When Does a Per-Example Difficulty Score Go Stale? Loss-Based Scorers Replicate Across Training Runs Far Better Than Forgetting-Based Ones

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Compute-allocation schemes that steer extra training budget toward high-value examples need a per-example value or difficulty score, and a practical design choice is whether that score must be recomputed online, in-loop, during the run it steers, or can instead be pre-scored once and reused. Prior work on a forgetting-decay score reports that the score is essentially uncorrelated across independent training runs (Spearman $\rho \approx 0.01$), implying that reuse is unsound and online recomputation is required. We test this at CIFAR scale with a synthetic value-weighted compute auction on ResNet-18 and DeiT-Small (a compact, from-scratch vision transformer), comparing static (frozen, pre-scored) against online (in-loop, recomputed) auctions for a forgetting-based score on the full matrix, and a loss-based score on a reduced, CIFAR-10-only robustness check; we separately measure cross-run score correlation across training-stage checkpoints on ResNet-18, for both scorer families and both datasets. We did not detect a nontrivial accuracy effect from the auction mechanism for the forgetting scorer: every static, online, and ensembled-static arm lands within 1.4 percentage points of a uniform-compute control, at n=3 seeds per cell. The correlation diagnostic tells a sharper story: across two datasets, three independent seed pairs, and four training-stage checkpoints, the loss scorer's mean cross-run correlation ($\rho \in [0.79, 0.92]$ across datasets, at the final checkpoint, ResNet-18 only) is consistently higher than the forgetting scorer's ($\rho \in [0.47, 0.58]$), with point estimates that do not overlap in any cell we measured (a descriptive comparison, not a significance test). Both estimates sit far above the near-zero correlation the forgetting-decay literature reports. This is a CIFAR-scale, small-n study with no formal significance testing throughout; within that scope, a team choosing between these two scorer families should expect the loss-based one to tolerate a stale, pre-computed score better than the forgetting-based one, and it is not the static/online distinction itself where the practical risk lives.

1 Introduction

Training pipelines that allocate extra optimizer steps to a subset of "high-value" examples need a value or difficulty score for every example in the dataset. That score can be computed two ways: *online*, i.e. recomputed in-loop as the run it steers progresses, or *static*, i.e. computed once by a separate scout run and reused unchanged. Online scoring is more expensive to implement

*Corresponding author: deep@lab.cloud

and to run [1, 2], so whether it is actually necessary is a real engineering and research question, not just a theoretical one: it turns on whether per-example difficulty is an intrinsic, transferable property of the data or is instead dominated by run-specific training stochasticity. A recent study of forgetting-decay scores in fine-tuned image classifiers reports a near-zero cross-run rank correlation [3], which would argue strongly for online recomputation.

We test this question directly with a synthetic value-weighted compute auction: batches scored in the top decile receive additional optimizer steps, and we compare a static auction (scored once by a frozen scout run) against an online auction (scored in-loop by the run itself) and a uniform-compute control, at matched total training budget. We run this for a forgetting-based score, on ResNet-18 [4] and DeiT-Small [5] trained from scratch on both CIFAR-10 and CIFAR-100, and a loss-based score, on the same two architectures on CIFAR-10 only, as a robustness check. We also measure the underlying diagnostic directly, the cross-run rank correlation between a frozen scout score and an independently trained run's own score, extending it beyond a single final-epoch estimate to a trajectory across four training-stage checkpoints, on ResNet-18.

Our contributions:

1. We show that, for a forgetting-based scorer, we did not detect either the static or the online variant of this auction beating a uniform-compute control by a nontrivial margin on either CIFAR-10 or CIFAR-100, on either architecture (§5.1, at n=3 seeds per cell, descriptive comparisons only) – which limits how much the static-versus-online comparison alone can teach us.

2. We measure the cross-run score-correlation trajectory across four training-stage checkpoints for two scorer families and two datasets, on ResNet-18 only, and find the loss scorer's correlation point estimate is higher than the forgetting scorer's at every checkpoint we measured, with non-overlapping point estimates across three independent seed pairs (§5.2) – both scorers replicate the field's counter-evidence to the near-zero-correlation claim, and at the final checkpoint the two scorer families differ from each other by roughly 1.3 to 2 times, growing larger (up to about 3.3 times) earlier in training when both correlations, and especially the forgetting scorer's, are smaller and the ratio is correspondingly less stable.

