MEASURING UNEVEN CLASS DAMAGE FROM DATA PRUNING: FOUR PITFALLS AND A CONTROL LADDER

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based data pruning damages some classes more than others, and the standard way of measuring that damage cannot tell three explanations apart: the class imbalance a difficulty score induces, the choice of *which* examples survive inside each class, and the plain fact that a pruned model is less accurate. We document four measurement pitfalls that produce false positives here, each demonstrated on a conclusion of our own that we then withdrew: a disagreement statistic reported without its mechanical floor, which inflated a churn ratio to 2.08 when the floor accounts for essentially all of it; a baseline fitted through the very group under test; two non-pruning control families that disagree at matched accuracy; and selection-seed variance as large as training-seed variance, in a setting where 16.3% of a canonical Forgetting coreset is decided by score ties and only 94.3% of its indices are stable across seeds. Three of the four apply to any study that measures how much two models disagree. We then build a ladder of five increasingly strict controls, fixing budget, then per-class counts (Gini), then the hard/random mixing fraction q, then final accuracy, and run the same question up it. An aggregate hypothesis dies at the second rung: against a non-pruning control trained to the same accuracy, the difference is $+0.0005$, CI $[-0.023, +0.023]$. A class-level hypothesis survives all five on the setting where we found it, bracketed at $[-1.26, -0.71]$ by two comparisons that fail in opposite directions, and persists under a second architecture. It does not generalize: borderline on Food-101, null under real human label noise, and sign-reversed on Tiny-ImageNet with an interval that excludes zero. Subsampling CIFAR-100's test set to Tiny-ImageNet's per-class count leaves the effect intact in 99% of draws, so the reversal is a boundary of the mechanism and not small-sample noise. We report the ladder, the failures, and the pitfalls as the portable contribution.

1 INTRODUCTION

Data pruning discards most of a training set while retaining most of its accuracy (Toneva et al., 2019; Paul et al., 2021; Sorscher et al., 2022). Aggregate accuracy hides how the remaining loss is distributed. Vysogorets et al. (2025) conducted the first systematic study of this effect, showed that standard pruning algorithms produce measurably biased classifiers, and proposed choosing class-wise quotas to counteract it. Their analysis contains a second, qualitative observation: robustness improves not only from setting appropriate class ratios, but also from sampling randomly within a class rather than cherry-picking hard examples.

That second observation is where this paper starts. It implies two distinct mechanisms, an *imbalance channel* (how many examples of each class survive) and a *within-class channel* (which ones), and it has never been separated from a third possibility: that hard-selection simply yields a less accurate model, and less accurate models damage classes unevenly for reasons unrelated to selection. Distinguishing a within-class mechanism from an accuracy artifact requires a control that fixes per-class counts *and* final accuracy, varying only which samples inside each class are kept. No such control exists in the literature. The comparison used instead, purposive method against random pruning at the same budget, is confounded badly enough that in our hands it reverses its own conclusion.

What this paper is. It is an audit. Section 6 states four pitfalls in how class-level pruning damage is measured; Sections 4 and 5 run a case study that shows what each one costs. Two of our own

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.

hypotheses die in that case study and we report the autopsies, because a pitfall demonstrated on a retracted conclusion is more useful than one asserted in the abstract.

Contributions.

1. **Four measurement pitfalls**, each demonstrated on a conclusion we withdrew: a disagreement statistic without its mechanical floor, a baseline fitted through the group under test, control families that disagree at matched accuracy, and unmeasured selection-seed variance in a setting where score ties decide 16.3% of the coreset. Three of the four apply to any study that measures how much two models disagree.
2. **A control ladder** that avoids them: a Gini-matched stratified-random control, a mixture parameter q holding per-class counts fixed bit-for-bit, and an accuracy-matched control obtained by shortening the training schedule rather than by changing data.
3. **A scale-invariant per-class coefficient** $\tilde{\beta}$, with a demonstration that its raw counterpart is a monotone function of accuracy and must not be compared across configurations differing in accuracy.
4. **A case study reported in full**, including the aggregate hypothesis that dies at rung 2, an effect bracketed by two comparisons that fail in opposite directions rather than given as a point estimate, and five generalization checks of which one is null and one reverses sign.

2 Related Work

Difficulty scores and coreset selection. Toneva et al. (2019) count how often an example flips from correct to incorrect during training and keep the most-forgotten. Paul et al. (2021) score by the norm of the error vector under an early proxy model. Jiang et al. (2021) characterise structural regularity through a consistency score, and Sorscher et al. (2022) study how such scores interact with scaling. A recurring finding is that keeping only the extreme tail under-represents the typical examples that stabilise optimization, which motivates coverage-aware selection (Zheng et al., 2023), selection from the middle of the difficulty distribution (Xia et al., 2023), message passing that balances diversity against difficulty (Maharana et al., 2023), and dynamic schemes that re-select during training (Qin et al., 2024). Hu et al. (2021) document regimes in which loss-based prioritisation fails outright. We use Forgetting, EL2N and CCS as our purposive methods because they span three different signals: training dynamics, early gradient magnitude, and coverage of the score distribution.

