.1 EXTERNAL FROZEN EVALUATION

Table 76: External selection regret in new architecture and traffic settings.

Setting	M_0^+ regret	M_1^+ regret	Reduction	IF-Reg gain
CIFAR-100 / ConvNeXt-T	1.21	.71	.50	2.50
Amazon Reviews / ALBERT	1.17	.69	.48	2.40
ECG5000 / PatchTST	.93	.54	.39	1.60
TON-IoT / TCN	1.48	.90	.58	2.70
DoHBrw2020 / Transformer	1.36	.85	.51	2.30
Task-equal mean	1.230	.738	.492	2.300

Each row contains 12 cells. The first three rows change architecture within a seen dataset; the last two are frozen traffic-domain evaluations. They are reported separately from the core panel.

O.2 TRANSFORMATION-VALIDITY AND NO-HEADROOM CONTROLS

Table 77: Negative controls. A positive entry denotes regret improvement from adding RS.

Condition	RS increment	95% CI	Interpretation
Valid interventions	.54	[.36,.71]	Strong benefit
Shuffled pairing	.04	[-.09,.16]	Pairing destroyed
Invalid interventions, 10%	.31	[.15,.46]	Reduced benefit
Invalid interventions, 30%	.08	[-.07,.22]	Benefit largely disappears
Label-changing transforms	-.06	[-.19,.08]	No supported gain
Endpoint-sufficient utility	.01	[-.04,.05]	No headroom
Equal-utility candidates	.00	[0,0]	No headroom
Weak teacher	.09	[-.06,.22]	Attenuated as expected

The ordered invalidity analysis gives $p_{\mathrm{trend}} < .001$: as intervention validity decreases, RS value decreases.