THE ILLUSION OF GLOBAL CALIBRATION: GROUP-AWARE RISK IN CLASS-INCREMENTAL LEARNING

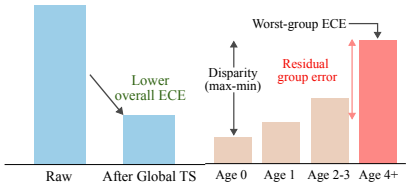
Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable class-incremental learning requires confidence estimates that remain informative after sequential training, because downstream decisions such as selective prediction and human referral depend on per-sample confidence rather than on accuracy alone. Calibration is typically summarized by aggregate scores, yet such averages can hide substantial errors within groups of classes introduced at different times during the training schedule. In this paper, we show that global temperature scaling can improve overall calibration while leaving large errors in specific schedule-age groups, and that the direction of these group-level changes differs across learners. Aggregate improvement therefore does not by itself certify uniformly reliable confidence. To make this discrepancy visible, we introduce a group-aware audit that evaluates aggregate and group-level calibration jointly, together with the direction of confidence bias, the source of calibration data, and the scale of the scores being calibrated. Across class-incremental benchmarks, the discrepancy persists even in high-accuracy models, although its magnitude depends on both the temperature search range and the data used for fitting. These findings support reporting group-level calibration alongside aggregate scores whenever continual learners are compared or deployed.

1 INTRODUCTION

Class-Incremental Learning (CIL) systems must acquire new classes sequentially while retaining previously learned capabilities Parisi et al. (2019); De Lange et al. (2021); Wang et al. (2024); Zhou et al. (2024). In applications that use confidence-aware decisions, including selective prediction, human referral, and uncertainty-aware planning, predicted confidence must reflect the true probability of being correct, a property known as *calibration* Guo et al. (2017); Minderer et al. (2021); Dong et al. (2025). This requirement is especially pressing for CIL. Non-stationarity and catastrophic forgetting create uneven error profiles, so classes encountered at different points of the schedule may exhibit very different reliability. Despite this heterogeneity, evaluation protocols for CIL calibration still rely mainly on aggregate metrics such as Expected Calibration Error (ECE), Negative Log-Likelihood (NLL), and Brier score Li et al. (2024); Hwang et al. (2025). A continual learner can therefore appear reliable under such summaries while retaining severe residual risk within specific subgroups.



(a) Overall ECE (b) Group-wise ECE

Figure 1: Conceptual illustration of aggregate and group-wise calibration disagreement, not a measured result. Global temperature scaling (TS) can reduce overall ECE while leaving uneven group errors.

The reason is that aggregate metrics average over heterogeneous reliability, so strong performance on large or well-calibrated subsets can offset failures in smaller, decision-relevant groups. This masking effect is particularly consequential in CIL because sequential training induces cohorts of classes with different learning histories. Schedule age, the number of experiences since a class was introduced, provides a natural axis for examining these differences, alongside output-derived and label-dependent diagnostics. Figure 1 illustrates this disagreement, and Figure 3 demonstrates it empirically: global temperature scaling (TS) sharply reduces overall ECE, while worst schedule-age group ECE remains

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.

several times larger. When downstream decisions depend on per-sample confidence, this disagreement distorts both model comparison and deployment assessment.

This paper studies calibration evaluation in CIL. Rather than introducing another post-hoc calibrator, we ask a prior question: whether aggregate-only evaluation can support the conclusions it is commonly used to draw. Answering it requires measuring what the aggregate conceals under explicitly reported conditions. We address this question with a *group-aware calibration audit* whose design choices are fixed in advance. The audit combines bounded temperature search with predefined groupings and reports worst-group errors, disparities, and signed confidence gaps alongside the usual aggregate scores. Recording the search boundaries and score scales makes these conditions explicit.

The audit reveals that aggregate improvement can accompany either improvement or deterioration in worst age-group ECE, changing which learner each criterion would select. This discrepancy is not an artifact of poorly trained models: it persists in high-accuracy pretrained models, although accounting for their native score scale reduces its magnitude. The data used for calibration also matter. A specialized class-incremental calibrator, T-CIL Hwang et al. (2025), improves both criteria under matched data, yet a scalar temperature fitted on independent validation data performs better in the evaluated setting. Taken together, these findings connect the interpretation of calibration improvements to the groups, scores, and fitting data behind them.