Compression damages a narrow subset. The question we study has a direct precedent in model compression. Hooker et al. (2021) show that networks with very different sparsity levels have comparable top-line accuracy while diverging sharply on a narrow subset of the data, which they name Pruning Identified Exemplars, and that the damage concentrates on the underrepresented long tail. Hooker et al. (2020) extend this to protected attributes and show compression amplifies existing bias. Our object of study is data pruning rather than weight pruning, but the structure of the problem is the same: an aggregate metric that moves little while a subpopulation moves a lot. What we add is the observation that this comparison is usually made without a control at matched accuracy, and that when one is supplied the aggregate part of the effect does not survive.

Fairness of data pruning. Vysogorets et al. (2025) study data pruning's effect on classification bias and propose DRoP, which sets class-wise pruning quotas from held-out class error rates and prunes randomly within classes. Their experiments cross quota schemes with within-class strategies, which makes DRoP the closest prior work to ours and the source of the qualitative claim we test. Our contribution is not the observation that within-class strategy matters. It is showing that the observation survives when accuracy is held fixed, quantifying it, and mapping where it fails.

Disagreement, churn, and multiplicity. That two models of equal accuracy can disagree substantially is an old observation. Breiman (2001) named it the Rashomon effect; Marx et al. (2020) formalised it as predictive multiplicity and argued that its severity should be measured and reported rather than assumed away; D'Amour et al. (2022) traced it to underspecification of the training pipeline. In applied settings the same quantity appears as prediction churn between successive model versions (Milani Fard et al., 2016). This literature supplies the vocabulary for our aggregate

branch, but it typically reports a disagreement rate without decomposing it into the portion mechanically implied by any accuracy difference between the two models and a residual attributable to genuine functional divergence. Section 3 makes that decomposition explicit, and Section 6 reports what happened to our own conclusions before we did.

3 MEASUREMENT

Disagreement and its mechanical floor. For models f, g on test set T, let $\delta(f,g) = |T|^{-1}\sum_x \mathbb{1}[\arg\max f(x) \neq \arg\max g(x)]$. If f and g have accuracies a_1, a_2 and make otherwise-independent errors over C classes,

$$\delta_{\text{indep}}(a_1, a_2) = 1 - \left[a_1 a_2 + \tfrac{(1-a_1)(1-a_2)}{C-1}\right],$$

strictly positive and growing as either accuracy falls. Any raw disagreement number must be read against this floor, so we report $\rho(f,g) = 1 - \delta(f,g)/\delta_{\text{indep}}(a_1, a_2)$, where $\rho = 0$ means the two models disagree exactly as much as independence predicts at those accuracies.

This matters immediately. In our setting (C=100, $a \approx 0.775$) the floor is $\delta_{\text{indep}} = 0.399$, so the $\delta_{\text{base}} = 0.189$ we measure between two full-data models differing only in seed corresponds to $\rho_{\text{base}} = 0.53$. Read raw, 18.9% seed-to-seed disagreement looks anomalous against the 5–9% commonly quoted; that figure comes from settings with ten classes and 95% accuracy, where the floor itself is an order of magnitude smaller. The apparent anomaly is entirely an artifact of not dividing by the floor.

Per-class coefficients. For a coreset S, let survive_c be the fraction of class c retained and $\text{NFR}_c = \Pr[f_S(x) \neq y \wedge f_D(x) = y \mid y = c]$, the rate at which the coreset model breaks predictions the full-data model gets right. The raw coefficient β regresses NFR_c on survive_c across classes, controlling for $\delta_{\text{base},c}$, the class's intrinsic seed-to-seed disagreement. Raw β is not comparable across configurations: a uniform multiplicative increase in NFR — which lower accuracy produces — steepens it with no class-selective mechanism at all. We therefore use

$$\log(\text{NFR}_c + \varepsilon) \;\sim\; \tilde{\beta} \cdot \text{survive}_c + \gamma \log(\delta_{\text{base},c} + \varepsilon) + \text{const}, \qquad \varepsilon = 10^{-4},$$

a semi-elasticity invariant to any uniform rescaling of NFR. No class in any dataset we study has $\text{NFR}_c = 0$ (Appendix A), so the regression is nowhere driven by ε-clipped outliers.

Controlling for $\delta_{\text{base},c}$ is not bookkeeping. The raw Spearman correlation between survive-rate and per-class disagreement is $+0.060$ (n.s.); as a partial correlation, with intrinsic class difficulty removed, it is -0.665. A placebo applying the identical construction to Random-0.3 gives -0.185, confirming the partial correlation is a property of purposive selection and not of the estimator.

Two seeds, not one. We separate `trainseed` (initialization, batch order, augmentation) from `selseed` (which samples fill non-hard slots, and which side of a tied score is kept). All intervals are hierarchical bootstraps resampling selseed, trainseed within selseed, and test index jointly.

3.1 EXPERIMENTAL SETUP

Holding gradient steps rather than epochs fixed matters for every comparison in the paper. Under a fixed epoch count a 30% coreset receives 30% of the updates, and any measured difference confounds the effect of the data with the effect of undertraining.

4 THE CONTROL LADDER

Rung 3 is the key construction. For a reference coreset we build a *stratified-random* control with the identical per-class survive-count vector, drawing within each class uniformly at random instead of by score. We generalize it to a continuum: for class c with target count k_c, $\text{round}(q\,k_c)$ samples are the highest-scoring in the class and the remaining $k_c - \text{round}(q\,k_c)$ are drawn at random from what is left. Per-class counts, and therefore Gini, are identical for every q, verified bit-for-bit at build time; q=0 is the stratified-random control and q=1 is the original coreset.

