WHAT DOES NEW-TASK ACCURACY MEASURE IN CONTINUAL LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Does lower new-task accuracy indicate a reduced ability to learn? In class-incremental learning, the answer depends on the update procedure and prediction problem. We test this inference by pairing actual learning with current-only adaptation from the same inherited parameters, then changing prediction candidates at fixed weights. Across ten paired CIFAR-10 seeds, experience replay trails SGD by 11.151 percentage points in actual new-task accuracy yet leads by 0.971 points after matched adaptation. Restricting actual predictions to current classes reverses the mean ranking without changing the probe scores. The apparent deficit thus coexists with better within-task discrimination and better matched adaptation. Retention gives the comparison its practical meaning: both standard probes have zero old-task accuracy, while a 20% retention requirement favors actual replay over the tested validation-selected probes on CIFAR-10.1/.2. Fixed-history continuations and a full-budget CIFAR-100 ResNet experiment show where the rankings change; in the latter, ER leads on both actual and probe accuracy in all five seeds. Together, these results distinguish an observed new-task deficit from weaker adaptation of inherited parameters and show why current and retained performance must be evaluated jointly.

1 INTRODUCTION

A class-incremental learner must learn new classes while preserving useful performance on earlier ones. When its new-task accuracy falls, a natural explanation is that it has become less able to learn. That inference matters: it motivates interventions for plasticity loss and shapes comparisons between continual learning methods (French, 1999; Dohare et al., 2024). We test this inference in class-incremental image learning on CIFAR-10/100 and matched natural test sets. New-task accuracy is the outcome of a particular update and prediction procedure, so a lower score does not by itself identify weaker adaptation of the inherited parameters.

Experience replay (ER) makes the distinction concrete. Its updates combine current and historical examples; with a mean loss, each current example receives less weight than in current-only SGD. At prediction time, current classes also compete with outputs preserved for older tasks. Replay allocation and classifier bias are established concerns (Cho et al., 2026; Wu et al., 2019). We ask what a lower endpoint score tells us about adaptation from the parameters that the learner has accumulated.

We compare three branches at saved task boundaries: actual learning continues the learner's algorithm, while a probe applies current-only SGD from its historical parameters and a fresh branch applies the same operation from the saved initialization. They share current examples, minibatch order and update counts. The comparison separates actual learning, matched adaptation and the benefit of the inherited state; re-scoring fixed logits with current-class candidates tests within-task discrimination.

Our central result jointly compares these measurements on the same histories. A discovery study of 36 histories leads to a forty-history CIFAR-10 follow-up with ten new paired seeds: ER has lower actual accuracy in all ten and higher probe accuracy in nine. Supplying task identity reverses the mean actual ranking while leaving the probe ranking unchanged. ER's all-class deficit therefore coexists with better within-task discrimination and matched adaptation.

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.

The probe's high current accuracy comes with zero old-task accuracy, and cannot be combined with retention from a different model. We therefore score old and current accuracy on the same endpoints and select under positive retention constraints. An absolute 20% old-class requirement favors actual ER over the tested validation-selected probes.

Fixed-history continuations and a full-budget CIFAR-100 ResNet experiment locate where rankings change. Together, the measurements separate an all-class decision deficit, inherited-state adaptation and its retention cost.

2 Related Work

Plasticity and adaptation-based evaluation. Checkpoint and fresh-reference probes are established tools for studying plasticity (Lyle et al., 2023; 2025). Inamdar et al. (2026) evaluate few-shot adaptation and show that per-shot measurements change assessments based on current-task accuracy. We examine the interpretation of that assessment change by combining matched adaptation with fixed-weight candidate restriction and the retained accuracy of each endpoint (Appendix Table 9). Our probes update both backbone and classifier, whereas frozen-feature probes measure the information available to a new readout (Davari et al., 2022). Task similarity also shapes transfer and forgetting (Lee et al., 2021; Hiratani, 2024), and continual evaluation reveals transient stability gaps that task-end measurements miss (De Lange et al., 2023).

Co-observation and replay. Hess et al. (2026) identify data co-observation as a further factor in the gap to joint training, using checkpoint ensembles and incremental joint-training references. Replay changes co-observation as well as the objective optimized on the current task. Cho et al. (2026) study memory allocation on CIFAR-100 with a residual network and compare against fresh retraining on available seen-task data. Our fresh reference uses current-task data, matching the question of current-task adaptation. DER++ adds distillation to replay (Buzzega et al., 2020); we include it in a ten-seed comparison at the same CNN budget as SGD and ER.

Prediction access and classifier bias. Task identity defines different continual-learning prediction problems (van de Ven et al., 2022). We study both separately trained task-aware models and evaluation-only masks applied to class-incremental logits. BiC learns affine corrections for new-class scores on held-out examples (Wu et al., 2019), and generalized zero-shot learning uses group calibration curves (Chao et al., 2016). Our scalar experiment instead fits a shared offset on previously trained-on examples, giving an explicit old/current operating-point comparison at fixed weights.

Plasticity interventions. Long-sequence studies demonstrate plasticity loss and evaluate continual backpropagation and other interventions (Dohare et al., 2024). Regenerative regularization, normalization and activation design target trainability through different mechanisms (Kumar et al., 2025; Lyle et al., 2025; Lillo & Cheney, 2026). These methods reinforce the need to distinguish the performance of an active update rule from adaptation after changing that rule. Our framework makes that distinction across SGD, ER, DER++, and an official-semantics CBP port.

3 Separating the Questions Behind a New-Task Score

Actual, probe, and fresh branches. For seed s and task t, let $\theta^-_{s,t}$ be the parameters immediately before the actual continual learner encounters that task, and let θ^0_s be its saved initialization. At adaptation prefix k, denote current-task test accuracies by $a_{s,t}(k)$ for the actual algorithm, $p_{s,t}(k)$ for a current-only SGD probe initialized at $\theta^-_{s,t}$, and $F_{s,t}(k)$ for a fresh current-only SGD branch initialized at θ^0_s. All three use identical current-task minibatch orders and the specified optimizer settings. The actual branch additionally uses replay when required by its algorithm, processing additional training examples at the same number of updates.

Define the following differences in percentage points:

$$A_{s,t}(k) = F_{s,t}(k) - a_{s,t}(k), \tag{1}$$

$$H_{s,t}(k) = F_{s,t}(k) - p_{s,t}(k), \tag{2}$$

$$C_{s,t}(k) = p_{s,t}(k) - a_{s,t}(k), \qquad A_{s,t}(k) = H_{s,t}(k) + C_{s,t}(k). \tag{3}$$

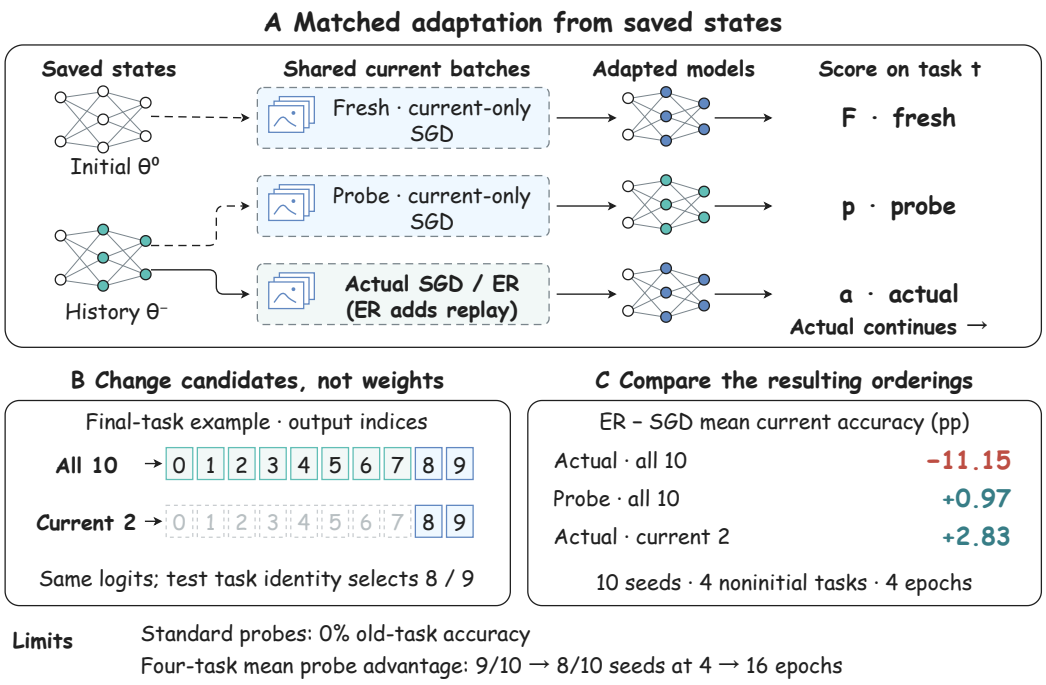


Figure 1: Diagnosing a new-task deficit through paired measurements. (A) Dashed arrows copy auxiliary states; only actual advances the history after four epochs. ER actual also restores memory and its random states; SGD stores none. Network and sample icons are schematic. (B) A final-task example of candidate restriction: output indices 8/9 denote current classes. The mask changes evaluation only and supplies task identity. (C) Descriptive ER-minus-SGD accuracy differences, averaged over ten seeds and four noninitial CIFAR-10 tasks. Positive values favor ER; Figure 2 shows the paired seed distribution and primary intervals. The budget annotation counts positive ER-minus-SGD probe differences after averaging the same four tasks within each seed.

A positive A indicates weaker actual adaptation than the specified fresh reference; positive H indicates weaker adaptation of the historical parameters under the specified probe. Negative values indicate the reverse. C compares complete adaptation procedures: $C > 0$ means that the current-only probe attains higher current accuracy than the actual continual update, while $C < 0$ means the actual update attains higher accuracy. The accounting identity $A = H + C$ expresses the actual gap as the historical-state gap plus this procedure contrast. Each term depends on the task, initialization, representation, classifier, optimizer and adaptation budget. In particular, $H < 0$ records a net benefit of the historical parameters under that probe; beneficial transfer and changes in trainability can both contribute to it.

Only the actual branch advances the subsequent history. Auxiliary branches start from saved states and never replace the next task's starting parameters or memory. Prefixes $k \in \{1, 2, 4\}$ are points along branches from the same pre-task checkpoint, whose preceding actual tasks each received four epochs. For plain SGD, actual and standard probe implement the same operation; equality of their final parameter hashes is an implementation check. Fresh and probe also coincide at task zero, giving a second implementation check.

Matched controls for the adaptation operation. For current/replay batch sizes n_c, n_r, ER's concatenated mean loss is $w_c\ell_c + (1 - w_c)\ell_r$, where $w_c = n_c/(n_c + n_r)$ and each ℓ is a mean cross-entropy. Weighted-current and extra-current probes test specified alternatives with their own fresh references. The unnormalized compensated ER branch uses $\ell_c + (1 - w_c)\ell_r$: it restores the current coefficient to one while preserving replay, and therefore increases the total coefficient from one to $2 - w_c$. It tests that complete update operation, including its larger gradient scale; it is not a pure

reweighting control. Appendix D.2 defines each control. All start from saved states and leave the continuing actual history intact.

4 Experimental Protocol

Data, heads, and optimization. The follow-up comprises forty completed histories: two evaluation settings, two actual algorithms (SGD and ER), and ten seeds, 201–210. It uses the official CIFAR-10 training and test splits (50,000 and 10,000 examples). For each seed, a random permutation of the ten classes defines five disjoint two-class tasks, each with 10,000 training and 2,000 test examples. A single class-to-output bijection is applied consistently to training and test labels: if the permuted original classes are $q_0, \ldots, q_9$, class q_i maps to output i, and task t uses rows $2t, 2t+1$ (zero-based). Initialization, current minibatch order, and class-order seeds are respectively $s + 10000$, $s + 30000$, and $s + 40000$, shared across the paired algorithms and evaluation settings. Images receive fixed channel normalization and no data augmentation.

We distinguish settings by their access to task identity (van de Ven et al., 2022). Our class-incremental variant uses a fixed ten-output head: all ten outputs participate in training and prediction. The task-aware setting retains the same ten-output parameterization but selects the two outputs assigned to each example's task. Replayed examples use their own stored task identities. These settings change both training and evaluation; Section 5.1 separately varies prediction candidates on the same trained models.

The CNN has three 3×3 convolutional layers with 32, 64, and 128 channels, each followed by ReLU and 2×2 max pooling, then a 256-unit ReLU layer and ten outputs. Each task receives four epochs of SGD with learning rate 0.01, batch size 64, zero momentum, and zero weight decay. Every branch constructs a new optimizer. ER maintains a total reservoir of 500 examples, draws up to 32 stored examples without replacement before each update, and then inserts current examples. Repeated presentations across epochs count as new reservoir arrivals, matching online reservoir sampling over the presentation stream and allowing repeated source images to occupy different slots. This has little effect on realized capacity here: at the final endpoint, the 500 slots contain 498.3±0.9 unique source examples on average across seeds (range 497–500). The last current batch has 16 examples: when $n_r = 32$, its current-loss coefficient is $1/3$, whereas a full batch has coefficient $2/3$. Coefficients are computed per update rather than assumed constant.

Endpoints and primary comparison. For any adaptation metric X, the seed-level value is $\bar{X}_s(k) = \frac{1}{4}\sum_{t=1}^{4} X_{s,t}(k)$; task zero is excluded. The primary comparison is class-incremental, prefix four:

$$(\Delta A_s, \Delta H_s) = (\bar{A}_s^{\mathrm{ER}} - \bar{A}_s^{\mathrm{SGD}}, \bar{H}_s^{\mathrm{ER}} - \bar{H}_s^{\mathrm{SGD}}). \tag{4}$$

The prespecified ordering pattern is $\Delta A > 0$ and $\Delta H < 0$. Methods are joined by explicit seed identity; incomplete or duplicate pairs are rejected. A shared 100000×10 bootstrap index matrix resamples entire paired seed vectors. The two 97.5% marginal Bonferroni-adjusted percentile intervals have nominal family coverage of 95%, with approximate coverage at ten seeds. The operational ordering criterion requires the ΔA lower bound to exceed zero and the ΔH upper bound to be below zero. Seed values, means, sample standard deviations, and both complete intervals accompany this rule.

Secondary comparisons report means and sample SDs; inferential intervals are reserved for this primary contrast. Secondary statistics, training-count conventions, and provenance checks are specified in Appendix D; numerical preprocessing, initialization and executable artifact specifications appear in Appendix G.

The discovery study selected the follow-up question. We froze the follow-up protocol before execution; its seeds exclude the discovery set, while the official test images are reused. The follow-up uses the full training set and randomized class order in place of the discovery configuration.

5 Lower Actual Accuracy, Better Matched Adaptation

At four epochs, ER learns the current task less accurately under its actual algorithm, but its inherited parameters adapt better under the common probe. Table 1 reports absolute endpoints; Figure 2

shows all paired seeds. We then test prediction candidates, calibration and retention-constrained selection on the saved models.

Setting	History	Fresh F	Probe p	Actual a
Class-incremental	SGD	77.14 ± 4.76	80.55 ± 5.22	80.55 ± 5.22
	ER	77.14 ± 4.76	81.52 ± 4.95	69.39 ± 7.54
Task-aware	SGD	79.46 ± 4.33	81.82 ± 4.67	81.82 ± 4.67
	ER	79.46 ± 4.33	82.13 ± 4.33	83.18 ± 4.34

Table 1: Follow-up accuracy at prefix four, in percent: mean $\pm$ sample SD over ten seeds. Each seed first averages its four noninitial tasks. The standard fresh reference F is shared by the paired algorithms; p is current-only adaptation and a is the actual algorithm. SGD actual and probe coincide by construction. Task-aware models use a different training loss and evaluation mask from class-incremental models.

The two rankings disagree on new seeds. For class-incremental adaptation at prefix four, the paired difference in actual gaps is $\Delta A = +11.151$ pp, with sample SD 6.752 pp and interval $[+7.430, +16.269]$ pp. The difference in historical-state gaps is $\Delta H = -0.971$ pp, with sample SD 0.726 pp and interval $[-1.480, -0.521]$ pp. The lower bound for ΔA exceeds zero and the upper bound for ΔH is below zero, meeting the frozen directional criterion. The bars are separate 97.5% marginal bootstrap intervals with nominal joint coverage of 95%.

ER has lower actual accuracy in all ten paired seeds and higher probe accuracy in nine; seed 202 has the opposite probe difference (Appendix Table 8).

The shared fresh reference cancels in the paired method contrasts: $\Delta A = \bar{a}_{\mathrm{SGD}} - \bar{a}_{\mathrm{ER}}$ and $\Delta H = \bar{p}_{\mathrm{SGD}} - \bar{p}_{\mathrm{ER}}$. Each is the negative of its ER-minus-SGD accuracy difference. The fresh reference answers a separate question: both class-incremental probes outperform initialization under the same adaptation procedure (Table 1). ER also achieves higher final accuracy across all five tasks: 42.964% versus 16.331% for SGD. Its lower actual new-task accuracy thus coexists with better overall retained performance. All-task accuracy and current-task adaptation measure distinct outcomes.

Sensitivity to the adaptation operation. Under weighted-current and extra-current probes, the ER-minus-SGD historical gap is -1.099 and -0.971 pp; each uses its own fresh reference (Appendix Figure 8). The unnormalized compensated ER operation reduces its actual-gap difference to $+5.104$ pp, but its coefficients sum to $2 - w_c$. After division by $2 - w_c$ preserves relative weighting at unit total scale, the gap remains positive at 1.881 ± 2.026 pp across ten seeds (Appendix L); gradient scale explains only part of the compensated result.

DER++ improves both endpoints relative to ER. Ten separately frozen DER++ histories use the same CNN, task order and four-epoch budget, with $\alpha = \beta = 0.5$ and two replay draws (Buzzega et al., 2020). Its mean actual/probe current accuracies are $78.365/82.624\%$: relative to SGD, actual accuracy is -2.181 pp and probe accuracy $+2.078$ pp. The actual difference has the opposite sign in three seeds, while every probe difference is positive. Relative to ER, both means improve ($+8.970/+1.106$ pp). The mean disagreement also occurs between DER++ and SGD, while DER++ leads ER on both endpoints. This comparison's prefixes, seed values, and actual retention appear in Appendix I. Section 6 finds that the final-task counterpart of this DER++–SGD disagreement disappears under continued training.

Continual Backpropagation separates the three outcomes. An official-semantics CBP port (Dohare et al., 2024) matches SGD actual accuracy (-0.0025 ± 0.043 pp) under the authors' CIFAR parameters, while its actual and probe endpoints differ by only 0.010 ± 0.016 pp across five paired seeds. Relative to ER, CBP is $+13.460$ pp on actual current accuracy, -0.770 pp under the common probe, and -25.142 pp on final all-task accuracy. A modern plasticity intervention therefore exposes a distinct current-adaptation– retention profile rather than collapsing the three measurements (Appendix J).