In summary, our contributions are as follows:

- We identify and characterize a decision-relevant blind spot in CIL calibration evaluation: large improvements in aggregate calibration can coexist with substantial schedule-age group residual error, and the direction of the group-level change is learner dependent.
- We introduce a group-aware calibration audit protocol that standardizes fixed-budget temperature search, predefined exact-age and output-derived groupings, worst-group and disparity reporting, signed direction analysis, and score-scale metadata, so that evaluation conditions are explicit and comparable across studies.
- We provide empirical evidence on what moderates the blind spot, covering the model accuracy regime and score scale, the provenance of calibration data, and the choice between generic and class-incremental-specific calibrators, evaluated on SplitCIFAR-100, TinyImageNet, and ImageNet-100.

2 Related Work

Calibration of Deep Neural Networks. Overconfidence in neural networks motivates post-hoc correction through temperature scaling Dong et al. (2025); Guo et al. (2017) and richer transformations such as Dirichlet calibration Kull et al. (2019). Reliability under distribution shift Ovadia et al. (2019), selective classification, and human referral Geifman & El-Yaniv (2017); Moon et al. (2020) also depends on confidence quality. Work on ECE estimation addresses statistical limitations of the metric Nixon et al. (2019); Vaicenavicius et al. (2019); Roelofs et al. (2022). We address a complementary question: what aggregate evaluation conceals when sequential training structures the evaluation population.

Calibration in Class-Incremental Learning. Forgetting affects both accuracy and confidence on older tasks De Lange et al. (2021); Wang et al. (2024); Li et al. (2024); Yu et al. (2024); Lai et al. (2026). Training-time methods modify representations, replay, or classifiers Singh et al. (2020); Lin et al. (2024); Ji et al. (2023); Liu et al. (2026), whereas continual post-hoc methods use replayed or held-out validation data Li et al. (2024); Hwang et al. (2025); Serra & Buettner (2025). DATS Serra & Buettner (2025) studies task heterogeneity and worst-task calibration through batch-specific prototype-distance estimates. Our evaluation study complements this perspective by separating criterion-dependent model selection, signed and operational errors, fitting-data sources, and score scales.

Hidden Stratification and Subgroup Auditing. Aggregate metrics can obscure subgroup failures, as shown by work on hidden stratification and conditional calibration Pleiss et al. (2017); Kumar et al. (2019); Oakden-Rayner et al. (2020); Kearns et al. (2018). Responses include multicalibration Hebert-Johnson et al. (2018) and worst-group robust optimization Sagawa et al. (2020). We bring this

measurement perspective to CIL through schedule-age auditing and evidence on the conditions that moderate residual group error.

3 Setup and Audit Protocol

3.1 Problem Setting

In class-incremental learning, a model f_θ observes a sequence of experiences $\{E_1, \ldots, E_T\}$, each introducing a disjoint set of new classes, and is evaluated after the final experience on the joint class space without task identity at inference. Although the test stream is evaluated as a single population, its classes entered training at different times. Let $\tau(c) \in \{1, \ldots, T\}$ denote the introduction experience of class c. Sample (x_i, y_i) has schedule age $a_i = T - \tau(y_i)$, with age zero denoting the most recent classes. Unlike age, replay exposure counts repeated presentations during optimization. Section 4.5 examines both variables.

Our primary reference calibrator is *global temperature scaling* Guo et al. (2017), which rescales logits $z = f_\theta(x)$ by a learned positive scalar t^*. Because a positive rescaling does not change the argmax, temperature scaling adjusts confidence while leaving accuracy unchanged. We measure calibration primarily with top-label ECE using 15 equal-width bins Guo et al. (2017); Nixon et al. (2019):

$$\mathrm{ECE} = \sum_{m=1}^{15} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{1}$$

where B_m is confidence bin m, n is the sample count, and acc and conf are empirical top-1 accuracy and mean confidence. Group ECE applies the same definition within each group separately, using the same equal-width boundaries on $[0, 1]$, with counts normalized within each group and empty bins contributing zero. We also report accuracy, NLL, and Brier score, defined in Appendix A.1.