3. We identify a secondary, architecture-specific effect: a loss-based value-weighted auction measurably harms DeiT-Small's accuracy while leaving ResNet-18 within about 1 percentage point of control (at n=2 seeds for ResNet-18, versus 3 for DeiT-Small; §7) under the same scorer and threshold (§5.3), which we flag as an open scorer times architecture interaction rather than a resolved explanation.

2 Related Work

Cross-run score stability is a contested question, not a settled one. A forgetting-decay score used to build a spaced-repetition training curriculum is reported to have near-zero cross-seed rank correlation ($\rho \approx 0.01$) between independently trained seeds [3], which is the specific claim our diagnostic tests. Several other studies, however, report the opposite: an ensembled hardness score used for offline DPO example selection labels a highly consistent set of examples as "difficult" across 40 runs that vary in model, seed, and training-data subset [6]; a frozen curriculum built from a separate scout model series transfers successfully across models and datasets [7]; and a coreset-selection method built on a static hardness score beats a dynamic in-loop baseline on

CIFAR-100 at higher pruning fractions [8]. Reported magnitudes in this literature range from near-zero [3] to above 0.9 [6] depending on scorer family and domain, with several studies clustering in a moderate band in between [9, 10]. Our results sit on the moderate-to-high side of this split for both scorer families we test, consistent with the latter group.

Scorer family matters more than the literature's framing often suggests. A single-snapshot score such as EL2N is markedly less cross-run-stable than training-dynamics or windowed scores such as forgetting counts [11], and a separate study finds forgetting/consistency-style scores *more* seed-robust than a loss-based score on CIFAR-10 [10] – the opposite ordering from what we observe. We take this as evidence that stability rankings between scorer families are not universal and must be measured per project rather than assumed from a single prior result, which is exactly what our study does for the forgetting-versus-loss pair.

Overhead accounting and threshold sensitivity are recurring pitfalls. MATES reports online-selection overhead shrinking from 18.0% to 7.3% of total compute as model scale grows [1]; RHO-LOSS folds proxy-model training cost into its overall FLOPs comparison [2]; and a temporal-influence study shows that restricting online rescoring to early and late training phases recovers about 96% of the benefit of continuous rescoring at more than five times lower overhead [12]. Separately, naive top-k thresholding on a value score can catastrophically collapse accuracy at aggressive allocation ratios, dropping below a random baseline [11]. Our auction's decile threshold and step multiplier are fixed rather than swept, a scope choice we return to in §7.

Static baselines are often tested at their weakest. Ensembling a scout score over multiple seeds or reference models measurably improves both cross-seed stability and downstream performance [10, 6], motivating the ensembled-static arm we include alongside a single-scout static arm.

CIFAR-10/CIFAR-100 with ResNet-18 is the dominant shared testbed for this literature, giving directly comparable baseline numbers [13, 11, 8, 2, 14], and a uniform or random-sampling control is the near-universal sanity baseline against which value-weighted methods are compared [14, 2]. Spearman rank correlation between per-example scores, computed mean-versus-mean, mean-versus-run, or run-versus-run [9], and averaged across independently varied seeds [10], or across many training checkpoints against an expensive ground truth [15], is the field's standard score-stability diagnostic and is what we adopt for our checkpoint-trajectory measurement.

3 Method

The auction mechanism is shared across every value-weighted arm. Within each training epoch, batches whose examples score in the top decile of a per-example value score receive two optimizer steps that epoch; the remaining 90% of batches receive one. This ratio is a fixed constant, not swept, across every arm and every job in this study. For a **forgetting-based** score, the number of times a training example's prediction flips from correct to incorrect across epochs [16], the same family of scores this project's motivating hypothesis is built on [3], we compare three arms against a uniform-compute control that applies no value weighting:

- **Static, single-scout.** The value score is computed once by a separate scout run (seed A) and frozen; the steered run (seed B) reuses it unchanged throughout training.
- **Static, ensembled-scout.** The frozen score is the mean over three independent scout runs on the same dataset/architecture cell, rather than a single scout, testing whether ensembling closes any gap to the online arm.