Table 1: Training configuration. Every configuration in this paper is trained from scratch; no checkpoint is reused across configurations, and the accuracy-matched control differs from its reference only in schedule length.

Architecture	ResNet-18 (CIFAR variant: 3×3 stem, no max-pool); VGG-16-BN for the ablation
Optimizer	SGD, momentum 0.9, weight decay 5×10^{-4}
Schedule	LR 0.1, cosine annealed to 0 at the horizon T
Batch size	128
Augmentation	random crop (32, pad 4) and horizontal flip; no mixup or label smoothing
Horizon T	$40{,}000$ steps (CIFAR-100, CIFAR-100-N, Food-101); $24{,}000$ (Tiny-ImageNet); $8{,}700$ (CIFAR-100-LT); shortened T used only for accuracy-matched controls
Budget control	steps held fixed across budgets, so every coreset receives the same number of gradient updates as the full-data reference
Seeds	3 training seeds per configuration; 2 or 3 selection seeds where stated
Hardware	single RTX 4090; $\approx$550 s per $40{,}000$-step run

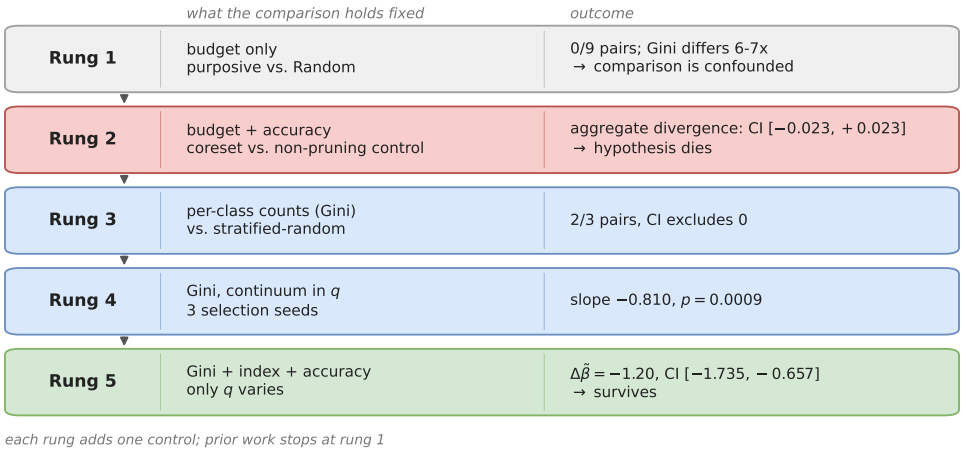

Figure 1: The control ladder. Each rung fixes one more nuisance variable. The aggregate hypothesis dies at rung 2; the within-class hypothesis survives to rung 5 on the discovery setting. Rung 1 is the comparison used in prior work.

Rung 5 additionally matches accuracy, and does so *without touching the data*: the $q{=}0$ coreset is trained with a shorter cosine-annealed schedule until its accuracy equals $q{=}1$'s. Candidate anchors are screened by ρ_{self}, the floor-adjusted agreement among the control's own seeds, which detects checkpoints that have not converged. The screening window is not load-bearing: three windows spanning $[0.40, 0.65]$ admit the same candidates and select the same anchor.

5 Results

5.1 The imbalance channel, and why rung 1 fails

On a 14-configuration grid (Forgetting/EL2N/CCS at budgets $\{0.3, 0.5, 0.7\}$, Random at five budgets), the standard comparison — purposive method against Random at the same budget — fails in *all nine* pairs, and at budget 0.7 all three purposive methods show *smaller* $|\beta|$ than Random. Budget is the wrong thing to match. At budget 0.3, Random has per-class Gini 0.037 against 0.209 for Forgetting and 0.270 for EL2N: a factor of six to seven. Matching budget compares two configurations that differ in the very quantity the imbalance channel operates through.

Figure 6 in Appendix A plots the whole grid on the (Gini, β) plane. A log-linear fit $\log|\beta| \sim \gamma \log(1-b) + \eta \log(\text{Gini})$ over the grid gives $\gamma \approx 2.7$–3.0 and $\eta \approx 1.8$ (CI $[0.88, 2.77]$). We report these as a trend, not a scaling law: they rest on twelve triples over a narrow Gini range.

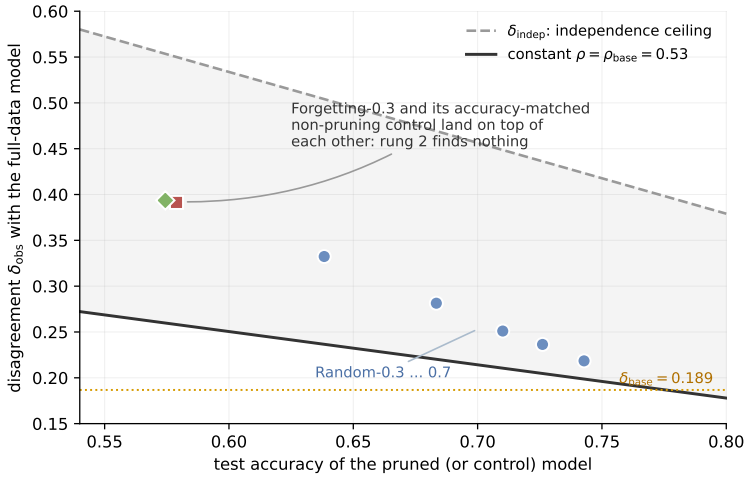

Figure 2: Pitfall 1 in one picture. Raw disagreement with the full-data model rises as the second model's accuracy falls, because the independence ceiling δ_{indep} rises with it. Forgetting-0.3 and a non-pruning control trained to the same accuracy land on top of each other, so the raw ratio $R = 2.08$ measures the accuracy gap and not the pruning.