The difference between the pooled and the group-level computation is not cosmetic. Opposing group biases can cancel within a shared bin, as explained in Appendix A.8. This cancellation matters even when age groups contain equally many examples. Within a confidence bin, one cohort can be overconfident and another underconfident, and pooling them can make mean confidence agree with accuracy although neither cohort is well calibrated. Group ECE takes the absolute discrepancy within each cohort before aggregation and therefore retains this heterogeneity. The audit asks whether a favorable pooled score also describes the groups on which a reliability claim depends, rather than treating any particular partition as a complete description of risk.

3.2 Calibration Data Sources

Post-hoc calibration requires data that represent errors on unseen examples, but continual-learning protocols differ in which data they retain. We distinguish a retrospective audit of the final test stream from comparisons of native calibration streams.

Retrospective stratified test-stream audit. The primary audit splits the final test stream by class-stratified sampling into a calibration split D^{calib} (20%) and an evaluation split D^{eval} (80%). Stratification ensures that every class, and therefore every schedule-age group, is represented in both splits. Calibrators are fitted only on D^{calib}, and all reported metrics are computed only on D^{eval}. This is a retrospective audit of the test stream, not an independent validation stream.

Native calibration streams. We compare final-experience validation data, training replay combined with current validation, and a held-out validation pool. The pool contains ten examples per class from every experience, excluded from learner optimization, for 1,000 examples in total. This design adapts calibration with replayed validation data Li et al. (2024) to a class-balanced pool, as detailed in Appendix C.

Section 4.3 compares these sources explicitly, without treating unequal data budgets as a matched algorithm comparison. Secondary coarse-age audits retain their chronological 20/80 test-stream split, which fits on the first two experiences rather than sampling every class.

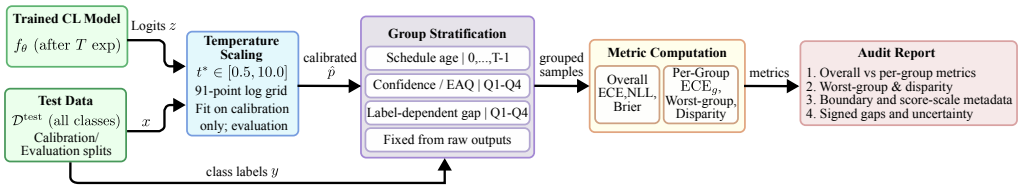


Figure 2: Group-aware calibration audit. Fitting uses only the calibration split. A separate evaluation split supports aggregate and group-level reporting with fixed group membership.

3.3 Group-Aware Audit Protocol

Figure 2 summarizes the group-aware audit. Its search budget, grouping axes, and reporting rules are fixed in advance. We report means and sample standard deviations across seeds unless stated otherwise, with primary per-seed results in Appendix Table 28.

Temperature search specification. The primary audit minimizes calibration NLL over 91 logarithmically spaced temperatures in $[0.5, 10]$. Prototype-score audits use a scale-aware 201-point log-grid in $[0.005, 100]$, with at most two tenfold endpoint extensions triggered by a boundary optimum on calibration data alone. Evaluation data never select the grid. Appendix A.2 describes the grid and the separate limited-memory BFGS (L-BFGS) comparison Liu & Nocedal (1989). Secondary binning and split-fraction checks appear in Appendix H.

Group definitions. The audit evaluates four grouping axes, each exposing a different aspect of the residual risk. **Schedule-age groups** use one exact age per experience, giving 800 evaluation samples per group on the primary benchmark, while secondary audits use coarse bins. **Confidence groups** are quartiles of $\max_c p_c^{\text{raw}}(x)$, where $p^{\text{raw}}(x) = \text{softmax}(z(x))$. **Gap groups** are quartiles of raw confidence minus the correctness indicator, making them label-dependent diagnostics. **Expected-age quantile (EAQ) groups** partition the inference-available score $s(x) = \sum_{c \in \mathcal{C}_{\text{seen}}} p_c^{\text{raw}}(x)[T - \tau(c)]$ into quartiles, where $\mathcal{C}_{\text{seen}}$ is the seen-class set. Unlike true age, this score can be computed at inference time without labels. These memberships remain fixed from raw predictions during pre/post comparisons. Two conditional fits then probe the value of group information. Age-Conditional Temperature Scaling (ACTS) fits a temperature per age group as an oracle diagnostic, whereas EAQ-conditional scaling uses the proxy groups, with quantile boundaries fitted on calibration data and transferred unchanged to evaluation. Appendix A gives further details.