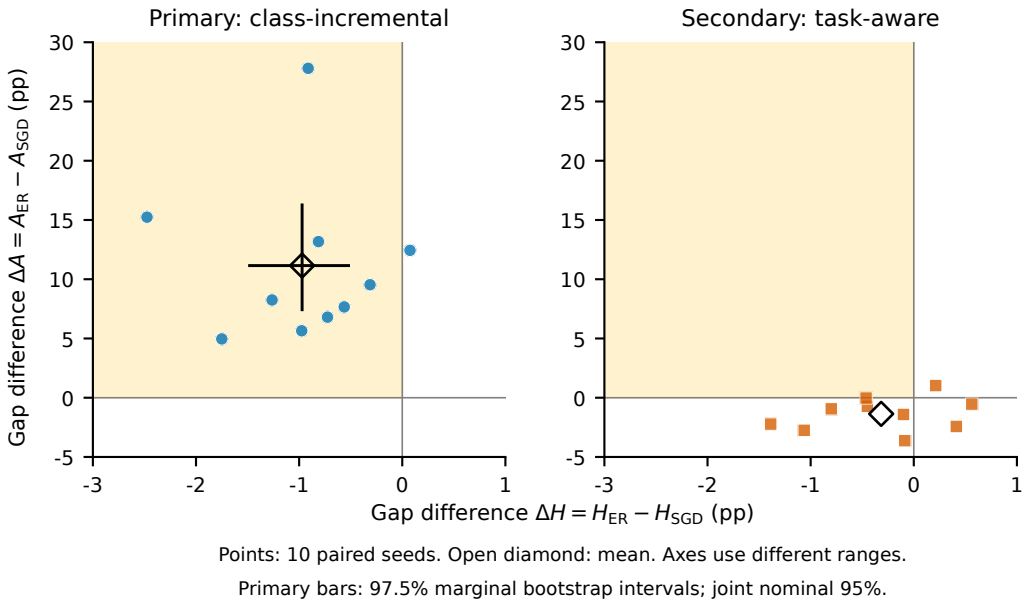


Figure 2: Paired ER-minus-SGD *gap* differences at prefix four. Points are seeds, open diamonds are means, and the shaded quadrant denotes a larger actual gap but smaller historical-state gap for ER. Only the prespecified class-incremental comparison has inferential bars: adjusted marginal bootstrap intervals from paired seed resampling. Task-aware points are secondary and descriptive. A positive ΔA means lower ER actual accuracy than SGD because $\Delta A = a_{\mathrm{SGD}} - a_{\mathrm{ER}}$; a negative ΔH means higher ER probe accuracy. The horizontal and vertical axes have different ranges; panels share the same ranges.

The actual-gap separation transfers across architecture. Keeping the CIFAR-10 stream, class order, four-epoch budget, optimizer and replay protocol fixed, we replaced the compact CNN with a pinned ResNet-32–GN8 on five paired seeds. The ER-minus-SGD actual-gap difference is $+23.07 \pm 17.38$ pp and is positive in all five seeds; the history-probe difference is -1.19 ± 3.25 pp and is negative in three. ER's final all-task accuracy is $30.63 \pm 5.21\%$ versus $16.29 \pm 3.30\%$ for SGD. The bridge therefore supports transport of the actual-ranking separation, while the probe component remains seed-sensitive (Appendix K).

The probe endpoint has an old-task cost. Post-hoc scoring of the same saved branches gives ER actual old/current accuracies of $43.560\%/69.395\%$, versus $0.000\%/81.517\%$ for its standard probe. SGD and ER standard probes have zero old-task correct predictions at every noninitial task and seed. The standard-probe comparison therefore measures current adaptation at zero retained classification accuracy. ER's actual retention and its probe's current advantage belong to different models; a joint retention–adaptation assessment must score both outcomes on the same branch. Appendix D.1 reports all branches, seed values and aggregation, including the weighted and extra-current controls.

Task-aware training favors ER on both gaps. Separately trained task-aware models favor ER on both descriptive gaps, $\Delta A = -1.364$ and $\Delta H = -0.316$ pp. Task identity changes both their loss and evaluation. The next comparison isolates the evaluation change on the original class-incremental models.

5.1 Changing Prediction Candidates

We reevaluate 350 saved four-epoch endpoints from all 20 class-incremental histories, including fresh, probe, actual and compensated ER branches. A single saved logit array supplies the all-ten

and current-two predictions; model parameters are unchanged. As in the primary analysis, task zero is recorded but excluded from seed means.

Table 2 reports the test-only change for actual adaptation. The ER-minus-SGD actual-gap difference changes from $+11.151$ pp with all ten candidates to -2.829 pp with the current two candidates, with paired-seed SD 2.125 pp for the latter. ER actual accuracy rises from 69.395% to 83.375%. Its 13.980 pp gain consists entirely of correcting predictions assigned to old classes, with no rescued future-class predictions. The fresh, current-only probe, and SGD actual accuracies are unchanged by this mask on these recorded endpoints, so ΔH remains -0.971 pp. The ordering difference between actual and probe adaptation therefore appears with all-class prediction and disappears at the mean under the local candidate set.

Test candidates	SGD actual	ER actual	ΔA	ΔH
All ten classes	80.55	69.39	+11.15	-0.97
Current two classes	80.55	83.38	-2.83	-0.97

Table 2: Evaluation-only comparison on the same class-incrementally trained models. Entries are ten-seed means, each averaging four noninitial tasks. Actual accuracy is in percent and paired ER-minus-SGD gaps in percentage points. Only prediction candidates change, using the test example's task identity for the local mask.

The local score uses the test example's task identity and measures within-task discrimination. Its contrast with all-class prediction locates the observed mean ranking change in cross-task output competition at these fixed endpoints. We fixed its scope and aggregation before masked evaluation. Section 6 also shows that the mean mask reversal does not persist through the longer final-task continuation.

5.2 All-Candidate Calibration

A scalar offset reverses the mean current-accuracy ranking while trading current accuracy for old accuracy. This check fits one shared, unconstrained scalar offset b to the current two logits of each of the twenty actual final-task checkpoints, retaining all ten prediction candidates. The fit uses 200 previously trained-on images, including old classes, and measures the effect of this readout adjustment with access to those training images. Appendix C gives the complete fit, every seed, and all boundary solutions.

Adaptive bracketing places every optimum in the interior ($b \in [-11.954, 0.338]$). The current-class ER-minus-SGD difference changes from -17.040 to 20.560 pp, but both current accuracies fall while old accuracy rises (Appendix Table 6). The reversal comes from SGD's larger current-accuracy decrease.

An oracle scalar enumeration compares current accuracy under common minimum old-accuracy requirements. ER exceeds SGD throughout their common attainable range in 9 of ten pairs; one pair crosses (Appendix E). These test-derived curves characterize the retention–adaptation tradeoff available through fixed-model readout changes; selecting an offset for deployment requires separate data. Group-calibration curves also occur in generalized zero-shot learning (Chao et al., 2016).

5.3 Validation-Selected Endpoints on New Images

At the 20% retention requirement, selection yields validation-feasible endpoints for seven paired histories on CIFAR-10.1. This evaluation uses CIFAR-10.1 v6's 2,000 new images (Recht et al., 2018). Before model evaluation, we fixed one class-balanced split: 100 images per class for validation and 100 for test, shared by all seeds. From each final-task pre-state, we replay the same 628 current-only SGD updates and save thirteen fixed candidate steps. For each history, selection maximizes validation current accuracy subject to validation old accuracy at least q; ties take the earliest step and an empty set is unavailable. All selections are sealed before test inference. Selection uses old validation labels and all saved candidates from the completed trajectory.

The sparse probe set contains steps $\{0, 1, 2, 4\}$ epochs, whereas the dense set adds the fixed intermediate update checkpoints listed in Appendix O. The within-history 5 pp drop limit admits only

candidates whose validation old accuracy is no more than five percentage points below that history's step-zero old accuracy.

q	History / endpoint	n	Test below q	Test old	Test current
20	SGD selected	7/10	2/7	20.48 ± 1.64	1.36 ± 3.59
20	ER selected	10/10	0/10	30.74 ± 4.61	18.15 ± 14.39
20	ER actual (fixed)	10/10	0/10	27.40 ± 4.43	55.85 ± 10.71
10	SGD selected	10/10	0/10	18.66 ± 2.56	14.45 ± 12.80
10	ER selected	10/10	0/10	22.45 ± 8.18	36.20 ± 17.50
10	ER actual (fixed)	10/10	0/10	27.40 ± 4.43	55.85 ± 10.71

Table 3: CIFAR-10.1 test accuracy (%). Rows labeled SGD or ER are validation-selected endpoints: available counts validation-feasible histories, and test-below-q counts their violations. Fixed actual ER rows evaluate the same ten prespecified actual checkpoints at each q; their test-below-q count records fixed-checkpoint outcomes. Entries are mean $\pm$ sample SD over the displayed number of seed histories; infeasible histories do not enter a selected row. For $q = 20$, the paired ER-minus-SGD current difference on the seven common seeds is 17.71 pp (SD 16.18), rather than the difference between the two displayed marginal means.

Candidate eligibility changes the selected endpoint within a history. A sealed CIFAR-10.2 check adds steps through 16 (Appendix O). At $q = 20\%$, dense versus sparse ER probes gain 15.925 pp current and lose 5.994 pp old accuracy, with two test violations; admitting actual selects it in all ten ER histories. Adding a within-history 5 pp drop limit reduces the dense-probe gain to 2.425 pp and makes nine of the ten actual ER checkpoints ineligible. On the seven common seeds, this rule changes the dense ER-minus-SGD current contrast from 26.571 to 0.214 pp; the old-accuracy contrast remains 15.045 pp. These policies therefore select different old/current operating points from the same histories.

On CIFAR-10.1 at $q = 20\%$, seven seed pairs are jointly available, with ER-minus-SGD current difference 17.71 pp (SD 16.18); both SGD test violations remain included. This paired contrast includes differing initial and attained retention. Actual ER attains mean old/current accuracy $27.40/55.85\%$, meets the 20% test requirement in all ten seeds, and exceeds every corresponding selected endpoint in current accuracy. Thus trajectory selection trades away current accuracy relative to the corresponding actual ER checkpoint. Per-task scores also reveal old-task failures hidden by pooled retention (Appendix H).

6 Longer Adaptation Budgets

The four-task mean ranking disagreement persists through sixteen epochs; the final-task comparisons change with continued training. Continuations extend saved epoch-four endpoints to epochs eight and sixteen. The final-task check covers SGD, ER and DER++ actual/probe branches and their shared fresh branch (70 trajectories); the SGD/ER four-task check adds 120 continuations and reuses 40 final-task branches. Current order, learning rate and branch operation continue unchanged, while actual replay restores its full reservoir and random states. Both retain ten seeds and fixed endpoints (Appendices M and N).

For the four-task aggregate, ER–SGD all-ten actual differences are -11.151, -8.445 and -2.618 pp; probe differences are $+0.971$, $+0.233$ and $+0.595$ pp. Mean directions remain opposed at all three budgets, but positive probe differences occur in 9, 7 and 8 seeds, respectively. The current-two actual mean stays positive, reaching $+1.116$ pp at sixteen epochs (Table 30). The four-task mean ranking disagreement therefore persists through sixteen epochs, while the number of seeds with a probe advantage varies with budget.

Final-task comparisons differ from that aggregate. ER's probe difference ends at 0.445 pp, with five positive and five negative seeds; its current-two actual difference becomes -0.870 pp, removing the masked mean reversal. DER++–SGD actual differences change from -2.115 to 0.230 pp, so that pair's final-task actual/probe mean disagreement does not persist.

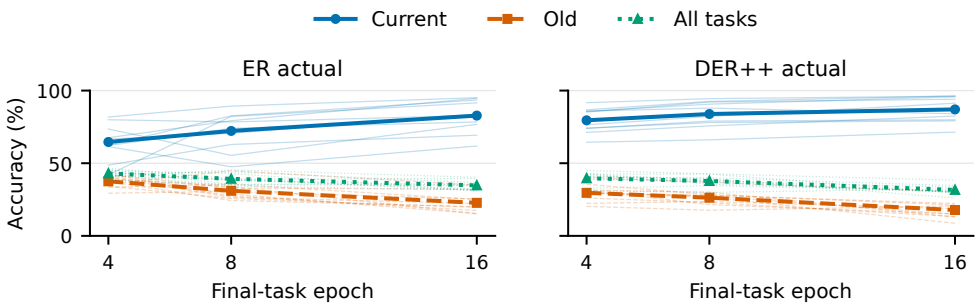


Figure 3: Final-task actual current gains accompany old/all-task losses. Thin curves show ten seeds and thick curves their means; each panel scores the same models over all ten candidates. Appendix Figure 14 gives every paired and candidate-set contrast; Table 26 gives absolute values and sample SDs.

Current gains accompany retention costs. Across the four-task aggregate, ER actual current accuracy changes by $+15.040$ pp from epochs four to sixteen, while old and seen-task accuracy change by -10.125 and -2.508 pp. Final-task old/all-task losses also occur (Figure 3). Because continued presentations enter the reservoir, its mean old-example fraction falls from 80.30% to 50.70%. Freezing the epoch-four reservoir isolates continued optimization: at epoch 16, ER reaches 81.625% current, 26.959% old and 37.892% all-seen accuracy. Relative to the evolving buffer, these are -1.155, +4.223, and +3.147 pp. Optimization drives most of the current gain; buffer evolution shifts the same-model current–retention tradeoff (Appendix L).

7 When the Measurements Agree

A full-budget comparison tests whether the opposing rankings extend to a different setting: CIFAR-100, ResNet-32 with group normalization, five paired seeds and ten ten-class tasks (He et al., 2016; Wu & He, 2018). Following pinned PyCIL schedules (Zhou et al., 2021), task 1 receives 200 epochs and later tasks 70. ER stores 20 examples per old class and pairs equal current/replay batches. Probe and fresh use the same 70-epoch current-only procedure at the final four boundaries; tasks are averaged within seed.

Endpoint	Current, seen	Current, ten	Old, seen
SGD actual / probe	64.39 ± 2.83	64.39 ± 2.83	0.00 ± 0.00
ER actual	71.07 ± 1.62	74.39 ± 1.40	14.69 ± 0.78
ER probe	77.68 ± 1.05	77.68 ± 1.05	0.00 ± 0.00
Fresh	61.74 ± 2.47	61.74 ± 2.47	0.00 ± 0.00

Table 4: Full-budget CIFAR-100 accuracy (%), mean $\pm$ sample SD across five paired seeds after averaging tasks 6–9. "Seen" uses all classes encountered so far; "ten" supplies current-task identity. Old accuracy belongs to the same endpoint as current accuracy.

All three measurements favor ER in every seed, by 6.675, 13.285 and 9.990 pp on average; no seed meets the frozen joint-direction criterion. ER actual retains 14.69% old-class accuracy, whereas current-only endpoints retain none. This experiment changes architecture, replay and budget jointly, so it establishes a boundary rather than attributing the reversal to duration. Appendix P gives all seed values and the protocol; Appendix F reports an opposing four-seed short-history memory-size comparison.

8 Discussion

New-task accuracy mixes adaptation, all-class competition and update-rule effects. Since current-only probes discard retention, deployment comparisons require old/current scores from one endpoint.

REFERENCES

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/b704ea2c39778f07c617f6b7ce480e9e-Abstract.html.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, 2016. URL https://schangpi.github.io/papers/eccv16_gzsl.pdf.

Dongkyu Cho, Taesup Moon, Rumi Chunara, Kyunghyun Cho, and Sungmin Cha. Forget forgetting: Continual learning in a world of abundant memory. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/17fc58b1ba92abbb72c5f457e140d851-Abstract-Conference.html.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Davari_Probing_Representation_Forgetting_in_Supervised_and_Unsupervised_Continual_Learning_CVPR_2022_paper.html.

Matthias De Lange, Gido M. van de Ven, and Tinne Tuytelaars. Continual evaluation for lifelong learning: Identifying the stability gap. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zy350cRstc6.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. doi: 10.1016/S1364-6613(99)01294-2.

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. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Timm Hess, Abhishek Jha, Gido M. van de Ven, and Tinne Tuytelaars. Forgetting, plasticity, and co-observation: a third facet of continual learning. In *Proceedings of the 5th Conference on Lifelong Learning Agents (CoLLAs)*, 2026. URL https://arxiv.org/abs/2608.18803.

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/05cdc7feee41e3572a9a3f4acb773891-Abstract-Conference.html.

Amogh Inamdar, Matthew So, Vici Milenia, and Richard Zemel. Re-evaluating continual learning with few-shot adaptation, 2026. URL https://arxiv.org/abs/2606.03843. Preprint.

Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity in continual learning via regenerative regularization. In *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 410–430. PMLR, 2025. URL https://proceedings.mlr.press/v274/kumar25a.html. CoLLAs 2024; proceedings published in 2025.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6109–6119, 2021. URL https://proceedings.mlr.press/v139/lee21e.html.

Lute Lillo and Nick Cheney. Activation function design sustains plasticity in continual learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9c8df8de46c1a1b39b30b9f74be69c02-Abstract-Conference.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. URL https://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf. Workshop paper.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211, 2023. URL https://proceedings.mlr.press/v202/lyle23b.html.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. In *Proceedings of The 3rd Conference on Lifelong Learning Agents (CoLLAs 2024)*, volume 274 of *Proceedings of Machine Learning Research*, pp. 750–783, 2025. URL https://proceedings.mlr.press/v274/lyle25a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10?, 2018. URL https://arxiv.org/abs/1806.00451.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022. doi: 10.1038/s42256-022-00568-3.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 374–382, 2019. URL https://arxiv.org/abs/1905.13260.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Yuxin_Wu_Group_Normalization_ECCV_2018_paper.html.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, and De-Chuan Zhan. PyCIL: A python toolbox for class-incremental learning, 2021. URL https://arxiv.org/abs/2112.12533. Toolbox preprint.

AI USE STATEMENT

AI tools assisted experiment implementation. Human authors formulated and reviewed the hypotheses and mathematical statements, reviewed the manuscript, and take responsibility for the content.

REPRODUCIBILITY STATEMENT

Appendix G gives protocols, environments, commands and the evidence map. Anonymous code and statistical evidence are included in the supplementary ZIP. The packaged commands reproduce the primary, trajectory, CIFAR-100, CBP and ResNet-32 statistics from compact raw evidence.

A Discovery Study and Selection of the Follow-up

The discovery matrix comprises three seeds, three algorithms (SGD, ER, and DER++), and four streams: target remapping, permuted MNIST, and the two CIFAR-10 settings. All histories contain five tasks with four epochs each; adaptation prefixes one, two, and four share the same pre-task states. The MNIST experiments use 6,000 training examples and the official 10,000-example test split. CIFAR-10 uses 4,000 official training examples per two-class task and all 2,000 corresponding official test examples. CIFAR class order is fixed in discovery. Target remapping reuses inputs with changing labels, producing a contradictory target stream.

Table 5 and Figure 4 show descriptive prefix-four results. The class-incremental ER-minus-SGD differences are $(\Delta A, \Delta H) = (+1.350, -1.604)$ pp. Their opposite directions motivated the primary follow-up comparison after the discovery results were observed. The follow-up uses a separate confirmatory sample.

Stream	Algorithm	A	H	C
Target remapping	SGD	-2.00 ± 0.29	-2.00 ± 0.29	$+0.00 \pm 0.00$
	ER	-0.10 ± 0.03	-1.92 ± 0.12	$+1.82 \pm 0.14$
	DER++	$+0.98 \pm 0.37$	-2.33 ± 0.16	$+3.31 \pm 0.46$
Permuted MNIST	SGD	-2.69 ± 0.33	-2.69 ± 0.33	$+0.00 \pm 0.00$
	ER	-2.22 ± 0.34	-2.71 ± 0.34	$+0.48 \pm 0.03$
	DER++	-3.09 ± 0.40	-2.97 ± 0.36	-0.12 ± 0.05
CIFAR-10 class-IL	SGD	-2.58 ± 1.30	-2.58 ± 1.30	$+0.00 \pm 0.00$
	ER	-1.23 ± 3.50	-4.18 ± 2.07	$+2.95 \pm 1.75$
	DER++	-5.59 ± 2.98	-6.00 ± 1.71	$+0.41 \pm 1.32$
CIFAR-10 task-aware	SGD	-0.33 ± 1.75	-0.33 ± 1.75	$+0.00 \pm 0.00$
	ER	-0.83 ± 1.80	-0.67 ± 1.92	-0.16 ± 0.20
	DER++	-3.68 ± 2.18	-2.23 ± 1.86	-1.46 ± 0.89

Table 5: Discovery at prefix four: mean $\pm$ sample SD over three seeds, in percentage points. Each seed first averages tasks one through four. SD summarizes seed variation in this exploratory study.

Mean historical gaps are negative in every discovery group across all prefixes. Under these short protocols, historical states therefore adapt better on average than their fresh references. This pattern motivates examining how the adaptation operation affects method comparisons within these short histories. The follow-up hypothesis was selected from these exploratory comparisons, then evaluated under the separately frozen protocol.