Practical consequence. CCS holds Gini near 0.06 at every budget and shows no detectable class-level effect beyond it. Where a coverage objective is available it removes most of the imbalance channel at no cost; where it is not, a stratified re-balancing pass after selection is a cheap substitute for this channel alone.

5.2 Rung 2: The aggregate hypothesis dies

We first asked whether pruning drives predictions away from a full-data model beyond its accuracy cost. The raw churn ratio $R = \delta_{\text{prune}}/\delta_{\text{base}} = 2.08$ suggested it does. It does not survive; Figure 2 shows why. Against a true non-pruning control — full data, cosine-annealed to convergence at a shorter horizon, trained to 57.44% to match Forgetting-0.3's 57.89% — the difference in ρ_{cross} is $+0.0005$, CI $[-0.023, +0.023]$. At the accuracy that matters, a model trained on 30% of the data is indistinguishable, in floor-adjusted agreement with a converged full-data model, from a full-data model merely trained less. **We report this as a negative result and do not recover it anywhere later.**

5.3 Rungs 3–5: The within-class channel

At fixed per-class counts (rung 3), Forgetting-0.3 and EL2N-0.3 both damage classes more than their stratified-random controls: in raw units, $\Delta\beta = -0.263$, CI $[-0.359, -0.171]$ and -0.583, CI $[-0.733, -0.447]$. Forgetting-0.5 does not (-0.034, CI $[-0.097, 0.030]$), consistent with the mechanism concentrating at aggressive budgets. These three comparisons involve configurations of different accuracy, so by our own argument raw β is the wrong unit for them; in elastic units the same Forgetting-0.3 comparison gives $\Delta\tilde{\beta} = -0.71$, which is the schedule-matched end of the bracket reported below. We keep the raw numbers only because they are what the corresponding grid was originally computed in.

Rung 4 traces the continuum. Table 2 gives $\tilde{\beta}$ at each q for every selection seed. Two things are visible at once: $\tilde{\beta}$ falls monotonically in q at fixed Gini (selseed-averaged slope -0.810, SE 0.147, $p = 0.0009$), and the spread across selection seeds at a fixed q is of the same order as the effect of moving one step in q. The second observation is pitfall 4 of Section 6 in tabular form.

Variance decomposition on this table gives selection-seed variance 0.034 against training-seed variance 0.035. A single-selseed design, which is the norm in this literature, therefore understates the standard error of any $\tilde{\beta}$ estimate by roughly $\sqrt{2}$.

Table 2: Rung 4: $\tilde{\beta}$ across the q continuum, by selection seed, at T=40,000. Per-class counts and therefore Gini (0.2086) are bit-for-bit identical in every cell; only which samples fill each class differs. Each cell averages three training seeds.

q	selseed 1	selseed 2	selseed 3	mean
0.0 (stratified-random)	−2.736	−2.281	−2.663	−2.560
0.5 (mixture)	−2.832	−2.788	−2.936	−2.852
1.0 (Forgetting-0.3)	−3.460	−3.154	−3.496	−3.370

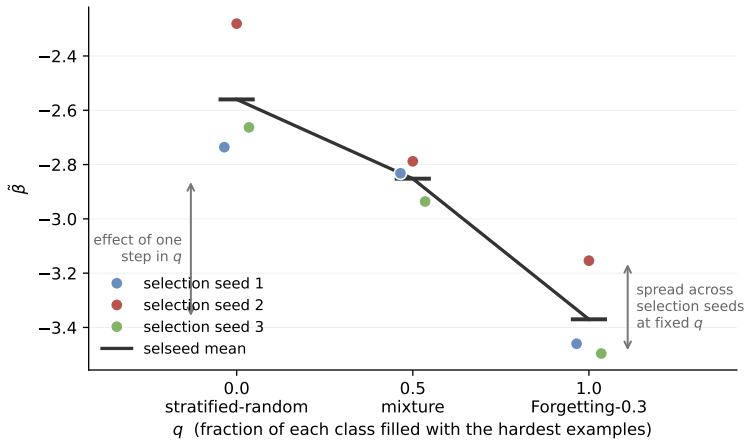

Figure 3: Rung 4 and pitfall 4 in one picture. Every point has bit-for-bit identical per-class counts; only which samples fill each class differs. The trend in q is monotone, and the spread across selection seeds at a fixed q is of the same order as the effect of moving one step in q.

At rung 5, $\tilde{\beta}(q{=}1) = -3.460$ against $\tilde{\beta}(q{=}0,\ T{=}9000) = -2.208$ at matched accuracy (57.89% vs. 58.63%), giving $\Delta\tilde{\beta} = -1.20$, CI $[-1.735, -0.657]$.

Bracketing rather than a point estimate. Rung 5 matches accuracy at the cost of un-matching the training schedule. The reverse comparison is also available: at the *same* schedule (T=40,000) but unmatched accuracy, the gap is −0.71. These two comparisons fail in opposite directions, and both exclude zero with the same sign. We therefore report the within-class effect as bracketed by $[-1.26, -0.71]$ rather than as the accuracy-matched number alone, which we consider the honest summary and which closes both lines of objection simultaneously.