Group-level metrics and protocol metadata. We report per-group metrics, worst-group ECE ($\max_g \text{ECE}_g$), and disparity ($\max_g \text{ECE}_g - \min_g \text{ECE}_g$). The signed gap $\text{conf}(g) - \text{acc}(g)$ distinguishes overconfidence from underconfidence. Fitted temperatures, boundary indicators, and score-to-probability mappings accompany the reported metrics, so that a constrained fit is not mistaken for an unconstrained optimum.

3.4 Experimental Setup

The primary **SplitCIFAR-100** Krizhevsky (2009) benchmark has ten experiences of ten classes, with class order shuffled by seed. We audit **Naive**, **EWC** Kirkpatrick et al. (2017), **LwF** Li & Hoiem (2018), and **DER++** Buzzega et al. (2020), trained from scratch on ResNet-18 over five seeds each. These learners span the main training mechanisms in class-incremental learning: unregularized fine-tuning, parameter regularization, distillation, and replay. The separate **calibration-stream benchmark** trains **Experience Replay (ER)** and DER++ with per-experience held-out validation data for the comparison in Section 3.2.

To examine whether the findings depend on the accuracy regime, we audit **ACMap** Fukuda et al. (2025), a high-accuracy method that combines a pretrained ViT-B/16 backbone with adapter merging and a centroid-prototype classifier, on SplitCIFAR-100 and **Tiny-ImageNet** (200 classes, 64×64) over three seeds each. Because ACMap produces cosine-prototype scores rather than softmax logits, it exercises the scale-aware branch of the audit described above. Matched comparisons with **CoCa** Ji et al. (2023) against its paired DER++ reference, and **DGC** Lin et al. (2024) with its calibration component enabled or disabled, compare complete training paths under their respective native implementations. Cross-dataset recurrence checks on Tiny-ImageNet and **ImageNet-100** with the

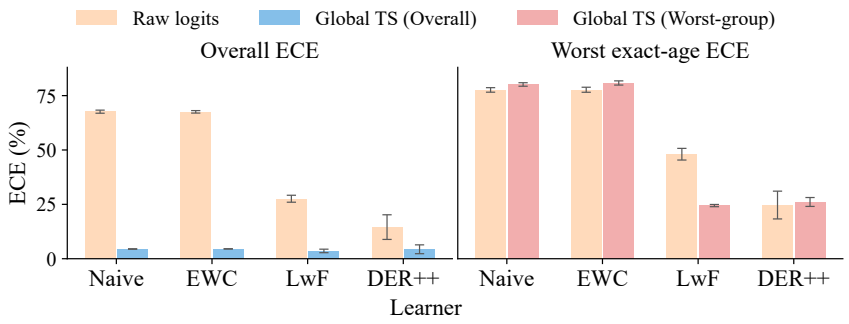


Figure 3: Primary stratified SplitCIFAR-100 audit before and after global TS. Bars show mean overall and worst exact-age ECE, with sample standard deviations over five seeds.

Table 1: Primary SplitCIFAR-100 audit under the retrospective stratified test-stream split of Section 3.2. Values are mean ± sample standard deviation over five seeds. Oracle ACTS conditions on exact-age information. Naive and EWC global fits reach $t_{\max} = 10$ in all seeds.

		Overall ECE (%)			Worst exact-age ECE (%)		
Learner	**Acc. (%)**	**Raw**	**Global TS**	**Oracle ACTS**	**Raw**	**Global TS**	**Oracle ACTS**
Naive	8.6±0.1	67.62±0.71	4.47±0.11	3.85±0.13	77.62±1.04	80.13±0.81	4.31±0.21
EWC	8.7±0.1	67.55±0.57	4.53±0.06	3.82±0.08	77.70±1.16	80.82±0.97	4.30±0.11
LwF	26.0±1.5	27.58±1.60	3.58±0.81	5.67±0.95	48.06±2.68	24.43±0.51	13.79±2.63
DER++	31.8±2.9	14.54±5.67	4.34±2.03	4.88±1.56	24.68±6.39	26.09±2.05	11.50±0.42

four primary learners, and extension baselines including L2P Wang et al. (2022), CL-LoRA He et al. (2025), IDER Liu et al. (2026), and X-DER Boschini et al. (2023), are reported in Section 4.4 and Appendix A.19.