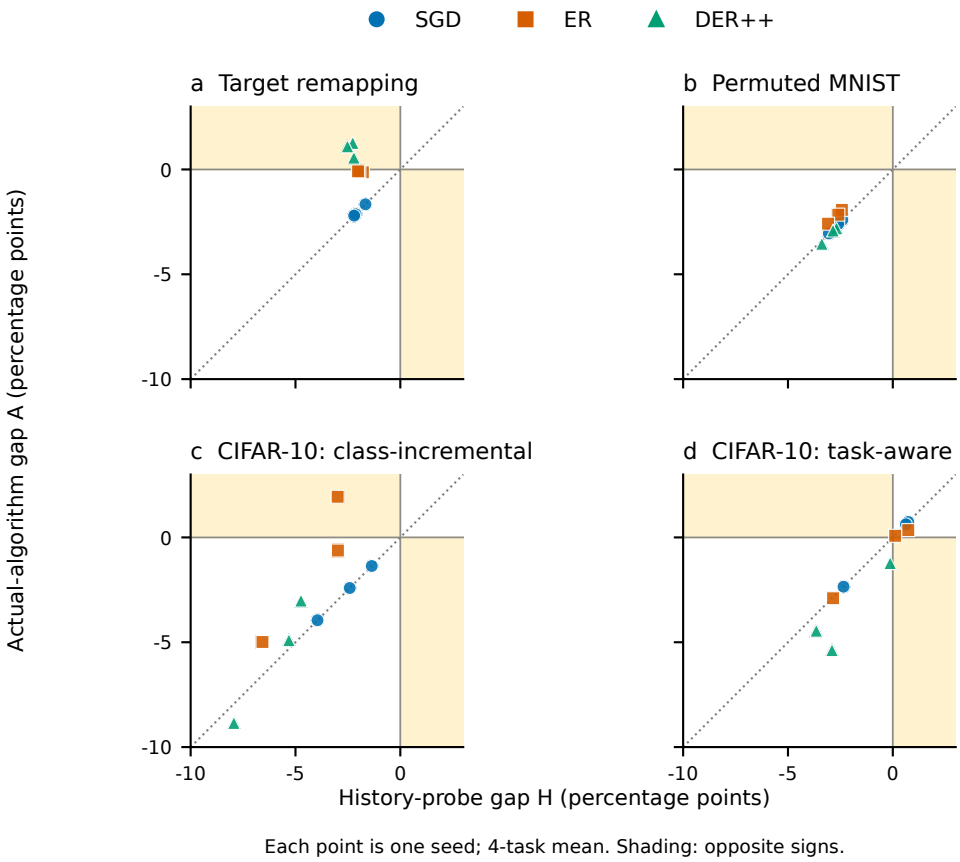


Figure 4: Discovery at prefix four. Each point is one seed averaged over the four noninitial tasks. Coordinates are historical-state gaps H and actual gaps A, in percentage points. The diagonal denotes $A = H$, and shaded quadrants mark opposite signs. Three seeds are shown per algorithm. Negative H indicates an advantage over the specified fresh reference.

B Discovery Budget Sensitivity

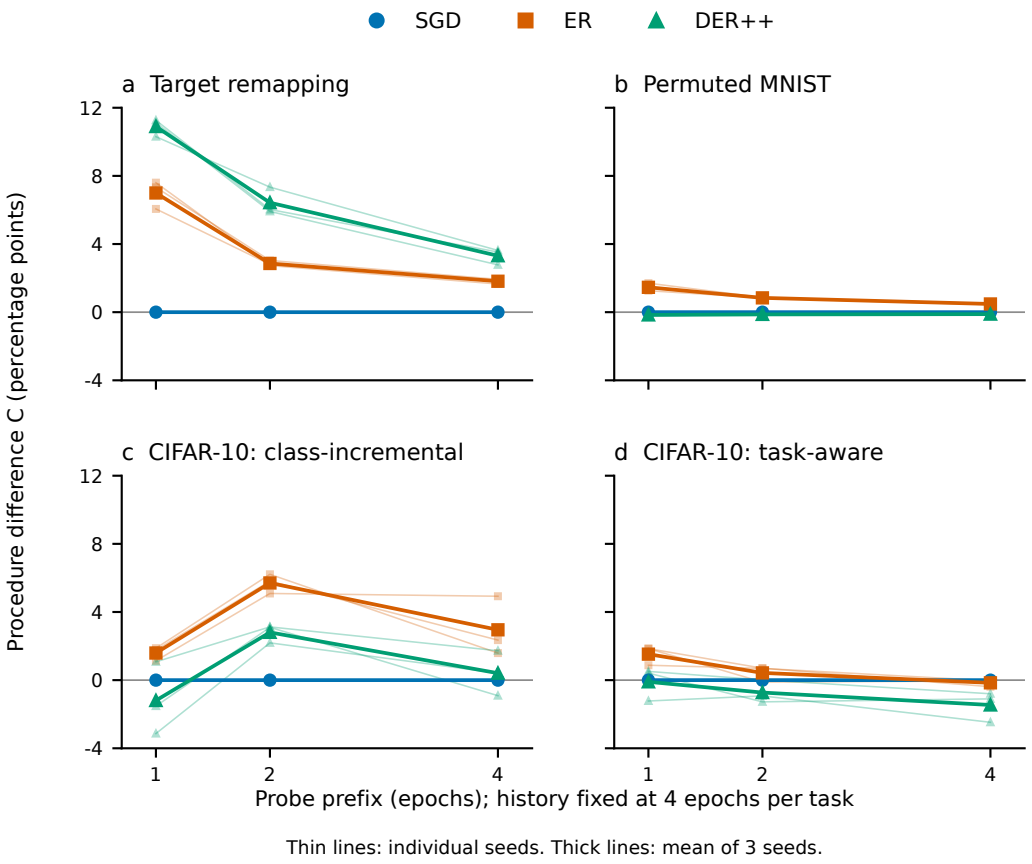


Figure 5: Discovery sensitivity to adaptation prefix. All prefixes share the same pre-task states from four-epoch-per-task actual histories. Thin lines trace individual seeds and thick lines their mean operator contrast C across prefixes. Prefixes are repeated observations of each run, and C can take either sign.

C Scalar Calibration Details

The diagnostic was specified after the follow-up and candidate-mask results were known. It uses seeds 201–210, ER and SGD, final task index 4 only. Calibration images are the twenty smallest training indices of each original class, then globally sorted. Training and test inputs use the original normalization and each history's archived label mapping. Hyperparameters are fixed before evaluating the calibration outcomes on test data.

Method	Offset	Old	Current	All
ER	Zero	37.55 ± 4.35	64.62 ± 12.51	42.96 ± 2.46
ER	Fitted	44.28 ± 3.73	46.69 ± 9.86	44.76 ± 1.98
SGD	Zero	0.00 ± 0.00	81.66 ± 11.54	16.33 ± 2.31
SGD	Fitted	18.82 ± 2.11	26.12 ± 10.79	20.28 ± 1.88

Table 6: Final-checkpoint calibration accuracy (%), mean $\pm$ sample SD over ten paired seeds. Each model evaluates all 10,000 official test examples: 8,000 old and 2,000 current. "All" weights old/current groups by 0.8/0.2. The calibration fit uses training images only; no model weights change.

For calibration logits z_i, define $\mathrm{LSE}(v) = \log \sum_j \exp(v_j)$ and $d_i = \mathrm{LSE}(z_{i,8:10}) - \mathrm{LSE}(z_{i,0:8})$. The derivative of calibration cross entropy is $g(b) = 200^{-1} \sum_i \sigma(d_i + b) - 0.2$. Equal sampling per class makes the current-group fraction 0.2. Starting from $[-1, 1]$, we double either endpoint until the derivative has opposite signs, then run 100 bisection iterations. All twenty solutions are interior; the three SGD fits formerly clipped at -10 move to -11.954, -10.078, and -10.387. Thus the fitted comparison is free of the previous search-boundary artifact. The calibration logits and fit are saved before test forward computation. Argmax ties select the smallest global class index. The same offset applies to all test examples, irrespective of their true group.

	ER			SGD		
Seed	b	Δ old	Δ current	b	Δ old	Δ current
201	-1.359	+7.85	-18.15	-8.464	+17.65	-36.30
202	-1.384	+8.64	-22.70	-11.954	+18.34	-69.65
203	-0.725	+4.16	-13.75	-9.494	+22.70	-55.65
204	-0.786	+4.09	-16.10	-10.082	+18.31	-65.60
205	0.338	-1.64	+8.90	-8.782	+15.64	-51.30
206	-1.434	+11.16	-30.70	-9.175	+21.95	-56.95
207	-1.476	+9.75	-24.55	-8.454	+19.21	-52.10
208	-0.754	+5.88	-20.70	-9.257	+17.94	-58.90
209	-1.767	+11.95	-25.65	-10.393	+17.36	-50.40
210	-0.982	+5.41	-15.90	-9.702	+19.06	-58.45

Table 7: Every seed's fitted offset and accuracy change in percentage points (fitted minus zero offset). Numerical means and SDs are descriptive; seeds, not individual examples, are the units of comparison. Full counts, per-class outcomes, and logits are retained in the reproducibility artifacts.

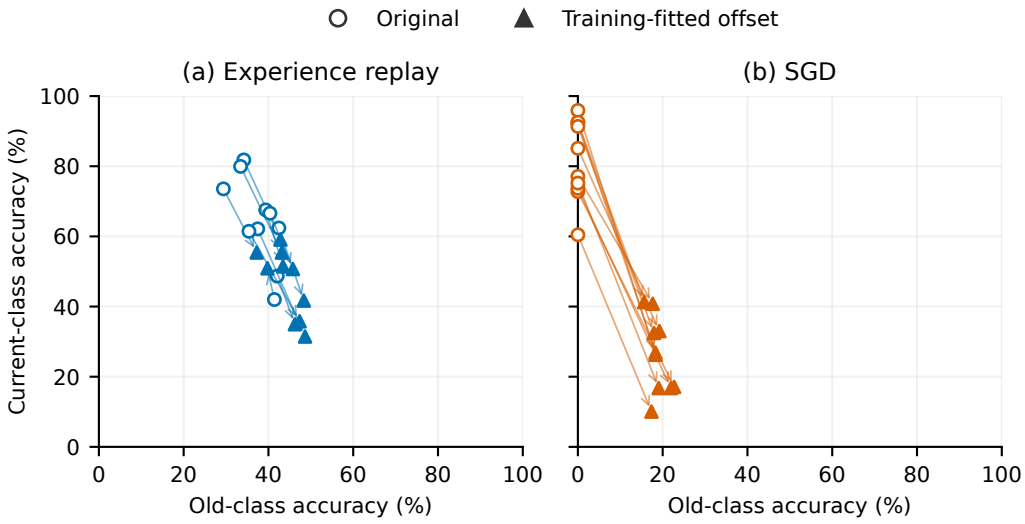


Figure 6: All-candidate accuracy before (circles) and after (triangles) training-fitted scalar calibration. Each arrow connects the same fixed checkpoint, with ten seeds per method; points can overlap. Both axes use 0–100% scales. Movement toward higher old-class accuracy is generally accompanied by lower current-class accuracy. Arrows display the paired change at each fixed checkpoint.

D Additional Protocol, Evidence, and Operator Controls

Secondary measurements. Prefixes one and two, the task-aware setting, and four leave-one-task-out averages remain secondary; leave-one-task-out aggregation is still performed within each seed. Sign-disagreement rates use margins $\delta \in \{0, 0.5, 1, 2, 5\}$ percentage points. At zero, signs are strict, so zero is not counted as disagreement; at positive margins we count $A \geq \delta$, $H \leq -\delta$, or the reverse. Operator-specific contrasts compare $\bar{H}^o_{\mathrm{ER}} - \bar{H}^o_{\mathrm{SGD}}$.

Let $R_{t,j}$ be actual accuracy on task j after completing all four epochs of task t. Final accuracy is $\frac{1}{5}\sum_j R_{4,j}$; forgetting is $\frac{1}{4}\sum_{j=0}^{3}(\max_{u=j,\ldots,4} R_{u,j} - R_{4,j})$. Forward transfer compares pre-task accuracy with the untrained fresh accuracy on that same task, then averages tasks one through four; the reference is measured initialization performance (Lopez-Paz & Ranzato, 2017). These full-history metrics are distinct from post-adaptation gaps and from a shorter-history campaign.

Complete primary seed values and existing prefixes. Primary resampling uses PCG64 with seed 2026090901 to generate the shared 100000×10 index matrix. Each endpoint uses the 1.25th and 98.75th percentiles with linear interpolation. Table 8 gives exact four-task seed averages for the class-incremental primary endpoint. Each displayed accuracy equals the sum of four integer correct counts divided by 80 (four tests of 2,000 images). Thus the four decimal places preserve the exact input to paired resampling. The package files `paper_r9/generated/followup_complete_seed_prefix.csv` and `paper_r9/generated/followup_integer_counts.csv` contain all 30 seed/prefix vectors and their 720 branch/task counts respectively. Figure 7 presents the existing frozen prefixes. Only prefix four uses the primary interval rule; prefixes one and two are descriptive. All use the same history trained four epochs per preceding task, with the seed as the statistical unit.

Seed	F	a_{SGD}	a_{ER}	p_{SGD}	p_{ER}	ΔA	ΔH
201	81.4875	83.4250	77.7750	83.4250	84.4000	5.6500	-0.9750
202	84.6250	91.6500	79.2125	91.6500	91.5750	12.4375	0.0750
203	78.1000	80.3500	67.1750	80.3500	81.1625	13.1750	-0.8125
204	76.6625	79.7875	71.5375	79.7875	81.0500	8.2500	-1.2625
205	76.0125	80.6375	52.8375	80.6375	81.5500	27.8000	-0.9125
206	73.1125	75.0250	67.3625	75.0250	75.5875	7.6625	-0.5625
207	82.3875	84.5750	75.0375	84.5750	84.8875	9.5375	-0.3125
208	77.6125	80.3500	65.1125	80.3500	82.8250	15.2375	-2.4750
209	70.5125	74.0125	67.2125	74.0125	74.7375	6.8000	-0.7250
210	70.9250	75.6500	70.6875	75.6500	77.4000	4.9625	-1.7500

Table 8: All ten primary paired seeds, averaged over four noninitial tasks. Accuracies are percentages and gaps are percentage points. Shared F cancels from the ER-minus-SGD differences: $\Delta A = a_{\mathrm{SGD}} - a_{\mathrm{ER}}$ and $\Delta H = p_{\mathrm{SGD}} - p_{\mathrm{ER}}$. Seed 202 retains its positive ΔH; seed 205 retains its large ΔA. No seed is excluded.

Provenance and training counts. The protocol requires a manifest fixing configurations, source files, analysis code, and protocol hashes before the follow-up. Completion checks require task records, integer accuracy numerators and denominators, model and memory checkpoints, and artifact hashes. Training counts distinguish current examples, replay examples, extra-current examples, and updates. Full-history totals include task zero; prefix totals cover tasks one through four and are reported separately for each branch. Evaluation counts are recorded separately. Group summaries require complete, consistent run records; failed or incomplete records are preserved for investigation.

CIFAR-100 execution checks. All eight histories completed in 30.82 minutes within the 55-minute watchdog. The independent auditor reconstructs all 47,400 update events, verifies checkpoint tensor hashes and inventory, and recomputes counts and seed statistics. A functional network independently replays fixed seed 1001 under both memories: six checkpoints, 26,000 image evaluations (including repeated images across branches) and 2.6 million logits agree exactly. This functional check covers those six endpoints.

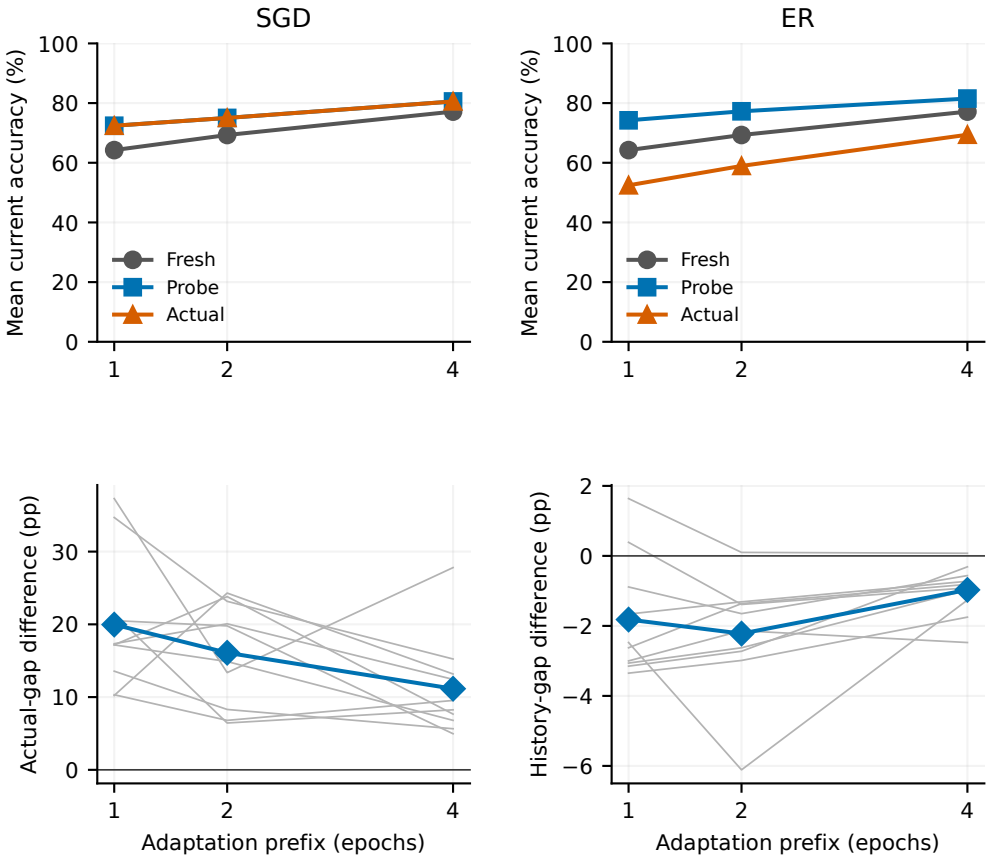


Figure 7: Existing class-incremental follow-up prefixes. Top: ten-seed mean fresh, probe and actual current-task accuracies, with shared 0–100% axes; SGD actual and probe coincide by construction. Bottom: all ten seed curves (thin gray) and their mean (blue diamonds) for ER-minus-SGD gap differences. The two gap panels have different vertical ranges. Connecting lines indicate the sampled endpoints on the same branch.

Comparison	Starting point and adaptation	Candidates and retention target
Inamdar et al. (2026)	Post-task checkpoints; 0–10 training examples and 10 adaptation steps on past or future tasks.	Task-/domain-incremental settings; backward and forward adaptation evaluated separately.
This CIFAR-10 study	Same pre-task histories; actual, matched current-only, and fresh branches at 1/2/4 epochs.	All ten outputs; a separate same-logit task mask. Old and current endpoints refer to the same adapted model.
Cho et al. (2026)	Replay allocation and fresh retraining with available seen-task data.	Seen-task continual-learning performance; our current-only fresh branch answers a different question.
Chao et al. (2016); Wu et al. (2019)	Group offsets on trained scores; BiC learns affine correction using held-out examples.	Group accuracy trade-offs / incremental classifier bias; the present scalar diagnostic uses previously trained-on data.

Table 9: Direct predecessors and the paired question examined here. Adaptation-dependent method assessment is established by Inamdar et al. Here the actual–probe mean ranking disagreement also disappears when only the prediction candidates change, with probe scores unchanged. This joint observation connects the adaptation comparison to fixed-model output competition; old/current pairs describe its retention costs.