- **Online, in-loop.** The score is recomputed from the steered run's own training trajectory as it progresses, using no scout run.

For a **loss-based** score, an exponential moving average of per-example training loss updated every epoch, we run a reduced two-arm robustness check on the same control, static single-scout and online in-loop, without the ensembled-scout variant. Both scorer families steer the same top-decile/2× auction mechanism, so any difference between them isolates the scorer, not the auction design.

For the checkpoint-correlation diagnostic (§5.2), run on ResNet-18 only (§7 explains why DeiT-Small was dropped for this diagnostic), we run a scout (seed A) and an independent (seed B) pass with no steered auction pass, and snapshot each run's per-example score at four points in training, 25%, 50%, 75%, and 100% of the schedule, computing the Spearman rank correlation between the two runs' scores at each matched checkpoint. This isolates the score-stability question from the auction's own effect on accuracy, and means no downstream accuracy is measured for these particular runs: the accuracy-side static-versus-online comparisons (Tables 1 and 3) and the checkpoint-correlation diagnostic (Table 2) come from separate sets of jobs.

4 Experimental Setup

Data. CIFAR-10 and CIFAR-100 [17], standard 32×32 color image classification benchmarks, 50,000 training and 10,000 test images each. We hold out a fixed, class-stratified 5,000-image validation slice from the training pool (45,000/5,000 split), shared across every arm and seed in a given cell so all comparisons within a cell use the same validation set; the canonical 10,000-image test split is reserved and not used in this study's comparisons.

Models. ResNet-18 [4] and DeiT-Small [5], both trained from scratch, with no ImageNet initialization for either architecture. Our DeiT-Small is a compact vision transformer built to consume CIFAR's 32×32 images natively (patch size 4), not the standard 224×224-input configuration typically initialized from ImageNet weights. Training both architectures from scratch is a deliberate deviation from DeiT-Small's more common ImageNet-pretrained usage, adopted specifically to keep both architectures on the same from-scratch/pretrained footing, since static scoring's reported failures and successes split along roughly that line in prior work [15] (§7 discusses a confound in that source we could not fully control for here).

Protocol. Each seed run pairs a scout run (seed A) with a steered run (seed B). Every dataset/architecture cell in the phase-1 forgetting-scorer matrix uses three independent seed pairs, with all arms in a cell sharing the same three seed-A scout runs; the loss-scorer robustness slice uses two seed pairs on ResNet-18 and three on DeiT-Small (a third pair was added after the first two showed a consistent effect). The checkpoint-correlation study (phase 2) uses ResNet-18 only, on both datasets and both scorer families, again with three independent seed pairs per cell. All jobs share one fixed hyperparameter set (learning rate, batch size, epoch count) within a comparison, so that final top-1 validation accuracy is compared across arms trained for the same nominal epoch schedule. This is not a matched-FLOPs comparison: value-weighted arms give top-decile batches twice the optimizer steps of the control, and we did not compute a separate accounting of the online arms' scoring-recomputation overhead against the static arms' one-time scout cost (§7).

Table 1: Mean validation accuracy (3 seeds), forgetting scorer: uniform control (C), static single-scout (A), online in-loop (B), and static ensembled-scout (D). Deltas are relative to the cell's control.

Cell	C (control)	A (static)	B (online)	D (ensembled)
CIFAR-10 / ResNet-18	0.9271	0.9230 (–0.4pp)	0.9285 (+0.1pp)	0.9255 (–0.2pp)
CIFAR-100 / ResNet-18	0.7215	0.7195 (–0.2pp)	0.7187 (–0.3pp)	0.7146 (–0.7pp)
CIFAR-10 / DeiT-Small	0.5426	0.5530 (+1.0pp)	0.5505 (+0.8pp)	0.5410 (–0.2pp)
CIFAR-100 / DeiT-Small	0.3278	0.3220 (–0.6pp)	0.3215 (–0.6pp)	0.3135 (–1.4pp)

5 Results

No formal significance test is used anywhere below; comparisons are descriptive (means and ranges over 2–3 seeds), and we flag where a finding rests on a thin seed count (§7 discusses this choice and its risks).