The primary from-scratch benchmark operates in a low-accuracy regime, which makes aggregate miscalibration large and therefore useful for diagnosing the audit itself. The high-accuracy ACMap audits provide the complementary regime, and the two are never pooled or ranked against each other. Method-specific regularization, optimization, and architectural details are provided in Appendix A.

4 Experimental Results

4.1 Aggregate Improvement Coexists with Group-Level Residual Risk

We begin with the primary audit, which asks whether aggregate calibration improvement implies group-level improvement. Figure 3 and Table 1 show the core discrepancy on SplitCIFAR-100 under the retrospective stratified audit. Global temperature scaling substantially reduces overall ECE for every learner, bringing the aggregate down to a few percent. Worst exact-age ECE remains much larger. For Naive and EWC it stays about as large as before scaling, and for LwF and DER++ it remains several times the calibrated aggregate.

The direction of the group-level change also depends on the learner. Appendix Table 28 shows a slight increase in worst exact-age ECE for Naive and EWC in every seed, a decrease exceeding 20 percentage points for LwF, and mixed directions for DER++. For Naive and EWC, Appendix Tables 30 and 31 show a shift from older overconfident cohorts to the most recent, strongly underconfident group. This pattern has a direct mechanistic reading: a large temperature fitted to the overconfident majority also flattens recent-class predictions. Thus, aggregate improvement establishes neither group-level improvement nor group-level harm.

This ambiguity is not only descriptive. It also affects model selection, because overall ECE and worst exact-age ECE select different learners on two of five seeds in Appendix Table 8. For Naive and EWC, the global temperature reaches the upper boundary in every run, so their reported fit is constrained by the declared search budget rather than being an unconstrained optimum.

Table 2: High-accuracy audit of ACMap (pretrained ViT-B/16) on two datasets. Values are mean $\pm$ sample standard deviation over three seeds. Scale-aware fits use $[0.005, 100]$. The last column reports global fitted-temperature ranges or conditional boundary hits/total fits. The restricted-grid global fits all hit their lower boundary.

Dataset	Acc. (%)	Calibration path	Overall ECE (%)	Worst age ECE (%)	Temperature fit
SplitCIFAR-100	89.5±0.3	Raw scores	87.37±0.29	91.50±1.85	–
		Global TS, restricted $[0.5, 5]$	85.23±0.30	89.24±1.84	0.5000–0.5000
		Global TS, scale-aware grid	1.94±0.21	6.55±1.55	0.0513–0.0539
		EAQ-conditional TS, scale-aware grid	1.66±0.06	6.15±1.39	0/12 boundary
		Oracle ACTS, scale-aware grid	1.78±0.36	6.56±1.33	0/30 boundary
Tiny-ImageNet	85.7±0.4	Raw scores	84.71±0.43	88.08±1.04	–
		Global TS, restricted $[0.5, 5]$	83.74±0.43	87.05±1.03	0.5000–0.5000
		Global TS, scale-aware grid	1.25±0.21	4.75±0.72	0.0539–0.0539
		EAQ-conditional TS, scale-aware grid	1.15±0.14	4.64±0.84	0/12 boundary
		Oracle ACTS, scale-aware grid	1.15±0.18	3.93±0.48	0/30 boundary

To test whether the residual error reflects missing group information, the audit includes the oracle ACTS diagnostic. ACTS substantially reduces worst exact-age ECE for all four learners in Table 1. For Naive and EWC, it also improves overall ECE relative to global TS, showing that group improvement need not sacrifice aggregate fit. This diagnostic uses true age to reveal the effect of additional group information, rather than providing a deployable calibrator or a performance bound.

The comparison separates two questions that an aggregate-only evaluation conflates. Global TS improves the pooled summary, whereas the age-conditioned fit tests how much residual error can be reduced when each cohort receives its own temperature. The large difference between these fits for Naive and EWC contrasts with the smaller difference for LwF and DER++. Together with the signed gaps, this contrast identifies where a shared temperature is insufficient without attributing every group residual to the same direction of bias.