Close to parallel, over a short overlap. Tracing $\tilde{\beta}$ against accuracy at four training horizons for each of q=0 and q=1 (Figure 4; the underlying numbers are in Table 4), the vertical gap at matched accuracy is −1.162 at the low end of the overlap and −1.252 at the high end. We state the limitation plainly: the two accuracy ranges overlap only over $[0.561, 0.579]$, so "parallel" is established across 1.8 accuracy points, not the eleven the two series jointly span. The fitted slopes differ (−8.8 vs. −13.7) and the q=0 series is non-monotone in its first three points. We do not claim the slopes are equal; we claim the vertical gap is stable where we can measure it.

A standard fairness metric agrees, after the accuracy confound is removed. Compared at the accuracy-matched anchor rather than at full budget — 0.74 points apart instead of 5 — Forgetting-0.3 still has lower worst-class accuracy (0.283 vs. 0.317) and a larger spread of per-class accuracy gap (std 0.105 vs. 0.077) than its random control. The effect is therefore not an artifact of the $\tilde{\beta}$ construction, and not an artifact of the accuracy difference that separated the two groups in the naive comparison.

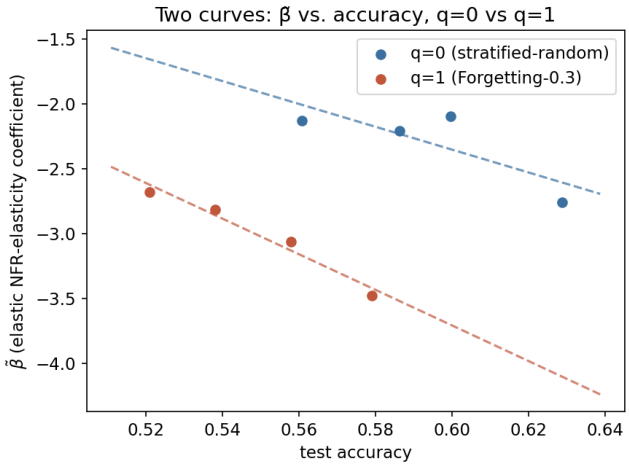

Figure 4: $\tilde{\beta}$ against accuracy for $q{=}0$ (stratified-random) and $q{=}1$ (Forgetting-0.3), each traced at four training horizons. The vertical gap is stable over the shared accuracy range, but that range spans only 1.8 accuracy points; the fitted slopes are not claimed to be equal.

Table 3: Generalization of the rung-5 comparison. Positive $\Delta\tilde{\beta}$ is the direction opposite to the within-class hypothesis.

Setting	Change from discovery	$\Delta\tilde{\beta}$ (95% CI)	Verdict
CIFAR-100 / ResNet-18	—	-1.20 $[-1.735, -0.657]$	discovery setting
CIFAR-100 / VGG-16	architecture	-0.726 $[-1.181, -0.270]$	replicates
Food-101	dataset	-0.233 $[-0.455, -0.0003]$	borderline
CIFAR-100-N	real label noise	$+0.215$ $[-0.188, 0.601]$	null
Tiny-ImageNet	dataset, 200 classes	$+0.573$ $[0.299, 0.874]$	**reverses**

5.4 Generalization: the failures are the informative part

We ran the rung-3/rung-5 grid on three further datasets and one further architecture (Table 3).

Three points deserve emphasis rather than burial.

First, Food-101's interval excludes zero by 3×10^{-4}, and one of its two selection seeds is null on its own. We treat it as borderline, not as a replication, and note that it would not survive any correction for the roughly ten intervals reported in this paper.

Second, CIFAR-100-N uses the real human annotations of Wei et al. (2022), which disagree with the clean labels on 40.2% of training samples, and the effect vanishes. A forgetting event under noisy labels conflates difficulty with label corruption, so the score no longer ranks what it is assumed to rank. This is a concrete, mechanistically plausible boundary condition.

Third, Tiny-ImageNet does not merely fail — it reverses, with an interval excluding zero, replicated across both selection seeds. Tiny-ImageNet provides 50 test images per class against CIFAR-100's 100 and Food-101's 250, so its per-class NFR estimates carry roughly $\sqrt{2}$ times the binomial standard error, and small-sample instability is the obvious explanation. We tested it and it is wrong. Subsampling the CIFAR-100 test set to exactly 50 images per class and repeating the comparison over 100 bootstrap draws leaves the estimate essentially unchanged (-0.677 against -0.712 on the full test set, CI $[-1.136, -0.160]$, with 99% of draws preserving the full-set sign). Halving the per-class test count does not destabilize the effect, so Tiny-ImageNet's reversal is a property of that setting, not of its test-set size. We report it as an open question rather than explain it away; the 200-class label space and the resulting change in what "hard" selects are the candidates we cannot yet distinguish.