5.1 The auction shows no detectable accuracy effect for a forgetting-based score

Table 1 reports mean validation accuracy for the uniform-compute control and the three forgetting-scorer arms, per dataset/architecture cell, averaged over three seeds. Every arm lands within 1.4 percentage points of its cell's control, which is inside the control's own seed-to-seed spread (0.2 to 2.1 percentage points across the three seeds). We did not detect a consistent, nontrivial improvement over uniform compute for any arm, nor did the online arm consistently beat the static arm; given the small seed count, this is an absence of a detected effect, not evidence the true effect is zero.

The forgetting-scorer diagnostic correlation over this same matrix, the final-epoch Spearman correlation between a cell's static score (arm A) and its matched in-loop score (arm B), ranges from 0.47 (CIFAR-100/ResNet-18) to 0.70 (CIFAR-100/DeiT-Small) for the single-scout static arm, and from 0.59 to 0.78 for the ensembled-static arm. Both ranges are moderate to high, not the near-zero correlation the forgetting-decay literature reports [3], and instead replicate the counter-evidence positions in §2 [6, 7]. These values are recorded by Table 1's arm-A jobs themselves; the independent, purpose-built correlation study in §5.2 uses separate jobs entirely and lands in a closely matching range for the cells the two studies share (e.g. CIFAR-10/ResNet-18: 0.583–0.591 here versus 0.570–0.603 in Table 2), which we read as an informal cross-check rather than a claim of statistical agreement.

5.2 Cross-run score correlation across training checkpoints

Our main diagnostic result extends the correlation measurement from a single final-epoch estimate to a trajectory across training, using a dedicated set of runs (scout and independent passes with no steered auction pass) rather than the accuracy-comparison jobs above. Using the ResNet-18-only checkpoint study (both datasets, both scorer families, three independent seed pairs per cell), Table 2 reports the final-checkpoint (100% of the 20-epoch schedule) Spearman correlation between a frozen scout score and an independently trained run's own score.

The loss scorer's correlation is higher than the forgetting scorer's on both datasets: the two point-estimate ranges do not overlap in either cell (a descriptive comparison over 3 seed pairs, not a significance test), and on CIFAR-100 the gap between the range midpoints is roughly two-fold (0.917–0.920 versus 0.470–0.475). Figure 1 shows this gap holds at every measured checkpoint, not

Table 2: Final-checkpoint cross-run Spearman correlation, mean (range across 3 seed pairs), ResNet-18 only.

Scorer	CIFAR-10	CIFAR-100
Forgetting	0.584 (0.570–0.603)	0.473 (0.470–0.475)
Loss	0.791 (0.775–0.800)	0.919 (0.917–0.920)

only at convergence: it is present from the earliest checkpoint (25% of the schedule) onward, though it narrows somewhat in absolute terms as the forgetting scorer's own correlation climbs through training. The forgetting scorer's earliest reading (0.198–0.216 across the two datasets) is the closest either scorer comes in this study to the near-zero correlation reported in prior work [3], but it is already climbing well past that range by the 50% checkpoint (0.361–0.415) and continues to rise for the remainder of training. A secondary pattern is visible in the same data: the loss scorer's correlation is higher on the harder dataset, CIFAR-100 exceeding CIFAR-10, at every checkpoint we measured, while the forgetting scorer's is lower on the harder dataset from the 50% checkpoint onward, the opposite direction; at the earliest (25%) checkpoint the forgetting scorer briefly runs the other way (CIFAR-100 0.216 versus CIFAR-10 0.198), so this reversal is a late- and mid-training pattern, not one that holds across every checkpoint we measured. We do not have an explanation for this reversal and flag it as an open question in §7. This diagnostic reports whole-distribution rank correlation; it does not directly measure how much the static and online scores agree specifically on which examples land in the auction's top-decile cutoff, which is the set the auction mechanism actually acts on (§7).