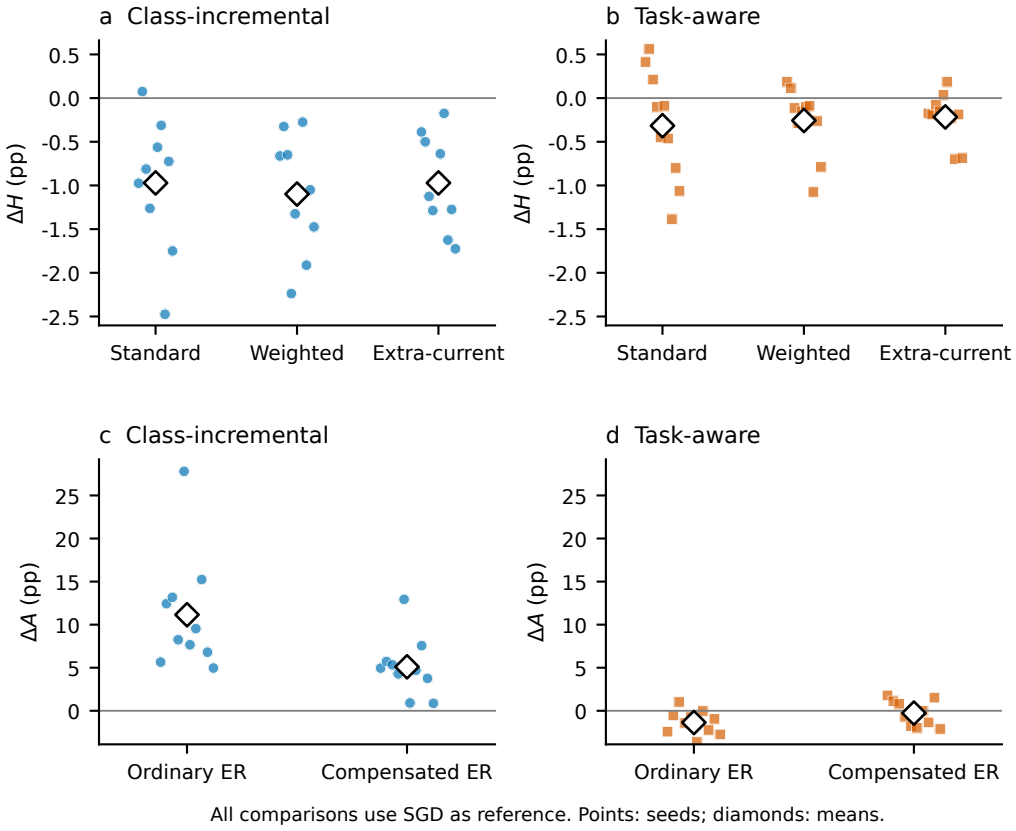


Figure 8: Operator sensitivity at prefix four. Top: historical-gap differences between ER and SGD for the standard, weighted current-only, and extra-current operators, each with its own fresh/probe pair. Bottom: actual-gap differences using ordinary or compensated ER against standard SGD. Compensation changes only a branch from the saved ER state; it does not replace the ER history. Points show all ten paired seeds and diamonds their means. Panels in each row share a vertical range. These secondary comparisons have no inferential intervals.

D.1 Old and Current Performance of the Same Adapted Model

This post-hoc diagnostic fixes all twenty class-incremental histories, ten paired seeds, noninitial tasks 1–4 and the final four-epoch endpoint. It evaluates all 680 saved checkpoints from the actual, standard probe, weighted probe, extra-current probe, each matching fresh branch, the pre-adaptation state, and ER's compensated branch. No training or task mask is applied. Every evaluation retains all ten output candidates, including future classes. The same endpoint is evaluated on each old task and its current task, providing matched old/current measurements.

Let $c^b_{t,j}$ be correct predictions among 2,000 test examples of task j from branch b at task t. Its old score is $100\sum_{j<t} c^b_{t,j}/(2000t)$, current score $100c^b_{t,t}/2000$, and seen score $100\sum_{j\le t} c^b_{t,j}/(2000(t+1))$. For each seed we take the equal mean across $t = 1, 2, 3, 4$, then report the mean and sample SD across ten seeds. Because the number of old tasks changes, the displayed seen mean cannot be reconstructed by applying one fixed old/current weight to the two displayed means. Before-adaptation scores describe the shared starting state of the adaptation branches.

History	Endpoint	Old	Current	Seen
SGD	Before adaptation	41.883 ± 2.197	0.000 ± 0.000	25.777 ± 1.153
SGD	Actual	0.000 ± 0.000	80.546 ± 5.224	25.644 ± 1.711
SGD	Standard probe	0.000 ± 0.000	80.546 ± 5.224	25.644 ± 1.711
SGD	Standard fresh	0.000 ± 0.000	77.144 ± 4.760	24.539 ± 1.719
SGD	Weighted probe	0.001 ± 0.003	81.784 ± 4.607	26.092 ± 1.540
SGD	Weighted fresh	0.000 ± 0.000	77.609 ± 5.063	24.810 ± 1.729
SGD	Extra-current probe	0.000 ± 0.000	82.586 ± 4.597	26.361 ± 1.435
SGD	Extra-current fresh	0.000 ± 0.000	79.689 ± 5.567	25.532 ± 1.740
ER	Before adaptation	61.404 ± 5.506	0.000 ± 0.000	40.193 ± 3.483
ER	Actual	43.560 ± 8.035	69.395 ± 7.535	51.535 ± 3.655
ER	Standard probe	0.000 ± 0.000	81.517 ± 4.955	25.879 ± 1.644
ER	Standard fresh	0.000 ± 0.000	77.144 ± 4.760	24.539 ± 1.719
ER	Weighted probe	0.005 ± 0.012	82.883 ± 4.311	26.382 ± 1.445
ER	Weighted fresh	0.000 ± 0.000	77.609 ± 5.063	24.810 ± 1.729
ER	Extra-current probe	0.000 ± 0.000	83.558 ± 4.377	26.610 ± 1.374
ER	Extra-current fresh	0.000 ± 0.000	79.689 ± 5.567	25.532 ± 1.740
ER	Compensated ER	34.252 ± 8.264	75.443 ± 5.111	47.106 ± 4.273

Table 10: All same-model old/current/seen accuracies (%), mean ± sample SD over ten seed units. Each unit averages four noninitial task boundaries. All outputs remain eligible. Standard and extra-current probes attain zero old-task correct predictions in every seed and task; the weighted probes retain only a few correct predictions. No inferential intervals are added.

The complete evaluation took 176.922 seconds under a 55-minute watchdog, with 4,760,000 image evaluations (repeated images across endpoints). Independent NumPy reconstruction checks official label mapping, indices, argmax counts and task/seed aggregation. A separately written functional CNN replays fixed seed 201, both histories, all 17 task-four branches: 1.7 million logits match exactly in this forward-output check. The supplement includes `paper_r9/generated/probe_retention_seed_values.csv` (170 seed/branch units), `probe_retention_task_values.csv` in the same folder (680 task/branch units), and the aggregate integer counts. The completed functional check used checkpoints and raw logits that are not bundled. These scores measure direct old-task prediction after the specified operation.

D.2 Matched Controls for the Adaptation Operation

Let n_c and n_r denote the current and replay batch sizes at an update. Writing mean cross-entropies as ℓ_c and ℓ_r, actual ER minimizes

$$\ell_{\mathrm{ER}} = w_c\ell_c + (1 - w_c)\ell_r, \qquad w_c = \frac{n_c}{n_c + n_r}. \tag{5}$$

On both SGD and ER histories, a *weighted current-only* control minimizes $w_c\ell_c$ without reading replay samples. Its virtual reservoir length follows the current-example presentations in the complete

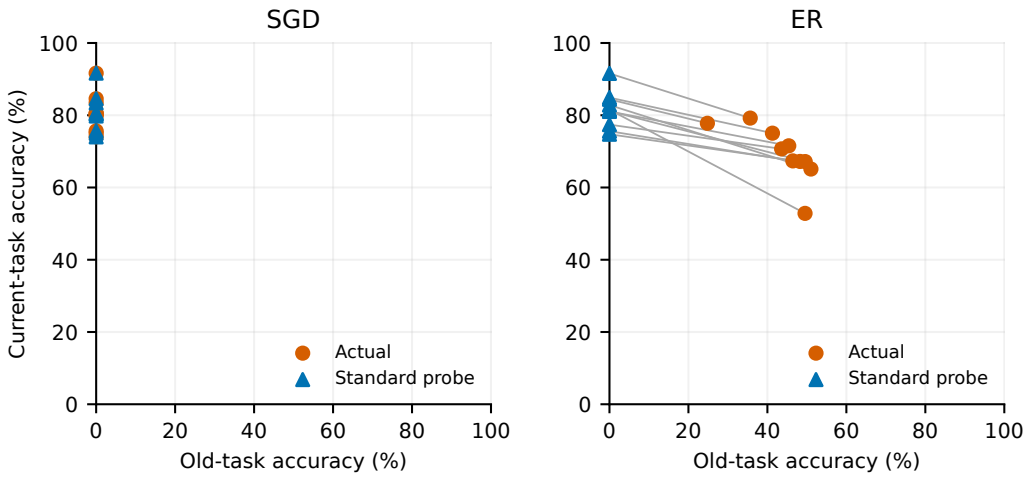


Figure 9: Old/current accuracies of actual and standard-probe branches. A line connects the two endpoints from the same seed's history, each averaged over four noninitial tasks. The two branches start from the same pre-task state. SGD endpoints coincide by construction. Both axes span 0–100%. ER's current-only probe has higher current accuracy and zero old accuracy.

preceding history and the current branch. An *extra-current* control replaces the n_r replay examples with draws, with replacement, from the current task's training set and uses their concatenated mean cross-entropy with the main batch. Its separate deterministic sampling stream is shared by its fresh and probe branches. It matches ER's update count and labeled forward-example count.

Each alternative operator has its own fresh and probe branches. For $o \in \{\text{weighted}, \text{extra}\}$, we use $H^o = F^o - p^o$, $A^o = F^o - a$, and $C^o = p^o - a$, retaining the standard actual endpoint a and reporting the absolute reference accuracies for each operator. A separate ER-only sensitivity branch copies the pre-task parameters, reservoir, and reservoir random states and minimizes $\ell_c + (1 - w_c)\ell_r$. Its coefficient sum is $2 - w_c$, so it changes both relative weighting and total gradient scale. The reviewer-requested normalized control in Appendix L divides this objective by $2 - w_c$. Standard fresh/probe references apply.

E Complete Readout Envelopes Under Old-Class Accuracy Requirements

The two training-fitted offsets optimize cross entropy. To compare readouts under a common old-class accuracy requirement, we conduct a separate post-hoc analysis of the same twenty final-checkpoint test-logit arrays. This is an oracle description of fixed model outputs on the already observed test set; offsets are chosen using test labels.

Let $O_m(b)$ and $N_m(b)$ be the numbers of correct old and current test examples after adding b to the current logits of checkpoint m. Define

$$U_m(k) = \max_{b\,:\,O_m(b)\geq k} 100N_m(b)/2000, \qquad k = 0, \ldots, \max_b O_m(b).$$

The old denominator is 8,000. A common minimum requirement allows different realized old accuracies. Every unique value of $\max z_{\text{old}} - \max z_{\text{current}}$ defines a switch: equality predicts an old class because old output indices are smaller. All tied samples move together. We enumerate these finite states over all real offsets, retain integer counts, and use no convex hull, interpolation, or finite bias bound. Unlike the earlier bounded CE fit, this analysis is invariant to a common positive rescaling of a checkpoint's logits in exact arithmetic.

Each paired curve is restricted to the smaller of the two attainable old maxima. We keep the complete method-specific support in Figure 11; inaccessible requirements remain undefined. The mean difference uses all ten seeds only, on their common support ending at 21.613% old accuracy. We summarize the complete curves descriptively. Offset sweeps and two-group accuracy curves are established in calibrated stacking (Chao et al., 2016); our application compares old/current supervised classes.

Seed	ER old max	SGD old max	Min gap	Max gap	Relation on paired common support
201	43.70	24.12	8.15	80.05	ER dominates
202	47.86	25.02	0.50	90.90	ER dominates
203	51.88	30.27	3.65	69.80	ER dominates
204	46.76	22.74	1.30	87.25	ER dominates
205	46.44	23.20	-0.75	82.70	Crossing
206	53.92	25.90	3.45	73.30	ER dominates
207	49.24	24.39	1.80	85.80	ER dominates
208	53.54	21.61	1.50	87.40	ER dominates
209	51.26	23.45	7.85	67.10	ER dominates
210	50.90	26.35	5.95	79.10	ER dominates

Table 11: Complete paired readout comparison. Old maxima are percentages; minimum and maximum ER-minus-SGD current gaps are percentage points over each seed's entire common support, including zero requirement. Relation describes each empirical pair on its common support. Seed 205's crossing is included in the table and mean curve.

Nine pairs favor ER throughout common support; one crosses. For those nine pairs, the ordering extends beyond the two CE operating points to every attainable common requirement. The attainable old maxima characterize the range of this single scalar readout at the fixed final-task checkpoints.

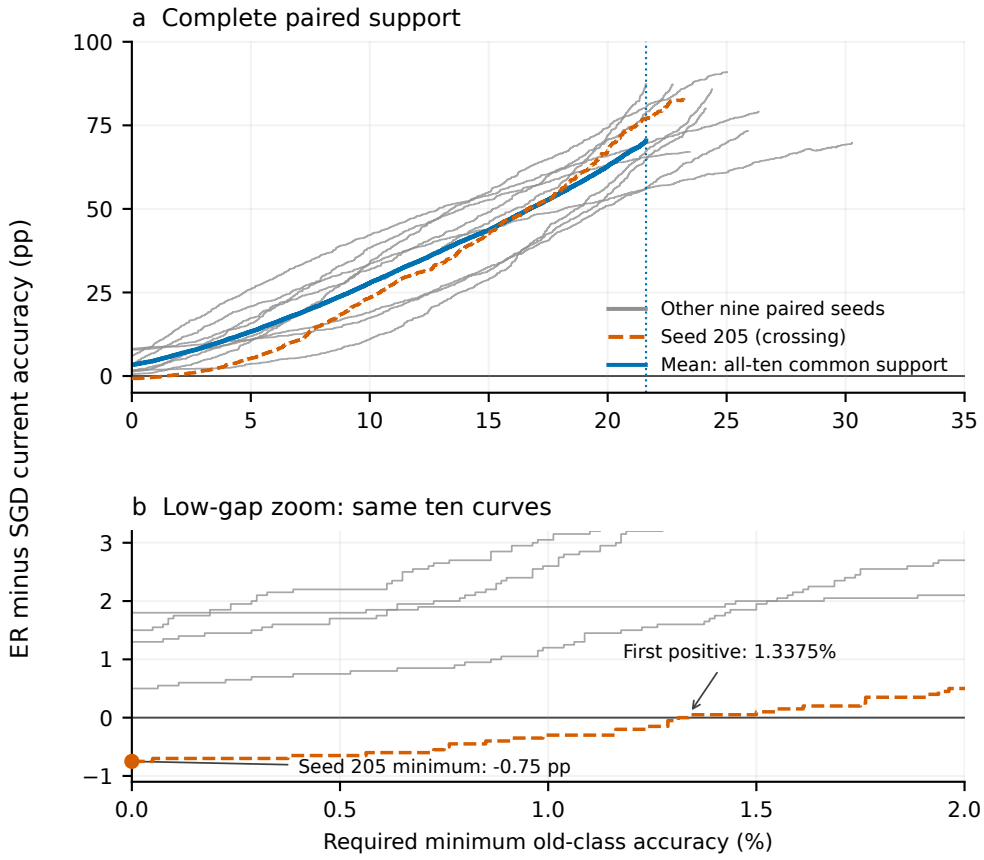


Figure 10: All ten ER-minus-SGD current-accuracy differences under common minimum old-class accuracy requirements. (a) Complete paired supports: gray curves show the other nine seeds and orange dashes highlight seed 205; each stops at its own paired support. The blue mean stops at the all-ten support (dotted line). (b) An explicitly cropped low-gap view repeats all ten seed curves over 0–2% old requirements and -1.1–3.2 pp differences, revealing seed 205's negative region and crossing. Curves outside that window remain visible in (a). Step connections display integer requirements, with no performance interpolation or confidence band.

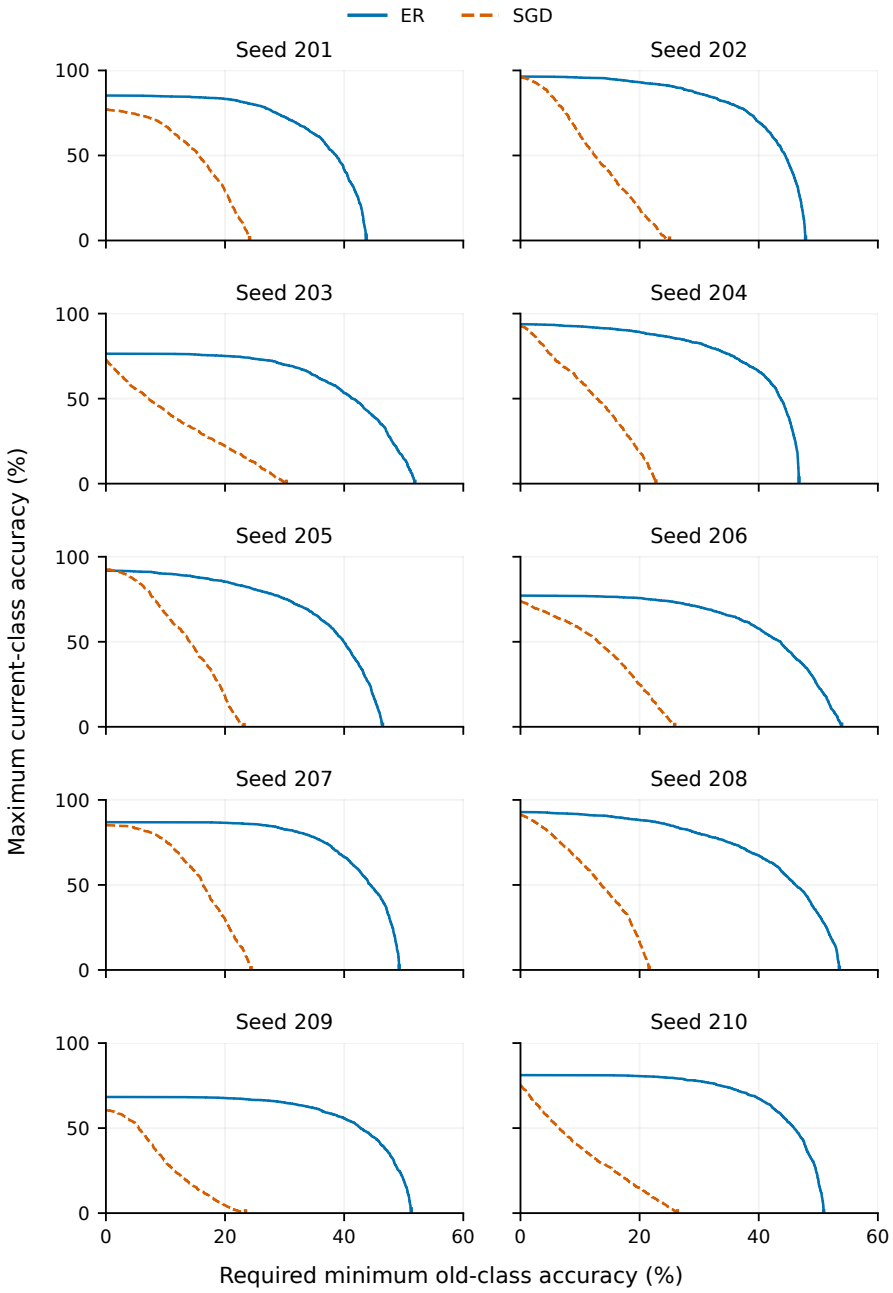


Figure 11: Complete empirical readout envelopes for every paired seed. ER is solid blue; SGD is dashed orange. Each curve stops at its largest attainable old-class requirement. Both methods use the same 10,000 test images and all ten output candidates. These are oracle fixed-output curves computed using test labels.

F CIFAR-100 Short-History Protocol and Complete Paired Values

The earlier short implementation violated its seen-output training protocol and omitted required artifacts. Its results are excluded and its test access is disclosed. The separately frozen repair is a descriptive short-history study conducted after that test exposure.