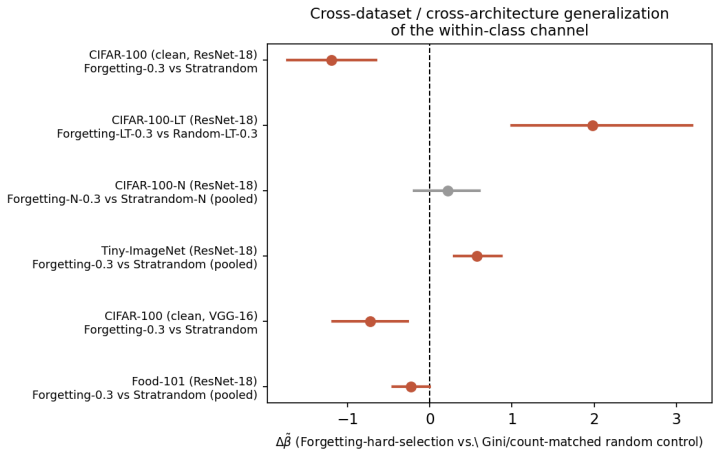

Figure 5: Cross-dataset and cross-architecture $\Delta\tilde{\beta}$ with hierarchical-bootstrap intervals. Intervals excluding zero are highlighted. The within-class channel replicates under an architecture change, is borderline on Food-101, null under real label noise, and reverses on Tiny-ImageNet.

What we do not claim. We do not count CIFAR-100-LT as a replication. Our long-tailed experiment compares Forgetting-LT-0.3 against Random-LT-0.3 at matched *budget*, which is rung 1 — exactly the comparison this section shows to be confounded by Gini. It is reported in Appendix B for completeness and excluded from Table 3.

6 Four Measurement Pitfalls

Each pitfall below produced a conclusion we held and later withdrew.

(1) A churn ratio with no mechanical floor. We computed $R = 2.08$ and read it as a pruning-specific effect. It is not: at C=100 and a 19.5-point accuracy gap, δ_{indep} alone accounts for essentially all of it. Any disagreement statistic reported without a floor correction, or without a matched-accuracy non-pruning control, cannot distinguish a method effect from an accuracy effect.

(2) A baseline fitted through the group under test. We next fitted ρ against accuracy through five Random configurations and measured residuals from that line. Random is itself a pruning method, so its own residuals are zero by construction, and the apparent excess of purposive methods was an artifact of extrapolating a fit from a narrow Gini range into a much wider one. A baseline must come from outside the family being tested.

(3) Control families that disagree at matched accuracy. We built two non-pruning controls at the same accuracy — shorter schedule, and narrower width. Two of three width checkpoints failed ρ_{self} screening, and the surviving one disagreed with the schedule control at the same nominal accuracy. Matched accuracy does not imply matched everything else that a disagreement statistic responds to, and a control family needs its own quality control.

(4) A coreset is not a well-defined object. Forgetting scores are low-cardinality integers. In Forgetting-0.3 the 30% cutoff lands on a tied score covering 3,771 samples of which only 2,439 enter the coreset, so 16.3% of the coreset is decided by tie-breaking; across selection seeds only 94.3% of indices are stable. Selection-seed variance in $\tilde{\beta}$ (≈ 0.034) is as large as training-seed variance (≈ 0.035). A single-selseed estimate understates its standard error by roughly $\sqrt{2}$, and produced a spurious non-monotonicity in our own q sweep that vanished under replication. A paper reporting results for "the" Forgetting coreset is reporting one draw from a distribution.

7 LIMITATIONS

The within-class effect is established on one dataset and confirmed under one architecture change. Of three new datasets, one is borderline, one null, and one reverses; we therefore do not claim it is a property of hard-sample selection in general. The boundary conditions we can name — label noise, and the 200-class regime — rest on one check each, and for the second we have ruled out the obvious explanation without establishing an alternative. Our intervals are per-comparison at 95% with no multiplicity correction across roughly ten reported; Food-101 in particular should not be read as significant under any correction. The exponents γ, η rest on twelve points over a narrow Gini range. $\tilde{\beta}$ is one of several possible scale-invariant normalizations and we compare it only against the raw β it is designed to replace. All experiments use ResNet-18 or VGG-16 at 32×32 or 64×64; nothing here speaks to ImageNet-scale training or to transformers.

8 CONCLUSION

Class-level damage from score-based pruning has at least two sources, and the comparison used to study them in practice cannot separate either from the other or from accuracy. We built a ladder that can. An aggregate divergence hypothesis dies on it. A within-class hypothesis survives on the setting where we found it, with an effect bracketed by two comparisons that fail in opposite directions, and then fails to generalize on two of three new datasets, reversing sign on one for reasons we can show are not a sample-size artifact. We consider the ladder, and the four pitfalls it exposed in our own work, the more portable contribution: three of the four apply verbatim to any study that measures how much two models disagree.

REPRODUCIBILITY STATEMENT

Every quantity is computed from stored test-set logits of trained models. The supplementary material lists each run, its selection and training seeds, and the JSON file from which each number in this paper is read. Coreset indices are released with the code so that the tie-breaking distribution of Section 6 can be reproduced directly.

DISCLOSURE OF AI ASSISTANCE

We used a large language model to check and polish the writing of this paper, and to assist with code, specifically the plotting and figure-generation scripts. All experiments were run on our own hardware, and every number reported here is computed from stored test-set logits of models we trained and was verified by the authors against the underlying data files.

REFERENCES

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, August 2001. doi: 10.1214/ss/1009213726.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models, 2020. URL https://arxiv.org/abs/2010.03058.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget?, 2021. URL https://arxiv.org/abs/1911.05248.