5.3 A loss-based auction shows a large accuracy drop on DeiT-Small

We separately tested the loss-based scorer on a reduced slice (CIFAR-10 only, both architectures). Table 3 shows the largest accuracy deviation from control in this study: the loss-based auction shows a mean accuracy drop on DeiT-Small (static arm 5.99 percentage points below control, all three seeds negative; online arm 2.69 percentage points below control on average, 2 of 3 seeds negative) while ResNet-18 stays within about 1 percentage point of control (at n=2 seeds, too few to characterize ResNet-18's own seed-to-seed spread). The forgetting-scorer arms on the same two architectures (Table 1) show no comparable architecture-specific drop. We treat this as the largest deviation among many cells examined across this study, not a pre-registered single comparison, and DeiT-Small's from-scratch training on this dataset is already the most fragile cell in our whole matrix (§7), so we hold this finding more loosely than the correlation result above. The loss scorer's static-versus-online correlation on this slice (0.74 to 0.79) is higher than the forgetting scorer's on the corresponding cells, which at minimum rules out one candidate explanation, that the static and online scores simply disagree with each other on DeiT-Small; it does not by itself establish what the shared ranking is doing wrong (§7).

On the individual per-seed level, the static loss-scored DeiT-Small arm lands below control on all three seeds (–7.06, –7.22, and –3.68 percentage points); the online arm lands below control on two of three seeds (–6.48 and –3.86 percentage points, with one seed at +2.28 percentage points).

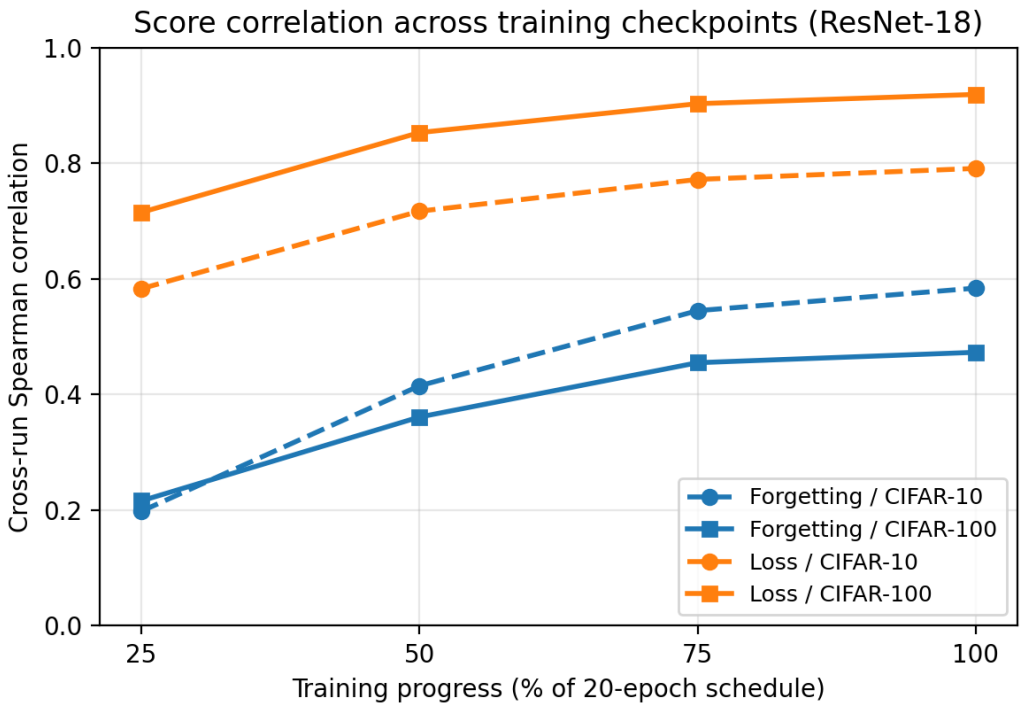

Figure 1: Cross-run Spearman correlation between a frozen scout score and an independent run's own score, at four training-stage checkpoints (25/50/75/100% of a 20-epoch schedule), for the forgetting and loss scorers on CIFAR-10 and CIFAR-100. Each point is the mean across 3 seed pairs; ResNet-18 only.