Data and update schedule. We use the official 50,000 training and 10,000 test images with fine labels. For each seed $s \in \{1001, 1002, 1003, 1004\}$, PCG64$(s + 40000)$ permutes 100 classes into ten disjoint tasks of ten classes, consistently mapped to output rows. Each task has 5,000 training and 1,000 test images. The model is the pinned PyCIL CIFAR ResNet-32 (Zhou et al., 2021) with all batch-normalization layers replaced by affine eight-group normalization, $\epsilon = 10^{-5}$, and a 100-way head: 470,004 parameters. It has option-A subsampled/zero-padded shortcuts. The shared CPU initialization seed is $s + 10000$.

The first task uses 20 epochs, the other nine use five each. Each epoch contains 79 updates over current examples, batch 64 with a final batch of 8. SGD uses learning rate 0.1, momentum 0.9, coupled weight decay 2×10^{-4}, no scheduler, no Nesterov, and a fresh optimizer at every task or auxiliary branch. Training cross-entropy uses only the outputs seen through the current task. At the start of each task, newly active head rows are reset to their paired initialization; inactive rows do not enter the preceding softmax.

Memory contains 20 or 200 images per old class, selected by a seed-specific random permutation of the official training indices for each class. The smaller memory is a strict subset of the larger one. Every old memory image is replayed exactly once per current epoch, with proportional integer allocation across updates. Current order and per-image augmentation keys are paired across memory sizes. This changes memory contents, total example exposure and the coefficient of the current loss together. It is therefore a joint protocol comparison, not an isolated memory-capacity effect. At task nine the old memories contain 1,800 and 18,000 examples respectively.

Images receive four-pixel zero padding, a random 32×32 crop, horizontal flip with probability 0.5, and brightness jitter $1 \pm 63/255$; normalization uses means $(.5071, .4867, .4408)$ and standard deviations $(.2675, .2565, .2761)$. All runs use deterministic FP32 CUDA without AMP/TF32. Private per-image generators and one-batch prefetch preserve the specified augmentation draws; exact dependencies and engineering witnesses are archived.

Branches, endpoint, and reporting. Only actual adaptation advances the ten-task history. At task nine, a current-only probe starts from the actual post-reset/pre-update checkpoint, and a fresh branch starts from the same seed's full initialization. Both run five epochs, with identical current events and all 100 outputs in the loss. Thus each history has 65 actual epochs and 10 auxiliary epochs, 5,925 updates; the complete eight-history matrix has 47,400 updates. Fresh parameters agree across memory sizes at every saved epoch, and its final logits agree exactly. We report both all-100 and current-ten argmax scores from the same endpoint logits. The latter supplies task identity.

The experimental unit is the paired seed, with four values, their mean and sample standard deviation. With only four seeds, these results are a descriptive boundary diagnostic and carry no inferential claim. All runs follow the frozen duration and endpoints. Table 12 gives every seed; the complete machine-readable table includes both candidate sets and all-task accuracy. This last-task endpoint differs from the four-task average used in the CIFAR-10 primary comparison.

Seed	a_{20}^{all}	a_{200}^{all}	a_{20}^{cur}	a_{200}^{cur}	p_{20}^{all}	p_{200}^{all}	F^{all}
1001	30.60	22.50	30.70	38.00	29.00	39.70	12.20
1002	43.30	35.30	43.30	50.90	48.20	55.10	10.90
1003	55.00	35.90	56.10	62.70	48.30	59.90	29.90
1004	51.10	35.40	51.60	57.90	42.30	53.40	26.50

Table 12: All four paired CIFAR-100 seeds. a, p, and F denote actual, current-only probe, and fresh accuracy (%). Superscripts distinguish all-100 and task-identified current-ten candidates; subscripts give memory examples per old class. Raw current-ten actual values are shown rather than a difference that cannot be derived from adjacent all-100 columns.

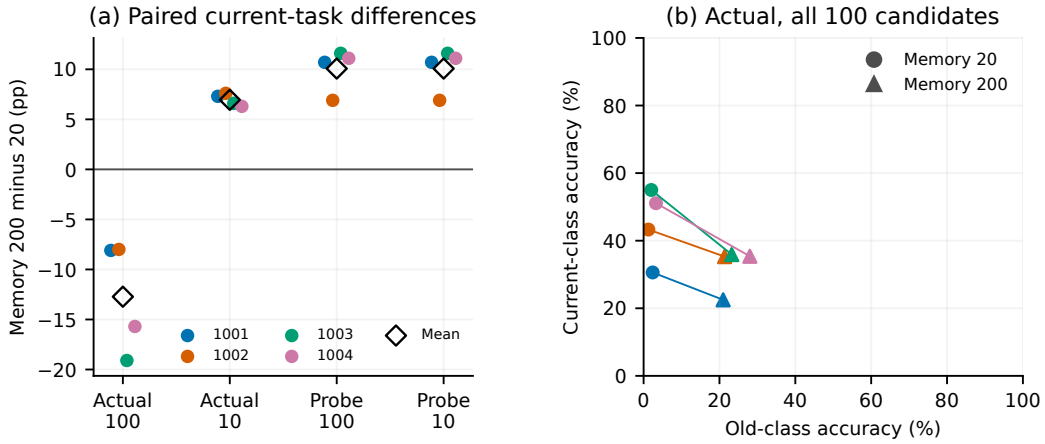


Figure 12: Complete four-seed CIFAR-100 comparison. (a) Current-task differences for actual and probe endpoints, with 100 or current-ten candidates; diamonds are means, with no inferential bars. (b) Actual all-candidate old/current accuracy: each line connects memory 20 (circle) to memory 200 (triangle) for one seed. Both axes in (b) span 0–100%. Color identifies the same seed in both panels. More memory improves retention while lowering current all-candidate accuracy in these runs.

Endpoint	Memory 20	Memory 200	Paired difference
Actual, 100 candidates	45.00 ± 10.76	32.27 ± 6.52	-12.73 ± 5.57
Actual, current 10	45.42 ± 11.16	52.38 ± 10.74	6.95 ± 0.60
Probe, 100 candidates	41.95 ± 9.08	52.03 ± 8.67	10.08 ± 2.15
Probe, current 10	41.95 ± 9.08	52.03 ± 8.67	10.08 ± 2.15

Table 13: Short-history CIFAR-100 final-task accuracy (%), mean $\pm$ sample SD across four paired seeds; memory-200 versus memory-20. Current-ten evaluation supplies task identity. These are separate short-history runs.

G Exact Implementation and Reproduction Specifications

CIFAR-10 numerical specification. Convert RGB bytes to FP32, divide by 255, then subtract channel means $(.4914, .4822, .4465)$ and divide by standard deviations $(.2470, .2435, .2616)$. There is no augmentation. NumPy PCG64 with seed $s + 40000$ generates the class permutation; its inverse maps original labels to output rows. Each convolution has stride one, zero padding one and a bias; each max pool has kernel and stride two. All linear layers have biases. The CPU PyTorch default generator, locally seeded with $s + 10000$, constructs the three convolutions, the 2048-to-256 layer, and the 256-to-10 head in that order. Their default initialization is uniform on $[-1/\sqrt{\text{fan_in}}, 1/\sqrt{\text{fan_in}}]$ for both weights and biases (implemented as Kaiming uniform with $a = \sqrt{5}$ for weights). The global CPU RNG is restored after construction; branch states are then loaded before transfer to CUDA. Current epoch order is a CPU `torch.randperm` with generator seed $(s+30000)10^5+100t+e$, where t and epoch e are zero-based. Saved initial tensors, per-epoch order digests and original-label mappings accompany every history.

CIFAR-100 source and exact replay allocation. The PyCIL source commit is `f3509b8ca3f20660ce4aa13f19d5283de81b4b35`. The archived upstream source is unmodified; the local GN8/head adaptation is specified by `experiments/r9_external/model.py`. Initialization first constructs the upstream network including its later-discarded ten-way head; upstream convolution weights are normal with standard deviation $\sqrt{2/(k_h k_w c_{\text{out}})}$, explicitly the upstream layer's `fan_out` rule. BN is replaced by GN8 with scale one and offset zero. After discarding the upstream head, a new 64-to-100 linear layer is constructed. Separately, its weights are overwritten by `kaiming_normal_` with $a = 0$, `mode=fan_in`, and `nonlinearity=leaky_relu`, giving variance $2/64$; its bias is zeroed. The fan rule is therefore intentionally layer-specific, not one shared convention. The discarded head and the new linear constructor both consume RNG draws; their order is preserved.

For u current examples already consumed in an epoch, next batch size n and memory size M, replay batch size is exactly

$$n_r = \left\lfloor \frac{(u+n)M}{5000} \right\rfloor - \left\lfloor \frac{uM}{5000} \right\rfloor .$$

Memory selection, current order and replay order use separate PCG64 generators. Their seeds are the first eight SHA256 bytes, read little-endian modulo 2^{63}, of compact ASCII JSON arrays respectively $[s + 50000, \texttt{``memory''}, c]$, $[s + 30000, \texttt{``current-order''}, t, e]$ and $[s + 60000, \texttt{``replay-order''}, t, e]$; $c$ is the original class ID. Each memory class contributes the first 20 or 200 entries of its permutation. Replay consumes a permutation of the concatenated memory exactly once per epoch. Per-image augmentation uses the same hash construction on $[s, \texttt{``augmentation''}, t, e, \text{source}, \text{position}, \text{sample_index}]$ and a private CPU PyTorch generator. Crop offsets, flip, ColorJitter's order draw and brightness factor are drawn in that order. The saved plans contain all indices and augmentation-key digests.

Environment and timing boundaries. Original run records identify Python 3.12.9, PyTorch 2.11.0+cu126, NumPy 1.26.4 and an NVIDIA GeForce RTX 4060 Laptop GPU. The CIFAR-100 records add Windows 11, torchvision 0.26.0+cpu, CUDA runtime 12.6, and two intra-op and two inter-op CPU threads; CIFAR-10 sets two intra-op threads. The full dependency lock is `requirements-r9-windows-lock.txt`. Training and forward evaluation use deterministic FP32, disabled cuDNN benchmarking and TF32, and `CUBLAS_WORKSPACE_CONFIG=:4096:8`. The 30.82-minute CIFAR-100 number measures the sequential eight-job runner from its entry through validation, training, evaluation and artifact writes; the 176.922-second retention number measures its complete 680-endpoint evaluation runner, including startup and validation. These are observed campaign wall times on this machine, excluding dataset download, installation, independent audits, manuscript production and packaging.

Frozen records and chronology. Table 14 identifies the configuration/source/analysis records by SHA256 prefix. Full hashes are in the package index and records. Times are machine-recorded UTC. The first follow-up run began on September 9 at 00:37:05 UTC, after its manifest; the head diagnostic began that day at 09:26:07 UTC and explicitly used completed follow-up evidence. Scalar calibration, envelopes and probe-retention are later post-hoc diagnostics. CIFAR-100 v2 began on

September 14 at 09:22:44 UTC, after its separate repair manifest and after v1 test exposure. The retention runner began at 23:14:38 UTC that day.

Record	Recorded UTC (2026)	SHA256 prefix (16 hex)
Follow-up manifest	Sep 09 00:36:51	`bb6b8e54d3ffb9ac`
Head diagnostic start	Sep 09 09:26:07	`ce9ceeb699285405`
Scalar manifest	Sep 11 08:09:19	`06fcc641ff7b3ac6`
Envelope manifest	Sep 13 22:54:17	`dd18092e66e13ab1`
CIFAR-100 v2 manifest	Sep 14 09:22:34	`863ec2639b19ac63`
Probe-retention manifest	Sep 14 23:10:41	`75325a7cd068258a`

Table 14: Frozen record identifiers. Manifest paths are `configs/`*campaign*`/manifest.json`, with campaign names `r9_followup`, `r9_scalar_calibration`, `r9_readout_envelope`, `r9_external_micro_v2`, and `r9_probe_retention`. The head start record is `results/r9_head_diagnostic/started.json`.

A short statistical reproduction path. The self-contained anonymous supplement includes the manuscript, code, and compact evidence used by the reported analyses. Its README and release manifest identify the enclosed manuscript and every evidence file. After extraction, use Python 3.12 from the extracted root (the output directory must not exist):

```
python -m pip install -r requirements-review.txt
python -I -X utf8 tools/recompute_reviewer_subset.py --output
reviewer-output
```

Only NumPy 1.26.4 is required for this command. It recomputes Tables 1 and 8 and the primary paired bootstrap from the forty follow-up integer-count records. It also recomputes all 160 candidate selections in Appendix O from 500 validation and 89 test endpoints, including unavailable histories, test violations, group summaries and every paired policy and conditional cross-history comparison. The attachment includes the manuscript's LaTeX dependencies; its README gives separate table-generation and TeX compilation commands. This lightweight path reproduces the specified statistics from archived counts and logits; the checks below additionally cover official labels, network inference and selected training runs.

Additional compact checks. The same extracted supplement provides three further CPU paths:

```
python -I -X utf8 tools/recompute_r15_compact.py --output
reviewer-output-r15
python -I -X utf8 tools/build_r15_evidence.py --analysis
reviewer-output-r15/analysis.json --output
reviewer-output-r15-tables
python -I -X utf8 analysis/recompute_r17_submission.py --raw
results/r17_review_controls_v1 --output
reviewer-output-r17.json
```

These commands reconstruct the full-budget CIFAR-100 decision statistics and the review controls from packaged per-example predictions or raw records. The official CBP and same-data ResNet-32 compact archives are under `evidence/r18/`; each includes its own evidence map, while the adjacent `audit/` reports record the completed integrity and replay checks.

Evidence boundary. The supplement includes frozen training source, protocols, manifests and the compact evidence required by the commands above. It excludes official image datasets, model checkpoints and the original full R15 logits. Thus the compact checks reproduce reported decisions and aggregates but do not rerun training or fresh network inference. Dataset sources and pinned checksums are recorded in the manifests; the README gives the separate R15 training command for a local CUDA environment.

Completed reproduction checks. Numerical implementations reconstruct counts and aggregates; separate functional implementations reconstruct selected network outputs. Isolated checks of the

base evidence included numerical audits, selected functional forwards, presentation regeneration and compilation of its archived manuscript. All 680 retention endpoints were separately re-executed: 7,140 arrays containing 57,120,000 elements agreed exactly. A limited training replay retrained the prespecified seed-1001/memory-20 history: all 5,925 updates, 88 model/optimizer checkpoints, 75 plans, 75 journals and 28 output arrays agreed with the archive. These completed checks reused the locked environment and verified the official data cache. The compact statistical commands were additionally executed in a new Python environment with only NumPy installed. Training replay covers the specified histories on the original hardware; the compact release preserves their audited outputs rather than the omitted checkpoints.

H External Validation-Selected Trajectory Diagnostic

This post-hoc diagnostic follows the original test results and the zero-retention probe finding. Its new-data selection rule was fixed before model inference. It reuses all twenty frozen class-incremental histories, seeds 201–210, at final task index four. Each current-only trajectory receives four epochs with the original batch size 64, order, learning rate, precision, and preprocessing: 157 steps per epoch, 628 total. Its terminal tensors match the original probe exactly in every history. The candidates are $0, 1, 2, 4, 8, 16, 32, 64, 128, 157, 314, 471, 628$ steps, defining the finite grid used for endpoint selection.

External data and selection access. The fixed CIFAR-10.1 v6 source commit is `d9982abb0bfc4846b8d13a11e66b887d946205d0`. For original class c, PCG64 seeded by $20260915 + c$ permutes its 200 official row indices. The first 100 are validation and the remainder test; both combined lists are sorted. All histories share this split. Original labels are mapped through each frozen class permutation. Each split has 800 old and 200 current examples; all ten prediction outputs participate. No external examples train the network. The validation labels include old classes and are an additional evaluation resource. Array decoding precedes selection, but no model receives test examples until all twenty validation selections have been sealed.

For $q = 20\%$ (primary descriptive requirement), 10% (sensitivity), and 0% (unconstrained reference), a candidate is feasible exactly when $100n_{\text{old,correct}} \geq 800q$ on validation. Among feasible candidates we maximize the integer current correct count; ties take the smallest step. An empty set is NA, with no fallback. The actual algorithm checkpoint is a fixed comparison anchor, not a candidate. Test forward evaluation is restricted to selected checkpoints and the pre, terminal, and actual anchors. The original image arrays are obtained separately from the authors; their data are not covered by the repository's MIT code license and are not redistributed in our package.

Complete results and their denominators. Tables 15–17 show all selections, including unavailable histories and test violations. Every old/current test value uses the same selected network. At $q = 20\%$, the jointly available seeds are 202, 203, 204, 205, 207, 209, and 210. ER and SGD current means on this common subset are 19.0714% and 1.3571%; their paired difference is 17.71 pp. ER's full ten-seed 18.15% mean summarizes a different set of available histories. Six of the seven available SGD endpoints have zero current accuracy.

At $q = 10\%$, all ten pairs are available and satisfy the test requirement; the mean current difference is 21.75 pp, with negative differences for seeds 203 and 205. The threshold itself equals ten-class uniform-random accuracy. At $q = 0\%$, both methods' selected endpoints have zero old accuracy; the paired current difference is 1.20 pp. Selection thus changes both the endpoint and the population available for comparison.

The actual ER models satisfy the 20% test requirement in all ten seeds (minimum 22%) and outperform their corresponding $q = 20\%$ and $q = 10\%$ selected endpoints in current accuracy. Their mean old/current accuracy is 27.40/55.85%. The $q = 20\%$ selection retains higher mean old accuracy (30.7375%), trading current accuracy for retention relative to actual ER.

Pooled retention and uncertainty. At $q = 20\%$, only two of ten ER selections achieve at least 20% on every old task. The seven available SGD selections contain 17 zero-accuracy old-task entries out of 28 (Table 18). Thus the pooled requirement permits uneven retention across old tasks. Starting old test accuracy also differs: ER 38.6625%, SGD 19.575%, while both have zero current accuracy. Attained old accuracy and selected update counts also vary across histories.

All differences and SDs are descriptive. Seeds jointly change class order and initialization and share the same external images; SD summarizes variation across these seeds on the fixed split. Validation maximization over thirteen candidates can be optimistic; held-out test violations are retained without reselection. NA denotes an empty feasible set on the specified candidate grid. Existing continual evaluation and adaptation assessment motivate these distinctions (De Lange et al., 2023; Inamdar et al., 2026).

Cost and evidence. The complete matrix used 12,560 new optimizer updates, 260 validation endpoints and 102 distinct test endpoints, and finished in 115.25 seconds under a 3,300-

second watchdog. Cost counts the entire 628-step trajectory, not the selected step. The separate manifest is `configs/r9_retention_trajectory/manifest.json`, SHA256 prefix `60ef2ac448e9e77c3`; it was frozen at 2026-09-15 23:32:31 UTC. Raw evidence includes all checkpoints, logits, labels, indices, exact counts, validation selection seals and artifact hashes. An independent NumPy implementation reconstructs mappings, counts, selection and aggregation; a functional CNN reproduces every test-scored checkpoint of both fixed seed-201 histories exactly in the locked environment.

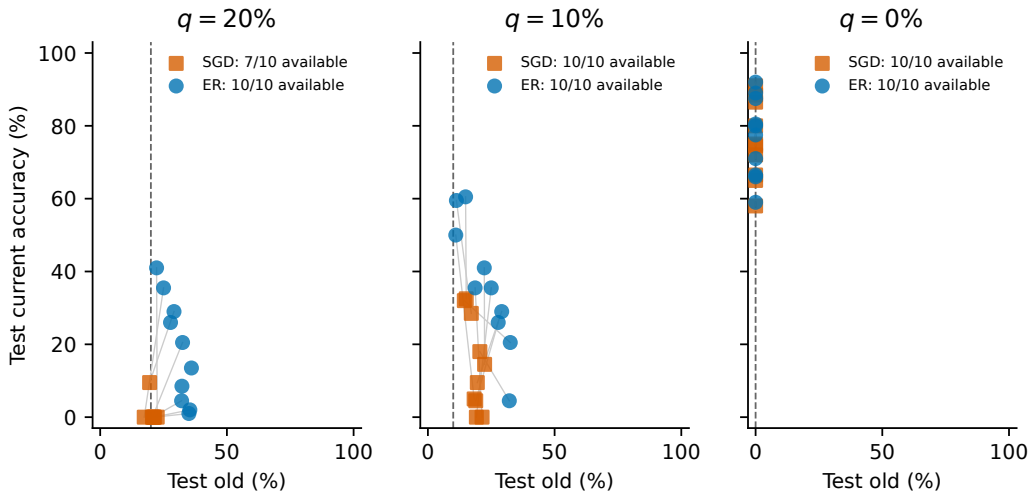


Figure 13: Selected test endpoints; grey segments connect paired seeds only when both validation selections are available. Dashed lines show the validation requirement. All points, including test violations, are retained. Symbols distinguish histories.