Niel Teng Hu, Xinyu Hu, Rosanne Liu, Sara Hooker, and Jason Yosinski. When does loss-based prioritization fail?, 2021. URL https://arxiv.org/abs/2107.07741.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael Curtis Mozer. Characterizing structural regularities of labeled data in overparameterized models, 2021. URL https://openreview.net/forum?id=3GYfIYvNNhL.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning, 2023. URL https://arxiv.org/abs/2310.07931.

Charles T. Marx, Flavio Du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/dc5c768b5dc76a084531934b34601977-Paper.pdf.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Uj7pF-D-YvT.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, xu Zhao Pan, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. Infobatch: Lossless training speed up by unbiased dynamic data pruning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=C61sk5LsK6.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Artem M Vysogorets, Kartik Ahuja, and Julia Kempe. DRop: Distributionally robust data pruning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fxv0FfmDAg.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations, 2022. URL https://arxiv.org/abs/2110.12088.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=7D5EECbOaf9.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QwKvL6wC8Yi.

A ROBUSTNESS CHECKS

Test-set subsampling. To test whether Tiny-ImageNet's sign reversal is small-sample instability, we reduced the CIFAR-100 test set to 50 images per class, matching Tiny-ImageNet exactly, and recomputed the rung-5 comparison over 100 independent bootstrap draws. Full test set (100/class): $\Delta\tilde{\beta} = -0.712$. Subsampled (50/class): bootstrap mean -0.677, CI $[-1.136, -0.160]$, excluding zero, with 99% of draws preserving the full-set sign. Halving the per-class test count does not destabilize the effect.

Zero-valued NFR. The elastic specification adds $\varepsilon = 10^{-4}$ before taking logs; a class with $\mathrm{NFR}_c = 0$ would contribute $\log \varepsilon \approx -9.2$ and could dominate the fit. No class produced $\mathrm{NFR}_c = 0$ in any dataset: 0 of 100 on CIFAR-100 (full test set and in every subsample draw), 0 of 200 on Tiny-ImageNet, 0 of 101 on Food-101.

Fairness metrics at the accuracy-matched anchor. Comparing Forgetting-0.3 against the stratified-random control at its accuracy-matched anchor (T=9000, accuracies 0.5789 vs. 0.5863) rather than at full budget (0.5789 vs. 0.6287): worst-class accuracy 0.2833 vs. 0.3167, mean per-class gap 0.1967 vs. 0.1893, std of per-class gap 0.1051 vs. 0.0771. The direction is unchanged once the five-point accuracy confound is removed.

Table 4: $\tilde{\beta}$ against accuracy at four training horizons, three trainseeds each. The accuracy ranges overlap only over $[0.561, 0.579]$.

T (steps)	acc (q=0)	$\tilde{\beta}$ (q=0)	acc (q=1)	$\tilde{\beta}$ (q=1)
6,000	0.5608	−2.127	0.5209	−2.676
9,000	0.5863	−2.208	0.5382	−2.815
12,000	0.5995	−2.095	0.5579	−3.057
40,000	0.6287	−2.755	0.5790	−3.475

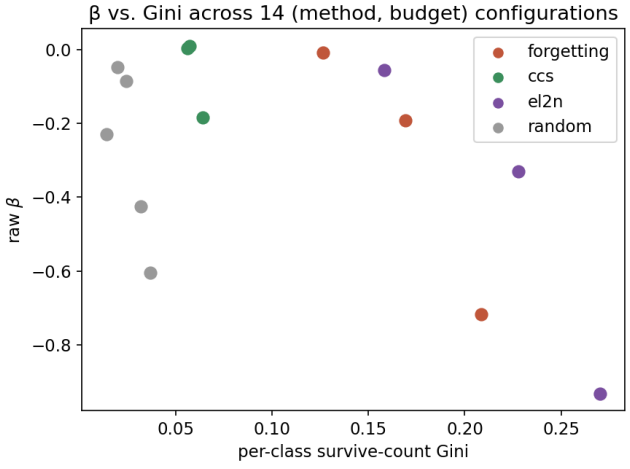

Figure 6: Raw β against per-class survive-count Gini across the 14-configuration grid. CCS sits at low Gini at every budget; Forgetting and EL2N reach much higher Gini at aggressive budgets. Matching budget, as prior work does, compares points far apart on this axis.

QC-window ablation. ρ_{self} at the three candidate anchors is 0.4848 (T=6000), 0.5010 (T=9000), 0.5116 (T=12000). Windows $[0.42, 0.58]$, $[0.45, 0.60]$ and $[0.40, 0.65]$ admit the same three candidates and select the same anchor; the decisive interval is unchanged under all three.

B LONG-TAILED CIFAR-100 (RUNG-1 COMPARISON, EXCLUDED FROM THE MAIN CLAIM)

We built a long-tailed CIFAR-100 (imbalance factor 100, $n_c = \mathrm{round}(500 \cdot 100^{-c/99})$, 10,899 samples) and compared Forgetting-LT-0.3 against Random-LT-0.3 at matched budget. $\tilde{\beta}$ is positive for both groups ($+2.141$ and $+0.674$), the base imbalance (Gini ≈ 0.59–0.67 even under random pruning) dominates the survive-rate covariate, and tail classes with five or six training examples make per-class NFR extremely noisy. More importantly, this is a rung-1 comparison: the two groups differ in Gini as well as in within-class selection, so it cannot separate the two channels. We report it for completeness and exclude it from Table 3.