6 Discussion

The results in this paper answer different questions and point in different practical directions. The auction-mechanism comparison (§5.1) asks whether recomputing a value score in-loop changes downstream accuracy relative to reusing a frozen score, and for the forgetting scorer at this scale and this fixed decile/2× threshold, we did not detect either variant doing much of anything relative to a uniform-compute control. That result limits how much this comparison alone can say about static-versus-online scoring, since a mechanism whose accuracy effect we could not detect cannot cleanly discriminate between two ways of feeding it a score (§7).

The correlation diagnostic (§5.2) is a more direct test of a closely related question, whether a per-example difficulty score is a stable, run-independent property of the data or an artifact of one particular training trajectory, and it gives a clear answer for both scorer families we tested: neither scorer replicates the near-zero correlation reported for forgetting-decay scores elsewhere [3], and the two scorer families differ substantially from each other, with the loss-based score consistently and considerably more cross-run-stable than the forgetting-based score. Of the two levers we tested, this suggests scorer-family choice is more likely to matter than the static-versus-online mechanism itself, at least within the two families and the compute scale we tested. This inference has two gaps we did not close: we measured whole-distribution rank correlation, not agreement specifically on the auction's top-decile cutoff (§5.2), and our one downstream-accuracy comparison for the loss scorer (§5.3) is CIFAR-10 only, so it does not directly test whether the CIFAR-100 cell, where the

Table 3: Loss-scorer robustness slice, CIFAR-10 only, mean validation accuracy vs. uniform control (C). ResNet-18 is averaged over 2 seeds; DeiT-Small over 3 (§7 discusses how the third DeiT-Small seed was added).

Architecture	C (control)	E (static)	F (online)
ResNet-18	0.9284	0.9184 (–1.0pp)	0.9252 (–0.3pp)
DeiT-Small	0.5426	0.4827 (–5.99pp)	0.5157 (–2.69pp)

correlation gap between scorers is largest, would show a correspondingly larger accuracy consequence.

The loss-scorer/DeiT-Small accuracy drop (§5.3) is the largest deviation from control we observed in this study, on one architecture. Because DeiT-Small's static and online loss-based rankings agree closely with each other, we can at least rule out a simple disagreement between the two scoring procedures as the explanation; that leaves open, and does not resolve, whether the shared ranking is itself a poor fit for how DeiT-Small trains from scratch at this scale, or whether the fragility of this particular cell (§7) is doing some of the work. We did not run the additional experiments that would be needed to isolate the mechanism.

7 Limitations

Statistical caveats (read first). Every comparison in this paper is descriptive: means over 2–3 seeds, with per-seed min–max ranges given where we report them (the correlation study, Table 2; and the loss-scorer robustness slice, Table 3) but not for the core forgetting-scorer matrix (Table 1), which reports only per-cell means against the control's own pooled seed spread. No significance test, confidence interval, or power analysis appears anywhere. A non-overlapping range between two groups of 3 is a much weaker claim than it may sound (it does not bound a false-positive rate the way a confidence interval does), and a "detected effect" or "no detected effect" at this sample size should be read as exactly that, not as evidence that a true effect is present or absent. Two seed counts are worth flagging specifically: the ResNet-18 loss-scorer cell in Table 3 has only 2 seeds, too few to characterize its own spread; and DeiT-Small's third seed in that same table was added *because* the first two seeds already showed the effect we report, an ad hoc, data-dependent stopping rule (not pre-registered) that plausibly biases the reported DeiT-Small effect size and its apparent unanimity upward. We did not apply a matching third seed to ResNet-18, so the table's two rows are not on equal evidentiary footing.