Seed	History	Step	Val. old	Val. current	Test old	Test current	Violation
201	SGD	NA	NA	NA	NA	NA	—
201	ER	2	30.75	6.50	32.25	8.50	no
202	SGD	0	20.62	0.00	21.25	0.00	no
202	ER	2	31.00	4.00	35.38	2.00	no
203	SGD	1	20.88	0.00	20.75	0.00	no
203	ER	2	31.75	3.00	32.12	4.50	no
204	SGD	1	20.50	0.00	22.50	0.00	no
204	ER	4	20.38	40.00	22.25	41.00	no
205	SGD	0	20.62	0.00	20.62	0.00	no
205	ER	2	32.00	18.50	32.50	20.50	no
206	SGD	NA	NA	NA	NA	NA	—
206	ER	4	28.38	30.50	27.75	26.00	no
207	SGD	0	20.88	0.00	21.38	0.00	no
207	ER	2	37.75	1.00	35.00	1.00	no
208	SGD	NA	NA	NA	NA	NA	—
208	ER	2	38.75	12.50	36.00	13.50	no
209	SGD	8	20.50	12.50	19.50	9.50	yes
209	ER	4	28.62	27.50	29.12	29.00	no
210	SGD	4	20.25	0.00	17.38	0.00	yes
210	ER	4	26.88	37.00	25.00	35.50	no

Table 15: All primary $q = 20\%$ selections. Accuracy is in percent; NA means no validation-feasible candidate. The test set is not used to remove violations.

Seed	History	Step	Val. old	Val. current	Test old	Test current	Violation
201	SGD	8	14.38	29.00	15.12	32.50	no
201	ER	4	13.62	53.50	14.88	60.50	no
202	SGD	8	17.25	38.00	17.12	28.50	no
202	ER	4	12.00	59.50	11.25	59.50	no
203	SGD	8	18.88	11.50	20.50	18.00	no
203	ER	2	31.75	3.00	32.12	4.50	no
204	SGD	8	19.88	8.50	22.38	14.50	no
204	ER	4	20.38	40.00	22.25	41.00	no
205	SGD	8	14.50	24.50	14.38	32.00	no
205	ER	2	32.00	18.50	32.50	20.50	no
206	SGD	8	18.88	8.50	18.88	4.50	no
206	ER	4	28.38	30.50	27.75	26.00	no
207	SGD	0	20.88	0.00	21.38	0.00	no
207	ER	4	19.62	36.50	18.62	35.50	no
208	SGD	8	19.25	2.50	18.25	5.00	no
208	ER	4	11.62	46.50	11.00	50.00	no
209	SGD	8	20.50	12.50	19.50	9.50	no
209	ER	4	28.62	27.50	29.12	29.00	no
210	SGD	0	19.88	0.00	19.12	0.00	no
210	ER	4	26.88	37.00	25.00	35.50	no

Table 16: All $q = 10\%$ sensitivity selections, with the same fixed split.

Seed	History	Step	Val. old	Val. current	Test old	Test current	Violation
201	SGD	628	0.00	74.00	0.00	74.50	no
201	ER	471	0.00	74.50	0.00	77.50	no
202	SGD	128	0.00	93.50	0.00	88.50	no
202	ER	64	0.00	94.50	0.00	89.00	no
203	SGD	314	0.00	77.00	0.00	66.50	no
203	ER	314	0.00	74.50	0.00	66.00	no
204	SGD	314	0.00	89.00	0.00	80.00	no
204	ER	128	0.00	90.50	0.00	80.00	no
205	SGD	471	0.00	90.50	0.00	91.00	no
205	ER	471	0.00	89.50	0.00	92.00	no
206	SGD	471	0.00	76.00	0.00	65.00	no
206	ER	471	0.00	77.50	0.00	66.50	no
207	SGD	314	0.00	80.50	0.00	75.00	no
207	ER	471	0.00	80.50	0.00	80.50	no
208	SGD	628	0.00	86.00	0.00	86.50	no
208	ER	314	0.00	91.50	0.00	87.50	no
209	SGD	628	0.00	59.50	0.00	58.00	no
209	ER	64	0.00	63.00	0.00	59.00	no
210	SGD	471	0.00	74.00	0.00	72.00	no
210	ER	64	0.00	72.50	0.00	71.00	no

Table 17: All unconstrained $q = 0\%$ validation selections. High current accuracy coexists with zero old accuracy in every selected test endpoint.

Seed	History	Old task 0	Old task 1	Old task 2	Old task 3
201	SGD	NA	NA	NA	NA
201	ER	24.00	20.50	30.50	54.00
202	SGD	0.00	0.00	0.00	85.00
202	ER	16.50	31.00	66.50	27.50
203	SGD	0.00	0.00	9.50	73.50
203	ER	24.00	49.00	15.00	40.50
204	SGD	0.50	0.00	0.00	89.50
204	ER	7.00	17.50	11.00	53.50
205	SGD	0.00	0.00	0.00	82.50
205	ER	17.50	40.00	48.50	24.00
206	SGD	NA	NA	NA	NA
206	ER	18.50	34.00	15.00	43.50
207	SGD	0.00	0.00	0.00	85.50
207	ER	28.00	43.00	37.50	31.50
208	SGD	NA	NA	NA	NA
208	ER	15.50	53.50	37.00	38.00
209	SGD	0.00	0.00	6.50	71.50
209	ER	30.00	31.50	15.00	40.00
210	SGD	0.00	0.00	1.00	68.50
210	ER	29.00	10.00	42.50	18.50

Table 18: Each old task's test accuracy at the primary $q = 20\%$ selected endpoint. Each task contains 200 test images. NA selections remain absent.

I TEN-SEED DER++ CHALLENGE

This subsequent descriptive comparison reuses the exact class-incremental SGD/ER seeds, original class permutation, CNN initialization, current minibatch order, data preprocessing and four-epoch history budget. It adds ten full DER++ histories, seeds 201–210. It was specified after the original results and uses previously evaluated CIFAR-10 test data. DER++ provides an additional established replay baseline from 2020. All ten histories are included in the descriptive comparison.

Algorithm and counts. At each step the loss is current ten-output CE plus 0.5 times mean logit MSE on one reservoir draw plus 0.5 times CE on a second draw. The draws have up to 32 examples each, each without replacement internally; they can overlap. Memory has 500 total slots, updated after the optimizer step with the current examples and their pre-update arrival logits. Repeated epoch presentations count as arrivals. Independent storage and sampling RNG streams preserve reservoir membership despite the extra draw. The coefficients were fixed without search on these outcomes. They differ from the discovery pilot's $\alpha = 0.1$. The existing frozen engine is reused, with ER-specific release controls disabled; actual, matched current-only probe and fresh branches remain. Each branch has 628 updates per task. Across five tasks and ten histories, all three branches use 94,200 updates and 8,008,960 training image forwards, including both replay draws.

Epoch	History	Fresh F	Probe p	Actual a	$C = p - a$
1	SGD	64.242 ± 4.414	72.410 ± 4.466	72.410 ± 4.466	0.000 ± 0.000
1	ER	64.242 ± 4.414	74.229 ± 3.818	52.444 ± 8.773	21.785 ± 8.975
1	DER++	64.242 ± 4.414	75.602 ± 4.254	66.793 ± 4.661	8.810 ± 4.271
2	SGD	69.304 ± 4.373	75.029 ± 5.272	75.029 ± 5.272	0.000 ± 0.000
2	ER	69.304 ± 4.373	77.249 ± 4.796	58.936 ± 7.387	18.312 ± 6.314
2	DER++	69.304 ± 4.373	78.843 ± 4.473	72.375 ± 4.649	6.468 ± 3.021
4	SGD	77.144 ± 4.760	80.546 ± 5.224	80.546 ± 5.224	0.000 ± 0.000
4	ER	77.144 ± 4.760	81.517 ± 4.955	69.395 ± 7.535	12.123 ± 6.813
4	DER++	77.144 ± 4.760	82.624 ± 4.671	78.365 ± 4.362	4.259 ± 3.182

Table 19: All specified prefixes, mean $\pm$ sample SD over ten seeds. Each seed equally averages its four noninitial tasks. Accuracy is percent; C is in percentage points. SGD/ER are the original frozen comparisons, not new retraining; DER++ is the subsequent challenge.

Mean disagreement and individual exceptions. At epoch four DER++ actual current accuracy is 2.18125 pp below SGD (paired SD 3.80482), while its probe accuracy is 2.07750 pp above (SD 1.07273). All ten probe differences are positive, but actual accuracy is higher for seeds 201, 204 and 209. Against ER, the actual mean improvement is 8.97000 pp (SD 5.73082), and the probe mean improvement 1.10625 pp (SD 0.70544); seed 210's probe difference is negative. The ordering therefore depends on the method pair and seed. The matched fresh reference is identical across methods and cancels in method contrasts.

Seed	DER++ a	DER++ p	Δa_{S}	Δp_{S}	Δa_{E}	Δp_{E}
201	84.9250	85.3750	1.5000	1.9500	7.1500	0.9750
202	84.9625	91.8250	-6.6875	0.1750	5.7500	0.2500
203	79.1500	83.0125	-1.2000	2.6625	11.9750	1.8500
204	81.5875	82.4500	1.8000	2.6625	10.0500	1.4000
205	73.7500	82.3500	-6.8875	1.7125	20.9125	0.8000
206	73.9125	77.2500	-1.1125	2.2250	6.5500	1.6625
207	75.8750	85.5250	-8.7000	0.9500	0.8375	0.6375
208	79.7500	84.4375	-0.6000	4.0875	14.6375	1.6125
209	74.9125	76.7125	0.9000	2.7000	7.7000	1.9750
210	74.8250	77.3000	-0.8250	1.6500	4.1375	-0.1000

Table 20: All epoch-four seed values. Differences subtract SGD (subscript S) or ER (E) from DER++; positive Δa or Δp favors DER++ in accuracy, the opposite sign convention from fresh-relative gaps.

Actual retention. DER++ improves current-task accuracy over ER while attaining lower final old and all-ten accuracy (Table 21). Old/seen/all values come only from actual epoch-four checkpoints, so each retention/current pair describes the same final adapted model.

History	Final old	Final current	Final all ten
SGD	0.000 ± 0.000	81.655 ± 11.541	16.331 ± 2.308
ER	37.551 ± 4.351	64.615 ± 12.510	42.964 ± 2.463
DER++	29.608 ± 5.847	79.540 ± 8.668	39.594 ± 4.474

Table 21: Actual final-task checkpoint accuracy, mean $\pm$ sample SD. Old/current use 8,000/2,000 official test images; all ten weights them 0.8/0.2. This final-task endpoint differs from the four-task means above.

Traceability and execution. The new manifest `configs/r9_derpp_challenge/manifest.json` has SHA256 prefix `b001c94f5594983b`. The full ten-history matrix completed in 589.188 seconds under one 3,300-second watchdog. Every history retains initial, pre, actual, probe, fresh and reservoir states, exact task/epoch counts, source snapshots, orders and artifact hashes. An independent implementation verifies 3,750 evaluation records and 630 model states across new histories and old comparisons. A separate functional CNN evaluates fixed seed 201, final task, all three branch checkpoints on all five test tasks: 30,000 images. Actual-task counts and prediction hashes, and the probe/fresh current counts and hashes, agree with the archive. Old probe/fresh outputs in that check are separately recorded additional diagnostics. Full integer counts for every seed, prefix and task are provided in `paper_r9/generated/derpp_current_counts.csv`. Old-task scores are reported at the measured final checkpoints. In a fresh minimal source directory, the prespecified seed-201 history was also retrained through all five tasks and three branches (9,420 updates). All 31 checkpoint files' model and reservoir tensors and all scientific task records matched, excluding execution times. This check reused the locked runtime and hash-verified public cache on the original hardware; its training-replay scope is the single specified seed.

J Continual Backpropagation Control

We port Continual Backpropagation (CBP) from the authors' repository at fixed commit `a6b79580d85f3025bdb601566d3627c5f489f13b`. The port preserves the official contribution utility, age correction, maturity gate, replacement accumulator, incoming-weight initialization and outgoing-weight reset. It runs after every SGD step in the primary compact CNN. Fresh and probe branches use the same current-only SGD operation as the primary comparison. Data, class orders, initializations, minibatch orders and four-epoch budgets match the five paired SGD and ER seeds 201–205.

Profile	a (%)	p (%)	C (pp)	Final all (%)	Replaced
Nature CIFAR	83.17 ± 4.95	83.18 ± 4.94	0.010 ± 0.016	17.22 ± 2.09	8.0 ± 0.0
Conv default	83.24 ± 4.86	83.23 ± 4.89	-0.005 ± 0.062	17.23 ± 2.12	143.0 ± 0.0

Table 22: CBP control, mean $\pm$ sample SD over five seeds after averaging four noninitial tasks within seed. $C = p - a$. "Replaced" is the number of hidden units reset over the complete five-task stream. The first row uses the authors' incremental-CIFAR parameters; the second tests the more aggressive official ConvCBP defaults.

The incremental-CIFAR parameters use replacement rate 10^{-5}, maturity 1000 and utility decay 0.99; each run replaces eight units. CBP and SGD have nearly identical actual accuracy (-0.0025 ± 0.043 pp), while CBP's a and p agree within 0.010 ± 0.016 pp. The official ConvCBP-default sensitivity uses rate 10^{-4}, maturity 100 and decay 0.9, replacing 143 units; its C remains -0.005 ± 0.062 pp. Thus the agreement is stable across two official replacement time scales.

The source snapshot, function-level port map, ten configs, seed-level JSON, per-example logits and artifact hashes accompany the submission. Tests verify feature-to-layer alignment, the convolution-to-linear index map, zero-rate identity and outgoing-weight resets. The comparison is descriptive because it was added after review; all five frozen seeds and both official parameter profiles are reported.

K SAME-DATA ARCHITECTURE BRIDGE

This bridge keeps the primary CIFAR-10 class-incremental stream fixed and changes only the network. We use a pinned PyCIL CIFAR ResNet-32 with eight-group normalization, the same five seeds (201–205), class orders, initialization and minibatch orders, four epochs per task, batch size 64, learning rate 0.01, zero momentum and weight decay, ten-output head, ER reservoir of 500 and replay batch 32. The architecture source is pinned at commit `f3509b8ca3f20660ce4aa13f19d5283de81b4b35`; the bridge audit verifies every task image/label hash against the compact-CNN runs.

Architecture	Fresh F	SGD $a{=}p$	ER a	ER p	ΔA	ΔH	Final SGD	Final ER
Compact CNN	79.38 ± 3.62	83.17 ± 4.94	69.71 ± 10.60	83.95 ± 4.48	$+13.46 \pm 8.59$	-0.78 ± 0.51	17.23 ± 2.09	42.37 ± 2.63
ResNet-32–GN8	80.39 ± 4.05	81.93 ± 5.79	58.87 ± 15.78	83.12 ± 6.23	$+23.07 \pm 17.38$	-1.19 ± 3.25	16.29 ± 3.30	30.63 ± 5.21
ResNet − CNN	–	–	–	–	$+9.61 \pm 10.69$	-0.41 ± 3.28	-0.94 ± 1.71	-11.74 ± 5.64

Table 23: Same-data architecture bridge. Entries are seed-level means over noninitial tasks 1–4, followed by sample SD. ΔA and ΔH use the paper's actual and history-gap definitions. The final columns score the last actual endpoint over all ten CIFAR-10 classes.

The ResNet bridge gives $\Delta A = +23.07 \pm 17.38$ pp, positive in all five seeds, and $\Delta H = -1.19 \pm 3.25$ pp, negative in three. The descriptive paired bootstrap intervals are $[8.48, 38.94]$ and $[-4.25, 1.12]$ pp. The actual separation therefore transfers across architecture; the probe direction is seed-sensitive. ER retains $37.41 \pm 8.18\%$ old-task accuracy at the noninitial endpoints and ends at $30.63 \pm 5.21\%$ all-task accuracy, versus $16.29 \pm 3.30\%$ for SGD.

The compact evidence archive `evidence/r18/r18_architecture_bridge_compact_v1.zip` contains only integer predictions, labels, official indices, configs and the recomputation script. It is independently rechecked against 100 current-task and 50 full-test checkpoint forwards; all logits match bitwise. The full local raw directory retains checkpoints and float32 logits for the audit trail.

L Reviewer-Requested Operator Controls

These post-hoc descriptive controls were specified in response to review and branch from the frozen CIFAR-10 checkpoints. They do not alter any continuing history or the prespecified primary analysis. Seeds remain the statistical units; tables report mean ± sample SD over all ten seeds.

Unit-sum compensated ER. The original compensated branch at each noninitial task minimizes $\ell_c + (1 - w_c)\ell_r$, whose coefficient sum is $2 - w_c$. The new branch uses

$$\ell_{\text{norm}} = \frac{\ell_c + (1 - w_c)\ell_r}{2 - w_c}.$$

It preserves the compensated branch's current-to-replay coefficient ratio, restores a unit coefficient sum at every update, and uses the same saved pre-task parameters, reservoir state, four epochs, current minibatch order, and replay random state. All 40 task branches retain per-example endpoint logits and checkpoints. The paired quantity is the four-task seed mean of $a_{\text{SGD}} - a_{\text{control}}$, matching ΔA's sign convention.

ER operator	ΔA vs. SGD (pp)	Seed n
Unnormalized compensation	5.10 ± 3.43	10
Unit-sum normalized	1.88 ± 2.03	10

Table 24: Gradient-scale control for compensated ER. Both variants use the same relative current/replay weighting. Normalizing the coefficient sum reduces the ER-versus-SGD gap but does not reverse its mean direction.

Frozen-reservoir continuation. For each ER seed, the final-task actual model, all 500 epoch-four reservoir rows, arrival count, and sampling RNG state are restored exactly. Training then continues through epoch 16 with the original learning rate and current minibatch orders. Replay draws continue from the restored sampling stream, but insertion is disabled, so reservoir contents and composition remain fixed. The evolving comparison is the previously reported continuation from the identical model and memory anchor, with normal reservoir insertion.

Epoch	Reservoir	Current	Old	All seen	Current-two
8	Fixed at epoch 4	68.00 ± 14.64	33.61 ± 6.39	40.49 ± 4.39	82.56 ± 11.30
8	Evolving	72.22 ± 13.21	31.02 ± 6.10	39.26 ± 3.80	82.91 ± 10.82
16	Fixed at epoch 4	81.62 ± 10.64	26.96 ± 6.02	37.89 ± 3.25	86.83 ± 9.45
16	Evolving	82.78 ± 11.36	22.74 ± 6.69	34.74 ± 3.34	85.94 ± 9.86

Table 25: Final-task ER accuracy (%) under fixed versus evolving replay contents. All-10 current, pooled old and all-seen scores come from the same models; current-two supplies task identity.

At epoch four, before either continuation, ER has 64.615% current, 37.551% old and 42.964% all-seen accuracy. With a fixed buffer, the epoch-16 current score rises by 17.010 pp while old accuracy falls by 10.593 pp. Additional optimization alone therefore produces a large current gain and a retention cost. Evolving the reservoir adds 1.155 pp all-class current accuracy relative to the fixed control but loses 4.223 pp old and 3.147 pp all-seen accuracy. The original continuation's qualitative tradeoff is not an artifact of changing buffer composition, although composition materially changes its location.

The raw directory `results/r17_review_controls_v1` contains every seed record, endpoint checkpoint and compressed per-example logits. The independent formatter `analysis/build_official_review_controls_evidence.py` verifies their hashes and regenerates Tables 24 and 25; its audit report is `audit/official_llm_revision/CONTROLS_EVIDENCE.json`.