C FULL MEASUREMENT PIPELINE

Figure 7 traces every quantity in the paper back to the trained-model logits it is computed from. Construction and training are shown at the top; the two measurement columns correspond to the aggregate branch (which produces the rung-2 null result) and the per-class branch (which produces β and $\tilde{\beta}$); the three branches at the bottom terminate in the pre-registered decision rules. The control ladder of Figure 1 is the abstraction of the lower half of this diagram: rung 2 is Branch A, rungs 3–5 are Branch B, and rung 1 together with the imbalance fit is Branch C.

D FULL EXPERIMENTAL RECORD

The supplementary material contains: the complete Phase 0–0e trajectory including the retracted aggregate-churn conclusions; ρ_{self} at every quality-control checkpoint; the full 14-configuration grid; both non-pruning control families; every q-mixture point at every selection seed; FLOPs estimates for selection and training separately; per-selseed cross-dataset results; the partial-correlation placebo; and a provenance table mapping every number in this paper to the JSON file it is read from.

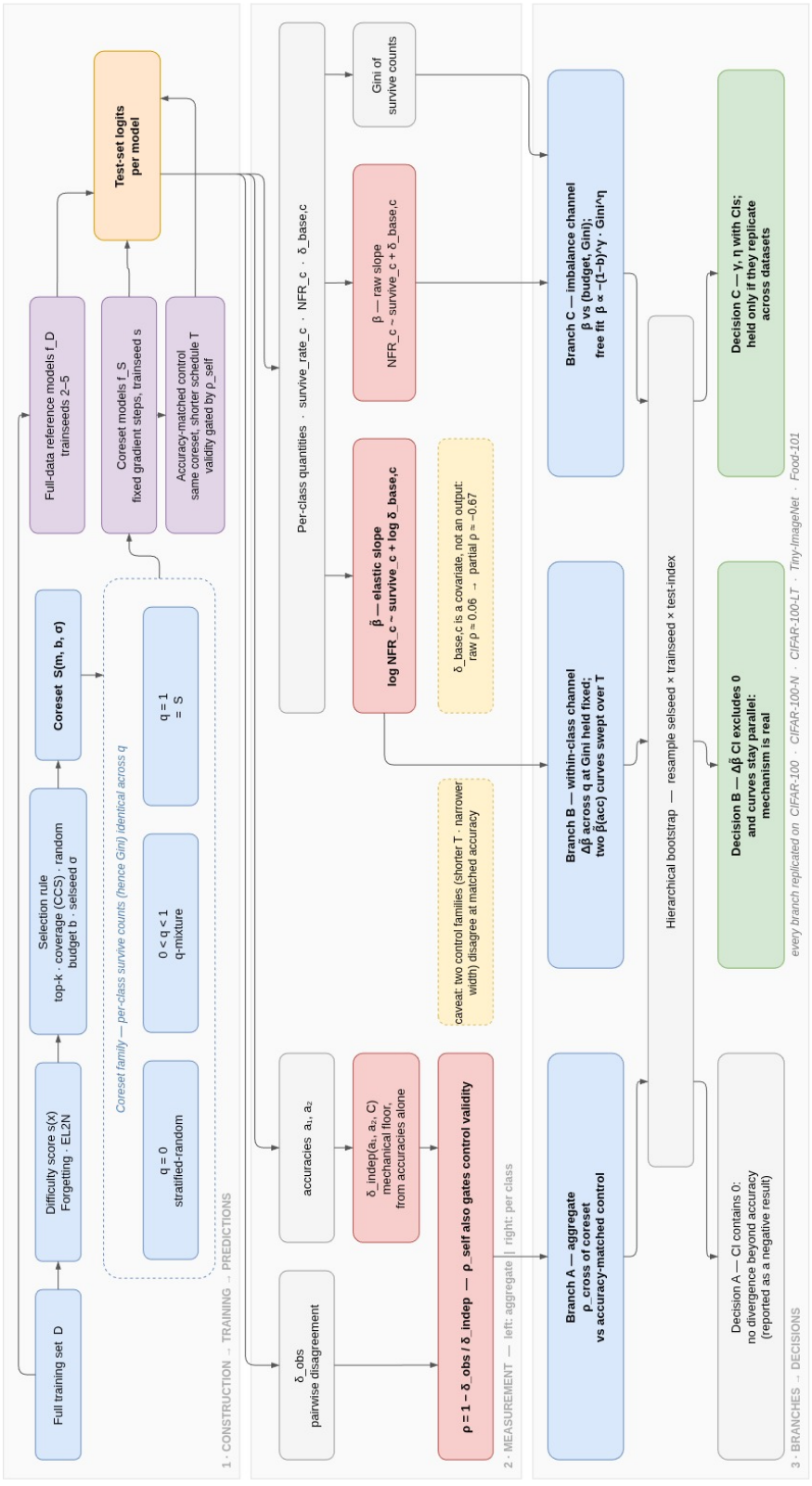

Figure 7: Full measurement pipeline (rotated; best viewed zoomed). Boxes are quantities or artefacts; arrows are computations. δ_{indep} is computed from accuracies alone, not from δ_{obs}; $\delta_{\text{base},c}$ enters the per-class regressions as a pre-treatment covariate rather than as an output; and Gini is a configuration-level summary used to compare β across configurations, not an input to any single β.