Internal validity. The forgetting-scorer core matrix landed at parity with its uniform-compute control on every cell we tested, which makes the static-versus-online comparison for that scorer close to untestable: a mechanism whose accuracy effect we could not detect cannot cleanly discriminate between two ways of computing its input score. The auction's own knobs (decile threshold, 2× step multiplier) were fixed rather than swept, by design, so we cannot rule out that a differently tuned auction would separate the static and online arms more clearly. The matched-total-FLOPs comparison did not include a separate accounting of the online arms' scoring-recomputation overhead against the static arms' one-time scout cost; only aggregate per-job wall-clock and GPU-hour totals were tracked, so we cannot report the field-standard scoring-versus-training compute breakdown that the closest related work provides [1, 2]. Separately, the checkpoint-correlation diagnostic (§5.2) reports Spearman correlation over the full score distribution; it does not report agreement specifically on which examples fall in the auction's top-decile cutoff, which is the only part of the distribution the

auction mechanism actually consumes. Two scorers could show similar whole-distribution correlation while disagreeing sharply near the cutoff, or vice versa, so the practical "tolerates a stale score better" framing in the Abstract and Discussion is an inference from the whole-distribution statistic we measured, not a direct measurement of top-decile-set stability.

DeiT-Small: disqualified from one analysis, retained for another. DeiT-Small was dropped from the checkpoint-correlation study (§5.2) because its from-scratch training regime is the most fragile cell in the phase-1 matrix (control accuracy near 0.53 to 0.55 on CIFAR-10, far below ResNet-18's 0.93) and we judged its inclusion would confound the correlation trajectory; the forgetting-versus-loss ordering we report for DeiT-Small therefore rests only on the phase-1 final-epoch estimate, not on the checkpoint trajectory. The same fragile cell is, however, the entire evidentiary basis for the loss-scorer accuracy-drop finding (§5.3). We do not have a principled way to say that fragility disqualifies one analysis but not the other; we flag both results as resting on this cell and let the reader weigh the accuracy-drop finding accordingly, rather than either suppressing it or presenting it with the same confidence as the ResNet-18-based correlation result.

External validity. We tested exactly two scorer families, forgetting-based and loss-based; other proxy families we did not test, gradient-norm and embedding-based scores among them, may behave differently and are out of scope here. Both architectures were trained from scratch rather than from an ImageNet-pretrained initialization; static scoring's reported failures are on from-scratch image classification, while its reported successes are on fine-tuning a pretrained backbone [15], though that comparison is confounded with a modality shift (image versus text) in the source, so our results may not transfer cleanly to a pretrained-backbone setting. All experiments are at CIFAR scale; we make no claim about how either finding scales to larger datasets or models.

Construct validity. No job in this study retained per-example predictions or logits, so no calibration or per-example error analysis was possible; our error analysis is limited to aggregate, cell-level patterns. The direction reversal we observed between the two scorers' response to task difficulty, the loss scorer's correlation rising on the harder CIFAR-100 task while the forgetting scorer's falls (§5.2), is reported as an open question, not a resolved mechanism.

8 Availability

No code, data, or checkpoints from this study are being made public at this time. Artifacts, including the training scripts, per-job configs, and the full run-level results underlying every table and figure in this paper, are available from the authors on request.

9 Conclusion and Future Work

We set out to test whether a per-example value score for a compute auction needs to be recomputed online or can be safely reused from a one-time scoring pass, motivated by a prior report of near-zero cross-run correlation for a forgetting-decay score. We did not detect an accuracy effect from the auction itself for a forgetting-based score at CIFAR scale, which limits what that comparison alone can teach us, but the underlying diagnostic is informative on its own: across two datasets, three seed pairs, and four training-stage checkpoints, on ResNet-18, a loss-based score's cross-run correlation ($\rho \in [0.79, 0.92]$ at the final checkpoint) is consistently higher than a forgetting-based score's ($\rho \in [0.47, 0.58]$), and both sit well above the near-zero correlation reported elsewhere for forgetting-decay scores. We also find a narrower, architecture-specific accuracy drop: a loss-based