M Final-Task Budget Continuation and Complete Values

This post-hoc diagnostic continues the saved task-four epoch-four endpoints of seeds 201–210. SGD/ER/DER++ each supply actual and current-only probe branches; one fresh branch per seed is shared after verifying identical initial endpoint tensors across methods. All 70 branches run twelve additional epochs on the complete 10,000-image current training task, with 157 updates per epoch. Earlier tasks are not retrained. The continuation uses the original SGD learning rate 0.01, batch 64, zero momentum and zero weight decay; ER keeps its concatenated mean loss and DER++ its two replay draws, pre-update teacher targets and $\alpha = \beta = 0.5$.

Actual replay restores all 500 memory rows, the arrival count and both storage/sample random states. Current order follows the original epoch formula with zero-based epoch indices 4–15. Probes and fresh use current examples only. All branch orders are paired within seed. Checkpoints are saved at every epoch from four through sixteen; complete five-task logits, labels and official indices are saved at the three fixed endpoints. The full group completed 131,880 additional updates in 14.128 minutes under one 55-minute watchdog, with all ten seeds retained.

Independent scoring reconstructs every endpoint from integer official labels and saved logits. A separately implemented functional network exactly reproduces forward outputs for all seven seed-201 branches at all three endpoints. SGD actual/probe tensors match across all ten seeds and thirteen checkpoints, an implementation invariant. The artifact includes all 210 endpoint records, every task's integer correct/total counts, and all 180 paired rows in `budget_endpoint_counts.csv` and `budget_paired.csv`. Tables below show means/sample SDs and every seed's current score as descriptive summaries of the fixed endpoints.

Epoch	History	Branch	Current	Current 2	Old	All ten
4	SGD	actual	81.66 ± 11.54	81.66 ± 11.54	0.00 ± 0.00	16.33 ± 2.31
4	SGD	probe	81.66 ± 11.54	81.66 ± 11.54	0.00 ± 0.00	16.33 ± 2.31
4	ER	actual	64.62 ± 12.51	84.99 ± 9.11	37.55 ± 4.35	42.96 ± 2.46
4	ER	probe	83.34 ± 10.94	83.34 ± 10.94	0.00 ± 0.00	16.67 ± 2.19
4	DER++	actual	79.54 ± 8.67	83.80 ± 9.88	29.61 ± 5.85	39.59 ± 4.47
4	DER++	probe	84.33 ± 10.25	84.33 ± 10.25	0.01 ± 0.03	16.87 ± 2.04
4	Fresh	fresh	78.01 ± 11.96	78.01 ± 11.96	0.00 ± 0.00	15.60 ± 2.39
8	SGD	actual	81.99 ± 11.93	81.99 ± 11.93	0.00 ± 0.00	16.40 ± 2.39
8	SGD	probe	81.99 ± 11.93	81.99 ± 11.93	0.00 ± 0.00	16.40 ± 2.39
8	ER	actual	72.22 ± 13.21	82.91 ± 10.82	31.02 ± 6.10	39.26 ± 3.80
8	ER	probe	82.05 ± 11.66	82.05 ± 11.66	0.00 ± 0.00	16.41 ± 2.33
8	DER++	actual	83.87 ± 9.20	86.46 ± 9.82	26.27 ± 5.27	37.79 ± 3.83
8	DER++	probe	83.71 ± 11.29	83.71 ± 11.29	0.00 ± 0.00	16.74 ± 2.26
8	Fresh	fresh	80.30 ± 13.73	80.30 ± 13.73	0.00 ± 0.00	16.06 ± 2.75
16	SGD	actual	86.81 ± 9.50	86.81 ± 9.50	0.00 ± 0.00	17.36 ± 1.90
16	SGD	probe	86.81 ± 9.50	86.81 ± 9.50	0.00 ± 0.00	17.36 ± 1.90
16	ER	actual	82.78 ± 11.36	85.94 ± 9.86	22.74 ± 6.69	34.74 ± 3.34
16	ER	probe	87.26 ± 9.11	87.26 ± 9.11	0.00 ± 0.00	17.45 ± 1.82
16	DER++	actual	87.04 ± 8.81	87.73 ± 8.50	17.78 ± 6.08	31.63 ± 3.32
16	DER++	probe	88.30 ± 8.12	88.30 ± 8.12	0.00 ± 0.00	17.66 ± 1.62
16	Fresh	fresh	86.43 ± 9.37	86.43 ± 9.37	0.00 ± 0.00	17.29 ± 1.87

Table 26: Complete continuation endpoints (%), ten-seed mean ± sample SD. Current, Old and All ten use the same all-ten model predictions; Current 2 only changes its current-task candidate set. Old pools four equally sized tasks; All ten weights old/current 0.8/0.2. Fresh is shared across methods. Rounding a tiny old accuracy to zero does not assert zero integer counts.

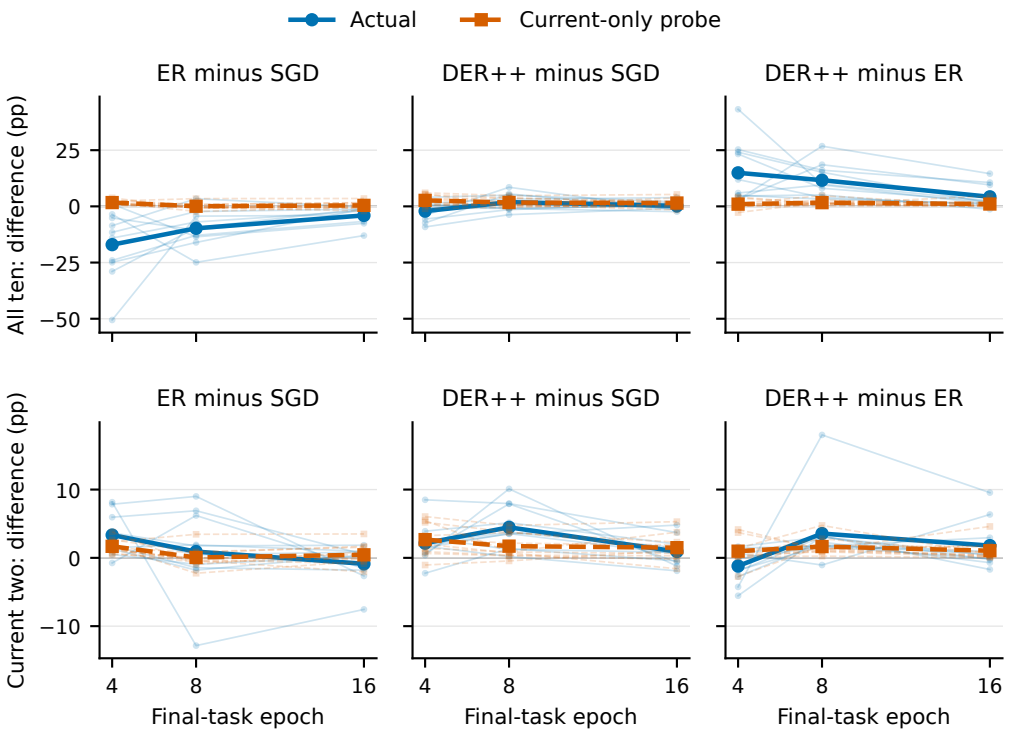


Figure 14: Complete paired current-accuracy differences. Thin curves are all ten seeds; thick curves are means, without inferential bars. Actual uses each algorithm's operation and probe is current-only. Lower panels restrict the same logits to current classes, supplying task identity. Axes are shared within each row; the rows have different vertical ranges. Accuracy differences are the negatives of corresponding differences in fresh-relative gaps A/H. Lines connect the three fixed final-task endpoints.

Epoch	Difference	Candidates	Actual difference	Probe difference
4	ER-SGD	All ten	-17.040 ± 15.400	+1.690 ± 0.995
4	DER++-SGD	All ten	-2.115 ± 4.221	+2.670 ± 2.334
4	DER++-ER	All ten	+14.925 ± 13.514	+0.980 ± 1.933
4	ER-SGD	Current two	+3.340 ± 3.075	+1.690 ± 0.995
4	DER++-SGD	Current two	+2.145 ± 2.774	+2.670 ± 2.334
4	DER++-ER	Current two	-1.195 ± 2.471	+0.980 ± 1.933
8	ER-SGD	All ten	-9.765 ± 7.886	+0.060 ± 1.484
8	DER++-SGD	All ten	+1.880 ± 3.753	+1.715 ± 1.778
8	DER++-ER	All ten	+11.645 ± 7.474	+1.655 ± 1.271
8	ER-SGD	Current two	+0.915 ± 6.118	+0.060 ± 1.484
8	DER++-SGD	Current two	+4.470 ± 3.277	+1.715 ± 1.778
8	DER++-ER	Current two	+3.555 ± 5.220	+1.655 ± 1.271
16	ER-SGD	All ten	-4.030 ± 3.999	+0.445 ± 1.559
16	DER++-SGD	All ten	+0.230 ± 1.923	+1.490 ± 1.976
16	DER++-ER	All ten	+4.260 ± 5.290	+1.045 ± 1.384
16	ER-SGD	Current two	-0.870 ± 2.742	+0.445 ± 1.559
16	DER++-SGD	Current two	+0.925 ± 2.135	+1.490 ± 1.976
16	DER++-ER	Current two	+1.795 ± 3.540	+1.045 ± 1.384

Table 27: All fixed paired method differences (pp), mean ± sample SD. Each difference uses the same ten seeds at the same final-task endpoint. The shared fresh reference cancels from method contrasts; differences in A/H are the negatives of the corresponding actual/probe differences.

Replay composition across budgets. A further post-hoc count of every actual ER/DER++ endpoint buffer finds mean old-example fractions 80.30%, 67.12% and 50.70%. Each presentation is a reservoir arrival: the cumulative count rises from 200,000 to 320,000 while earlier tasks contribute the same 160,000 presentations. Buffer task counts match between ER and DER++ for each seed because their storage random streams and arrivals match. This count was added after viewing the complete budget scores. The continuation jointly changes optimization time, reservoir contents, replay exposure and model states; the observed retention and ranking changes characterize this complete protocol.

Epoch	Seed	SGD $a = p$	ER a	ER p	DER++ a	DER++ p	Fresh
4	201	77.10	73.55	80.90	76.45	82.30	70.75
4	202	95.90	81.80	96.40	86.75	96.55	91.20
4	203	72.75	48.70	74.05	74.00	78.20	69.65
4	204	92.45	67.55	93.50	91.65	93.85	90.30
4	205	92.55	42.00	93.25	85.20	93.45	89.60
4	206	73.70	62.15	76.35	74.10	76.95	70.35
4	207	85.10	79.95	86.45	85.90	87.10	82.65
4	208	91.35	62.40	92.85	85.70	94.20	89.85
4	209	60.45	61.45	62.80	64.45	66.50	59.15
4	210	75.20	66.60	76.90	71.20	74.15	66.65
8	201	80.30	55.40	79.35	82.20	84.10	78.00
8	202	96.25	89.30	96.30	92.55	96.55	96.45
8	203	75.80	62.90	75.50	79.00	76.70	71.10
8	204	95.30	79.25	95.25	94.45	96.15	94.60
8	205	86.55	82.20	85.95	92.10	86.95	89.75
8	206	76.35	74.05	77.25	78.00	80.05	76.70
8	207	88.95	78.40	88.55	87.95	89.70	85.60
8	208	92.10	82.60	92.85	90.65	94.25	92.05
8	209	61.10	47.55	64.55	66.10	65.85	59.30
8	210	67.20	70.60	64.95	75.70	66.75	59.50
16	201	89.75	76.75	90.65	91.30	91.15	88.95
16	202	97.10	95.10	96.85	96.30	96.90	96.95
16	203	76.00	69.30	75.15	79.90	79.75	76.35
16	204	95.80	94.70	96.30	96.15	96.45	95.45
16	205	95.50	91.55	95.45	94.05	95.80	94.55
16	206	79.60	78.40	81.15	79.15	81.90	78.35
16	207	86.35	82.90	84.30	84.00	84.75	86.50
16	208	94.60	93.50	96.30	95.70	96.50	93.55
16	209	69.40	61.85	72.90	71.40	74.70	68.95
16	210	84.00	83.75	83.50	82.45	85.10	84.70

Table 28: All-ten current-task accuracy (%) for every seed and fixed continuation endpoint. Values come from integer counts out of 2,000 current examples. SGD actual/probe coincide. These are task-four scores, not the noninitial-task averages reported for the original follow-up.

Epoch	Seed	SGD $a = p$	ER a	ER p	DER++ a	DER++ p	Fresh
4	201	77.10	85.25	80.90	81.00	82.30	70.75
4	202	95.90	96.40	96.40	93.65	96.55	91.20
4	203	72.75	76.40	74.05	74.60	78.20	69.65
4	204	92.45	93.75	93.50	95.40	93.85	90.30
4	205	92.55	91.80	93.25	93.25	93.45	89.60
4	206	73.70	77.15	76.35	75.20	76.95	70.35
4	207	85.10	86.90	86.45	86.80	87.10	82.65
4	208	91.35	92.85	92.85	93.55	94.20	89.85
4	209	60.45	68.30	62.80	68.95	66.50	59.15
4	210	75.20	81.15	76.90	75.60	74.15	66.65
8	201	80.30	67.45	79.35	85.45	84.10	78.00
8	202	96.25	95.30	96.30	97.60	96.55	96.45
8	203	75.80	77.55	75.50	79.30	76.70	71.10
8	204	95.30	93.50	95.25	96.55	96.15	94.60
8	205	86.55	92.75	85.95	94.45	86.95	89.75
8	206	76.35	76.80	77.25	79.90	80.05	76.70
8	207	88.95	87.50	88.55	89.15	89.70	85.60
8	208	92.10	94.00	92.85	95.85	94.25	92.05
8	209	61.10	70.10	64.55	69.05	65.85	59.30
8	210	67.20	74.10	64.95	77.30	66.75	59.50
16	201	89.75	82.20	90.65	91.75	91.15	88.95
16	202	97.10	97.35	96.85	96.65	96.90	96.95
16	203	76.00	77.85	75.15	80.80	79.75	76.35
16	204	95.80	96.25	96.30	96.60	96.45	95.45
16	205	95.50	93.95	95.45	94.90	95.80	94.55
16	206	79.60	80.45	81.15	80.20	81.90	78.35
16	207	86.35	84.55	84.30	84.45	84.75	86.50
16	208	94.60	95.55	96.30	96.10	96.50	93.55
16	209	69.40	66.75	72.90	73.10	74.70	68.95
16	210	84.00	84.50	83.50	82.80	85.10	84.70

Table 29: Current-two accuracy (%) for the same endpoints as Table 28. The candidate mask supplies task identity, yielding a task-aware score. Every seed is retained.

N Rechecking the Four-Task Aggregate at Longer Budgets

To extend the final-task continuation in Appendix M to the primary four-task aggregate, we perform a post-hoc descriptive check on the same ten seeds. For each of the four noninitial tasks, each method's actual and current-only probe branches continue from their own saved epoch-four endpoints to epochs eight and sixteen. Earlier histories remain the original four-epoch histories: extended branches are never propagated into subsequent tasks. The forty existing final-task SGD/ER branches are reused with verified provenance; the remaining 120 branches are newly continued. Every seed and fixed endpoint is included.

All learning rates, current minibatch orders, update operators and restored replay random states follow the original implementation. SGD has zero momentum and weight decay, so continuation requires no additional optimizer state. The two identical SGD branches are both retained as an implementation check. Each seed remains one statistical unit. The shared fresh reference cancels in the between-method accuracy contrast, so continuation requires only actual and probe branches.

Epoch	Actual, all ten	Probe, all ten	Actual, current two	Probe, current two
4	-11.151 ± 6.752	$+0.971 \pm 0.726$	$+2.829 \pm 2.125$	$+0.971 \pm 0.726$
8	-8.445 ± 5.351	$+0.233 \pm 0.541$	$+0.825 \pm 3.311$	$+0.233 \pm 0.541$
16	-2.618 ± 2.610	$+0.595 \pm 0.891$	$+1.116 \pm 1.275$	$+0.595 \pm 0.891$

Table 30: ER-minus-SGD current accuracy, in percentage points: mean $\pm$ sample SD over ten seeds, after averaging four noninitial tasks within each seed. These are accuracy differences, opposite in sign to the gap differences in the primary test. All three fixed budgets are reported. The current-two mask uses test-task identity. Summaries are descriptive.

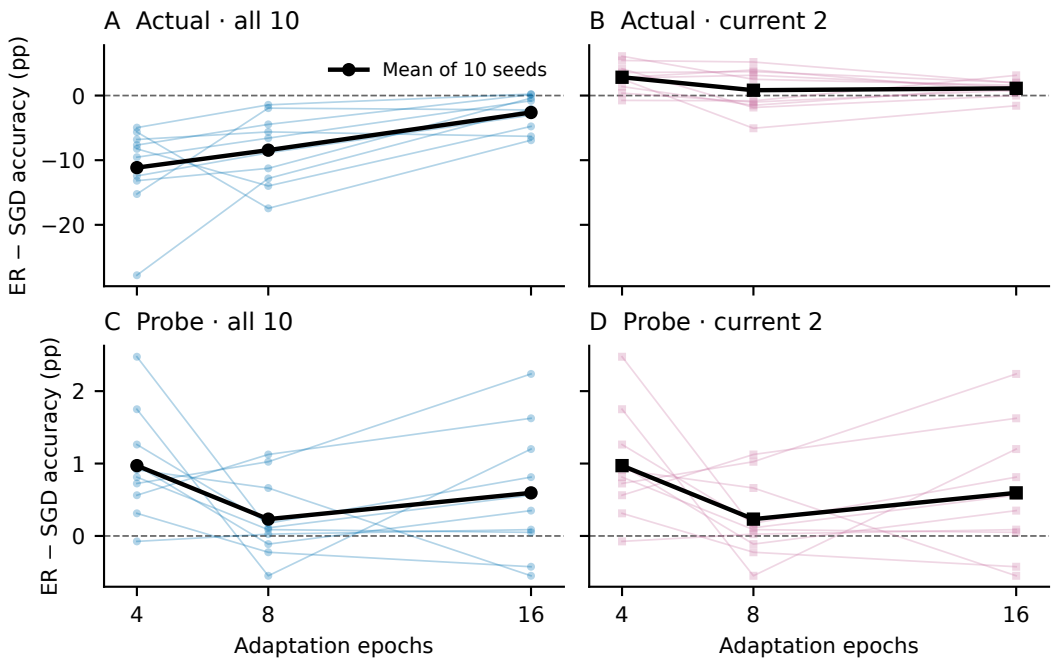


Figure 15: Four-task accuracy contrasts across adaptation budgets. Thin curves show all ten paired seed means; black curves show their mean. Columns share a vertical scale within each row; actual and probe rows use different scales. Zero is included in every panel. The task histories are fixed, while the actual ER reservoir continues to receive arrivals. These curves characterize the complete continuation protocol, including the evolving replay composition.

Each fixed endpoint stores all five tasks' logits, labels and official sample indices. Independent NumPy code recomputes current, pooled-old, seen-task and all-five-task accuracy, task-specific

scores, and current-two scores using the appropriate output rows for each task. Epoch-four actual scores are matched to the original all-five-task records; noncurrent probe scores are explicitly identified as new reevaluations (or reused reevaluations for the final task). The artifact retains every per-seed and per-task endpoint, including paired current/old/seen changes relative to epoch four. All scores associated with a checkpoint refer to that same model. Accuracy differences and within-seed averages are computed as exact rational numbers from integer correct counts; signs are assigned before conversion to floating point. This corrects an analysis-version-one rounding artifact: seed 204 has exactly zero actual current-two contrast at epoch sixteen (taskwise correct-count differences $4, 8, -21, 9$). The corresponding seed counts are eight positive, one negative and one zero. Frozen raw outputs and the original analysis snapshot are retained.

The numerical audit checks the complete 480-endpoint matrix, original model and memory anchors, batch order, update counts, and reservoir random state transitions. An additional independent functional CNN forward pass checks all 36 newly continued seed-201 endpoints across tasks one–three, both methods and branches, and all three budgets on the original hardware. Table 31 provides the complete paired values for the all-ten comparison. The protocol is `protocol/R12_FOUR_TASK_BUDGET_PROTOCOL.md`; the manifest fixes exact inputs, and the artifact README provides invocation commands.

	Epoch 4		Epoch 8		Epoch 16	
Seed	Actual	Probe	Actual	Probe	Actual	Probe
201	-5.6500	+0.9750	-17.4500	-0.1125	-6.9125	+0.3500
202	-12.4375	-0.0750	-8.8625	+0.0250	-3.0375	+0.0875
203	-13.1750	+0.8125	-11.2625	+0.0875	-0.4000	+0.0500
204	-8.2500	+1.2625	-13.9875	+0.1750	-4.7750	+0.8125
205	-27.8000	+0.9125	-12.8250	+0.6625	-2.0750	-0.5500
206	-7.6625	+0.5625	-4.4625	+1.1250	+0.1500	+1.6250
207	-9.5375	+0.3125	-6.5875	-0.2250	-0.8125	-0.4250
208	-15.2375	+2.4750	-1.9500	+0.1125	-2.2250	+0.5625
209	-6.8000	+0.7250	-5.6250	+1.0250	-6.3125	+2.2375
210	-4.9625	+1.7500	-1.4375	-0.5500	+0.2250	+1.2000

Table 31: Complete paired ER-minus-SGD all-ten current-accuracy differences (percentage points). Each value averages the same four noninitial tasks within a seed. Actual and probe columns refer to different models; repeated budgets and tasks are not additional independent seeds.

O EARLY-STEP RESOLUTION AND CANDIDATE ELIGIBILITY

The thirteen-step diagnostic in Section 5.3 could miss early feasible models and excluded actual checkpoints from selection. This subsequent check holds the twenty final-task histories fixed and changes candidate eligibility in a post-hoc study motivated by the CIFAR-10.1 results.

Fixed policies and new evaluation images. S denotes the original steps $\{0, 1, 2, 4, 8, 16, 32, 64, 128, 157, 314, 471, 628\}$; D adds all integer steps through 16 (24 candidates total). S+A and D+A additionally admit the corresponding original actual checkpoint. Rule 20 requires validation pooled-old accuracy at least 20%. Rule 20+5 also limits its decrease from that history's validation step zero to 5 pp. Both maximize validation current correct counts; ties take the earliest step, with a step-628 probe preceding actual. Empty feasible sets remain NA. These pooled-old constraints allow histories to differ in initial states, attained retention and individual old-task scores.

Selection reuses the original 1,000 CIFAR-10.1 validation images. All 160 selections were sealed before the first model inference on the official 2,000-image CIFAR-10.2 test set (Lu et al., 2020). Only distinct winners and pre/actual/terminal anchors were tested (89 endpoints), not the full candidate grid. The same images are shared across seeds; old/current test denominators are 1,600/400. CIFAR-10.2 was collected from Tiny Images, as was CIFAR-10; its collection and annotation procedures differ. Our byte-level comparison found no exact duplicates with CIFAR-10 train/test or CIFAR-10.1, or within this test set; near-duplicate overlap remains unmeasured.

History	Rule	Policy	n	Old	Current	Violations
SGD	20	S	7	20.46	0.93	3
SGD	20	D	7	20.40	3.54	2
SGD	20	S+A	7	20.46	0.93	3
SGD	20	D+A	7	20.40	3.54	2
SGD	20+5	S	7	20.46	0.93	3
SGD	20+5	D	7	20.40	3.54	2
SGD	20+5	S+A	7	20.46	0.93	3
SGD	20+5	D+A	7	20.40	3.54	2
ER	20	S	10	30.26	17.15	0
ER	20	D	10	24.27	33.08	2
ER	20	S+A	10	28.59	54.70	0
ER	20	D+A	10	28.59	54.70	0
ER	20+5	S	10	36.88	0.85	0
ER	20+5	D	10	36.14	3.28	1
ER	20+5	S+A	10	36.45	6.75	0
ER	20+5	D+A	10	36.09	7.68	1

Table 32: CIFAR-10.2 test means (%). S/D are sparse/dense probe candidates; +A admits actual. Rule 20+5 adds the within-history 5 pp drop limit. n counts validation-feasible histories; violations count test failures of the corresponding rule, retained in all means. All SGD rows omit the same unavailable seeds 201, 206, 208; ER rows include all ten. Do not subtract unequal-denominator method means.

Candidate sensitivity. Under rule 20, D changes four of seven available SGD winners and seven of ten ER winners without changing availability. The ER current gain accompanies lower retained accuracy and two new absolute test violations (seeds 206, 209). Including actual selects it in all ten ER histories, under either grid. Thus the original probe-only comparison depends on its candidate restriction: admitting actual ER recovers the existing actual model.

Relative retention changes the question. Under rule 20+5, ER's D-minus-S current gain is 2.425 pp and eight of ten seed outcomes are unchanged. Actual is feasible and selected only for ER seed 204; nine ER actual checkpoints fail this additional validation constraint. Dense ER policies have one relative test violation (seed 209). The relative rule constrains the two histories differently: SGD has already lost the first three old tasks, whereas ER retains more of them. The comparison therefore remains conditional on those different starting states.

History	Rule	Contrast	n	Δold	Δcurrent	SD current
SGD	20	D−S	7	-0.054	+2.607	3.275
SGD	20	S+A−S	7	+0.000	+0.000	0.000
SGD	20	D+A−D	7	+0.000	+0.000	0.000
SGD	20+5	D−S	7	-0.054	+2.607	3.275
SGD	20+5	S+A−S	7	+0.000	+0.000	0.000
SGD	20+5	D+A−D	7	+0.000	+0.000	0.000
ER	20	D−S	10	-5.994	+15.925	12.156
ER	20	S+A−S	10	-1.675	+37.550	20.228
ER	20	D+A−D	10	+4.319	+21.625	14.605
ER	20+5	D−S	10	-0.738	+2.425	5.289
ER	20+5	S+A−S	10	-0.431	+5.900	18.657
ER	20+5	D+A−D	10	-0.056	+4.400	13.914

Table 33: Within-history paired policy effects on CIFAR-10.2 (pp). Each comparison uses histories feasible under both policies. SD is the sample standard deviation of paired current differences, not a confidence interval. Integer-count fractions are aggregated before decimal rounding.

Rule	Policy	n	Δold	Δcurrent
20	S	7	+9.473	+17.357
20	D	7	+4.536	+26.571
20	S+A	7	+8.393	+54.786
20	D+A	7	+8.446	+52.179
20+5	S	7	+16.045	-0.643
20+5	D	7	+15.045	+0.214
20+5	S+A	7	+15.429	+7.786
20+5	D+A	7	+14.964	+6.500

Table 34: Conditional ER-minus-SGD paired test differences (pp), always on seeds 202, 203, 204, 205, 207, 209, 210. Their initial retention levels differ across methods.

For these seven common seeds, pre-state old accuracy averages 36.679% for ER and 19.839% for SGD. The dense current difference changes from 26.571 pp under rule 20 to 0.214 pp under rule 20+5, but their old-accuracy difference remains 15.045 pp under the latter. The near-equal current scores occur at different initial and attained retention levels. These descriptive policy comparisons use the shared images and validation split specified above.

Seed	History	S	D	S+A	D+A
201	SGD	NA	NA	NA	NA
202	SGD	0: 20.7/0.0	0: 20.7/0.0	0: 20.7/0.0	0: 20.7/0.0
203	SGD	1: 21.9/0.0	7: 21.5/6.3	1: 21.9/0.0	7: 21.5/6.3
204	SGD	1: 21.7/0.0	7: 22.0/3.5	1: 21.7/0.0	7: 22.0/3.5
205	SGD	0: 19.7/0.0*	0: 19.7/0.0*	0: 19.7/0.0*	0: 19.7/0.0*
206	SGD	NA	NA	NA	NA
207	SGD	0: 21.3/0.0	11: 20.4/0.8	0: 21.3/0.0	11: 20.4/0.8
208	SGD	NA	NA	NA	NA
209	SGD	8: 18.4/6.5*	8: 18.4/6.5*	8: 18.4/6.5*	8: 18.4/6.5*
210	SGD	4: 19.4/0.0*	10: 20.1/7.8	4: 19.4/0.0*	10: 20.1/7.8
201	ER	2: 31.2/6.5	3: 24.4/34.5	A: 23.4/56.8	A: 23.4/56.8
202	ER	2: 35.3/1.0	3: 28.1/33.5	A: 27.8/72.0	A: 27.8/72.0
203	ER	2: 30.4/3.3	3: 21.4/20.8	A: 33.0/42.5	A: 33.0/42.5
204	ER	4: 21.9/41.0	4: 21.9/41.0	A: 27.3/59.0	A: 27.3/59.0
205	ER	2: 30.6/17.8	2: 30.6/17.8	A: 31.3/33.5	A: 31.3/33.5
206	ER	4: 25.6/23.0	5: 18.8/45.8*	A: 27.0/56.3	A: 27.0/56.3
207	ER	2: 37.4/0.5	3: 28.5/18.0	A: 25.4/71.0	A: 25.4/71.0
208	ER	2: 36.4/14.0	3: 24.9/39.8	A: 33.5/44.0	A: 33.5/44.0
209	ER	4: 27.6/29.5	5: 17.8/44.8*	A: 27.4/53.3	A: 27.4/53.3
210	ER	4: 26.4/35.0	4: 26.4/35.0	A: 29.9/58.8	A: 29.9/58.8

Table 35: Rule 20: all seed-level selections. Each cell is selected step: test old/current accuracy (%); A denotes actual and NA no validation-feasible candidate. A star marks a test violation; all starred rows remain included. Display rounding to one decimal does not determine constraint membership.

Seed	History	S	D	S+A	D+A
201	SGD	NA	NA	NA	NA
202	SGD	0: 20.7/0.0	0: 20.7/0.0	0: 20.7/0.0	0: 20.7/0.0
203	SGD	1: 21.9/0.0	7: 21.5/6.3	1: 21.9/0.0	7: 21.5/6.3
204	SGD	1: 21.7/0.0	7: 22.0/3.5	1: 21.7/0.0	7: 22.0/3.5
205	SGD	0: 19.7/0.0*	0: 19.7/0.0*	0: 19.7/0.0*	0: 19.7/0.0*
206	SGD	NA	NA	NA	NA
207	SGD	0: 21.3/0.0	11: 20.4/0.8	0: 21.3/0.0	11: 20.4/0.8
208	SGD	NA	NA	NA	NA
209	SGD	8: 18.4/6.5*	8: 18.4/6.5*	8: 18.4/6.5*	8: 18.4/6.5*
210	SGD	4: 19.4/0.0*	10: 20.1/7.8	4: 19.4/0.0*	10: 20.1/7.8
201	ER	2: 31.2/6.5	2: 31.2/6.5	2: 31.2/6.5	2: 31.2/6.5
202	ER	0: 38.9/0.0	0: 38.9/0.0	0: 38.9/0.0	0: 38.9/0.0
203	ER	0: 39.1/0.0	0: 39.1/0.0	0: 39.1/0.0	0: 39.1/0.0
204	ER	0: 31.6/0.0	3: 27.8/15.0	A: 27.3/59.0	A: 27.3/59.0
205	ER	0: 35.6/0.0	0: 35.6/0.0	0: 35.6/0.0	0: 35.6/0.0
206	ER	0: 40.6/0.0	0: 40.6/0.0	0: 40.6/0.0	0: 40.6/0.0
207	ER	2: 37.4/0.5	2: 37.4/0.5	2: 37.4/0.5	2: 37.4/0.5
208	ER	0: 41.5/0.0	0: 41.5/0.0	0: 41.5/0.0	0: 41.5/0.0
209	ER	2: 34.3/1.0	3: 30.6/10.3*	2: 34.3/1.0	3: 30.6/10.3*
210	ER	2: 38.6/0.5	2: 38.6/0.5	2: 38.6/0.5	2: 38.6/0.5

Table 36: Rule 20+5: the same complete seed display. Stars mark failure of either the absolute test threshold or the within-history test drop limit. The relative test check uses the same history's test pre-state.

Seed	History	Pre	Actual	Terminal probe
201	SGD	19.8/0.0	0.0/69.3	0.0/69.3
202	SGD	20.7/0.0	0.0/91.8	0.0/91.8
203	SGD	19.1/0.0	0.0/70.3	0.0/70.3
204	SGD	20.1/0.0	0.0/90.0	0.0/90.0
205	SGD	19.7/0.0	0.0/90.8	0.0/90.8
206	SGD	17.2/0.0	0.0/69.5	0.0/69.5
207	SGD	21.3/0.0	0.0/81.0	0.0/81.0
208	SGD	18.3/0.0	0.0/84.0	0.0/84.0
209	SGD	18.1/0.0	0.0/54.3	0.0/54.3
210	SGD	19.9/0.0	0.0/72.0	0.0/72.0
201	ER	32.7/0.0	23.4/56.8	0.0/72.0
202	ER	38.9/0.0	27.8/72.0	0.0/91.5
203	ER	39.1/0.0	33.0/42.5	0.0/69.3
204	ER	31.6/0.0	27.3/59.0	0.0/91.5
205	ER	35.6/0.0	31.3/33.5	0.0/90.5
206	ER	40.6/0.0	27.0/56.3	0.0/73.5
207	ER	37.8/0.0	25.4/71.0	0.0/82.3
208	ER	41.5/0.0	33.5/44.0	0.0/86.8
209	ER	37.1/0.0	27.4/53.3	0.0/55.3
210	ER	36.6/0.0	29.9/58.8	0.0/76.3

Table 37: Fixed CIFAR-10.2 anchors for all histories: test old/current accuracy (%). These are evaluated independently of validation availability. SGD actual equals its terminal current-only probe.

Execution and independent reconstruction. The frozen R13 protocol adds only sixteen original current-only updates per history (320 total). All model tensors at steps 0, 1, 2, 4, 8, 16 match the archived trajectory, joining the new prefix to the old continuation. The run reuses 260 validation endpoints and measures 240 new ones; the complete campaign took 52.141 seconds with a 3,300-second hard limit. An independent implementation reconstructed all raw counts, 160 selections, test violations and exact-fraction seed aggregates. A separate functional CNN reconstructed every newly evaluated validation and actually scored test state for seed 201 under both methods: all 400,000 logit elements agreed exactly. This functional check verifies selected forward outputs.

The manifest is `configs/r13_early_retention/manifest.json`, SHA256 `0eb4d79e250749185e8963d11e3c541f7d86eb897d618858569afd36781e7fb8`; raw files are in `results/r13_early_retention_v1`. The short NumPy-only entry in Appendix G recomputes all selections and their reported statistics from archived logits. The compact release includes the frozen audit and replay implementations, but not the checkpoints required to repeat functional forwards. The fixed CIFAR-10.2 source and checksum are in `third_party/r13_cifar102_reference/SOURCE.json`; `tools/prepare_r13_external_data.py` verifies a local cache. The packaged numerical analysis runs from the extracted root using the locked environment:

```
python -s -X utf8 analysis/audit_r13_early_retention.py
--output audit/new_r13_audit
python -s -X utf8 tools/build_r13_early_retention_evidence.py
```

Output paths must be new; the final command regenerates display tables from the archived independent audit and records its hash. These commands reproduce the numerical analysis. The reported functional forwards were completed in the locked environment but cannot be repeated from the compact release.

P FULL-BUDGET CIFAR-100 EXTERNAL TEST

Question and fixed comparison. Before these runs, we fixed seeds 3101–3105, SGD and ER, diagnostic tasks 6–9, all endpoints and a joint-direction criterion: negative ER-minus-SGD actual accuracy, positive probe accuracy and positive masked-actual accuracy. The official CIFAR-100 test had been accessed in earlier project experiments; these are new training runs on reused test images.

Seed	Actual	Probe	Masked actual	Joint direction
3101	5.775	13.200	9.575	no
3102	10.075	16.125	13.100	no
3103	9.025	14.925	12.625	no
3104	1.650	10.275	4.800	no
3105	6.850	11.900	9.850	no
Mean	6.675	13.285	9.990	0/5

Table 38: All paired contrasts, ER minus SGD in pp, averaged over tasks 6–9. Actual and probe use all seen candidates; masked actual uses the current ten. The joint direction requires actual < 0, probe > 0 and masked actual > 0 simultaneously. Zero differences would fail the criterion.

Data, network and training. PCG64$(s + 40000)$ permutes the 100 fine classes into ten disjoint tasks of 5,000 training and 1,000 test images each. Initialization uses seed $s + 10000$. The pinned PyCIL ResNet-32 and replacement of batch normalization with eight-group normalization follow Appendix F; the network has 470,004 parameters. At each task boundary, newly activated head rows are restored from the paired initialization. This checkpoint starts both actual and probe; only actual advances the history. Current batches contain 64 images, with a final batch of eight, giving 79 updates per epoch. Each task or auxiliary branch starts an SGD optimizer with momentum 0.9, without Nesterov. Task zero uses 200 epochs, initial learning rate 0.1, multiplicative 0.1 decays after epochs 60, 120 and 170, and coupled weight decay 5×10^{-4}. Later tasks and auxiliary branches use 70 epochs, decays after 30 and 50, and weight decay 2×10^{-4}. Decay includes GN affine parameters and head biases. Numerical settings follow pinned PyCIL replay; current epochs, GN and the sampling below are local adaptations, rather than an official reproduction.

ER stores 20 training examples per old class by fixed seeded random priority. It replays the same number of examples as each current batch, cycling memory permutations as needed. Concatenated batches use mean cross-entropy over all seen outputs. Matching current examples and updates does not match total exposure or computation. Current order and augmentation keys are paired across methods and branches. Augmentation uses four-pixel zero padding, random crop, horizontal flip with probability 0.5, continuous brightness multiplier uniform on $[1-63/255, 1+63/255]$, clipping to $[0, 1]$ and CIFAR-100 normalization. CUDA arithmetic differs from the short study's PIL pipeline. Training uses BF16 autocast with FP32 master parameters and SGD state; evaluation uses FP32. Both use channels-last tensors, deterministic algorithms and disabled TF32. Train-only engineering measurements selected this implementation before formal runs.

Branches and reporting. At tasks 6–9, probe starts from the actual pre-task checkpoint and fresh from the complete saved initialization. Both receive 70 current-only epochs with the same optimizer schedule and seen-class loss. SGD actual and probe are identical operations and explicitly share an endpoint. Fresh is trained once per seed/task and reused across methods with exact checkpoint identity; reuse is recorded separately from executed updates. All ten histories and auxiliary training branches finish before test evaluation. Raw 100-way logits, official indices and labels, class maps and checkpoint hashes are saved. Current and old metrics use only classes and images seen at that boundary. The current-ten mask supplies task identity without changing weights; old accuracy is measured on each same model.

The unit is the paired seed: tasks are equally averaged before seed aggregation. Strict replication requires all five seeds to meet all three signed inequalities. Actual ER-minus-SGD accuracy is positive in every seed, so the joint criterion fails in all five. We report every seed and sample mean/SD. The machine-readable analysis also enumerates all 5^5 seed bootstrap resamples for approximate percentile intervals; coverage is not guaranteed at five seeds. The result delineates the ranking pattern

without isolating a causal budget effect: architecture, replay, optimizer and precision differ from the primary experiment.

Execution and reproduction. A CUDA driver failure interrupted seed 3102 SGD before test evaluation. The failed attempt was retained; two completed seed-3101 jobs were copied with byte-identity checks, and seed 3102 was rerun from initialization under the unchanged protocol. Its completed task-prefix parameter hashes match the original attempt. No partial endpoint or seed exclusion was used. The anonymous supplement provides frozen source, derived per-example decisions, recovery records and NumPy-only analysis. From its root, run:

```
python -I -X utf8 tools/recompute_r15_compact.py
--output NEW_DIRECTORY
```

The compact analyzer recomputes counts and contrasts from the packaged decisions. Checkpoints and original logits are not bundled. Fresh training needs the locked GPU environment and official CIFAR-100 data described in the README.