auction shows a large mean accuracy decline on DeiT-Small (the most fragile cell in our matrix, §7) while ResNet-18 stays within about 1 percentage point of control (n=2 seeds) under the same scorer and threshold. All of this is descriptive, small-n evidence with no formal significance testing (§7); we present it as a set of directions worth a larger follow-up, not a settled result. For future work, we rank: (1) extending the checkpoint correlation study to DeiT-Small with a training recipe suited to vision transformers from scratch, to test whether the loss-versus-forgetting ordering we report holds across architectures; (2) sweeping the auction's decile threshold and step multiplier, since the current fixed setting may be why we could not detect an accuracy effect; (3) adding a gradient-norm or embedding-based scorer family to test whether the loss-versus-forgetting stability gap generalizes to a third scoring approach; and (4) measuring top-decile-set overlap directly, alongside whole-distribution correlation, to test whether the practical "tolerates a stale score" claim holds at the specific cutoff the auction uses.

References

[1] Zichun Yu, Spandan Das, and Chenyan Xiong. MATES: Model-Aware Data Selection for Efficient Pretraining with Data Influence Models. 2024. URL https://arxiv.org/abs/2406.06046.

[2] Andreas Kirsch. Advancing Deep Active Learning & Data Subset Selection: Unifying Principles with Information-Theory Intuitions. 2024. URL https://arxiv.org/abs/2401.04305.

[3] Miit Daga and Swarna Priya Ramu. Not All Forgetting Is Equal: Architecture-Dependent Retention Dynamics in Fine-Tuned Image Classifiers. 2026. URL https://arxiv.org/abs/2604.11508.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 2015. URL https://arxiv.org/abs/1512.03385.

[5] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training Data-Efficient Image Transformers & Distillation Through Attention. 2020. URL https://arxiv.org/abs/2012.12877.

[6] Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and Zhiqiang Xu. Principled Data Selection for Alignment: The Hidden Risks of Difficult Examples. 2025. URL https://arxiv.org/abs/2502.09650.

[7] Zhenwei Tang, Amogh Inamdar, Ashton Anderson, and Richard Zemel. Level Up: Defining and Exploiting Transitional Problems for Curriculum Learning. 2026. URL https://arxiv.org/abs/2603.13761.

[8] Pranav Ramesh, Arjun Roy, Deepak Ravikumar, Kaushik Roy, and Gopalakrishnan Srinivasan. The Easy Path to Robustness: Coreset Selection using Sample Hardness. 2025. URL https://arxiv.org/abs/2510.11018.

[9] Devin Kwok, Nikhil Anand, Jonathan Frankle, Gintare Karolina Dziugaite, and David Rolnick. Dataset Difficulty and the Role of Inductive Bias. 2024. URL https://arxiv.org/abs/2401.01867.

[10] Simon Rampp, Manuel Milling, Andreas Triantafyllopoulos, and Björn W. Schuller. Does the Definition of Difficulty Matter? Scoring Functions and their Role for Curriculum Learning. 2024. URL https://arxiv.org/abs/2411.00973.

[11] Yeseul Cho, Baekrok Shin, Changmin Kang, and Chulhee Yun. Lightweight Dataset Pruning without Full Training via Example Difficulty and Prediction Uncertainty. 2025. URL https://arxiv.org/abs/2502.06905.

[12] Jiachen T. Wang, Dawn Song, James Zou, Prateek Mittal, and Ruoxi Jia. Capturing the Temporal Dependence of Training Data Influence. 2024. URL https://arxiv.org/abs/2412.09538.

[13] Muhammad Asif Khan, Ridha Hamila, and Hamid Menouar. Accelerating Deep Learning with Fixed Time Budget. 2024. URL https://arxiv.org/abs/2410.03790.

[14] Patrik Okanovic, Roger Waleffe, Vasilis Mageirakos, Konstantinos E. Nikolakakis, Amin Karbasi, Dionysis Kalogerias, Nezihe Merve Gurel, and Theodoros Rekatsinas. Repeated Random Sampling for Minimizing the Time-to-Accuracy of Learning. 2023. URL https://arxiv.org/abs/2305.18424.

[15] Ziao Yang, Longbo Huang, and Hongfu Liu. Layer-Aware Influence for Online Data Valuation Estimation. 2025. URL https://arxiv.org/abs/2510.16007.

[16] 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. 2018. URL https://arxiv.org/abs/1812.05159.

[17] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.