4.2 High-Accuracy Learners and the Score Scale

A natural objection is that the primary benchmark operates in a low-accuracy regime, where calibration failure may simply reflect failed learning. Table 2 answers this objection with ACMap, a pretrained ViT-B/16 method that reaches high accuracy on both SplitCIFAR-100 and Tiny-ImageNet. Its prototype classifier also changes the nature of the scores being calibrated, which turns out to matter as much as the accuracy regime itself.

The discrepancy persists at high accuracy, although its magnitude shrinks. Under the scale-aware grid in Table 2, worst exact-age ECE remains about three to four times the calibrated aggregate. EAQ-conditional scaling and oracle ACTS stay within about one percentage point of global TS, so their additional group information produces only small changes after appropriate rescaling.

The same table exposes a second, distinct phenomenon: a measurement failure caused by score scale. ACMap uses cosine-prototype scores with fitted global temperatures near 0.05. A restricted comparison grid $[0.5, 5]$, distinct from the primary grid $[0.5, 10]$, excludes that range entirely. Every restricted fit reaches its lower boundary and leaves overall ECE close to the raw value, whereas the scale-aware grid gives interior global optima and much smaller errors. A poorly matched search range can therefore manufacture an apparent calibration failure in a highly accurate model.

The useful distinction is between the residual after an appropriate fit and the error imposed by an inappropriate search range. The former remains visible in the worst-age column, while the latter dominates the restricted-grid result. Thus, recurrence at high accuracy supports the need for group reporting, but not the transfer of the large error magnitudes from the from-scratch benchmark. Boundary indicators make these two interpretations distinguishable.

4.3 Calibration Data Source Changes the Outcome

Calibration data substantially change the outcome even with checkpoints and evaluation data fixed. In Table 3, global TS on the held-out pool achieves much lower overall ECE than the alternatives. Current-validation-only TS can instead increase DER++ ECE relative to raw predictions. Among scalar TS sources, training replay combined with current validation gives the largest error, consistent with a mismatch between optimized training examples and unseen test errors. The held-out and replay-

Table 3: Calibration-source comparison on the calibration-stream benchmark (ER and DER++, five seeds each). Calibrators share final checkpoints and the evaluation test stream but differ in calibration data and transformation. Values are mean ± sample standard deviation; ECE and accuracy are percentages. ECE is lower-is-better and accuracy higher-is-better. Held-out-pool temperatures range from 2.47 to 3.02 for ER and 1.15 to 1.77 for DER++, with zero boundary hits on $[0.5, 10]$.

Learner	Calibrator	Calibration data	n	Overall ECE (%)	Worst exact-age ECE (%)	Acc. (%)
ER	Raw logits	–	0	55.76±4.39	68.81±5.09	23.4±1.8
	Global TS	current validation	100	51.34±2.67	64.13±3.27	23.4±1.8
	Global TS	training replay + current val.	2100	63.22±2.99	76.73±3.89	23.4±1.8
	T-CIL adaptation	training replay + current val.	2100	33.66±5.13	45.57±4.70	23.4±1.8
	Vector scaling	training replay + current val.	2100	58.05±3.17	72.76±2.46	26.5±1.5
	Global TS	held-out validation pool	1000	13.98±1.81	27.26±2.03	23.4±1.8
DER++	Raw logits	–	0	23.78±10.68	34.54±12.04	25.9±5.5
	Global TS	current validation	100	35.89±8.55	47.18±10.92	25.9±5.5
	Global TS	training replay + current val.	2100	44.02±8.88	55.77±10.58	25.9±5.5
	T-CIL adaptation	training replay + current val.	2100	22.70±4.62	33.35±5.07	25.9±5.5
	Vector scaling	training replay + current val.	2100	56.47±3.43	68.06±2.89	18.0±2.2
	Global TS	held-out validation pool	1000	5.19±2.45	29.89±6.82	25.9±5.5

Table 5: Overall and age-group ECE on ImageNet-100 under coarse age bins and the chronological calibration split. Values are mean ± std over three seeds. Bold and underline indicate the best and second-best results among methods, respectively. The last column reports the fitted temperature range and upper-bound saturations ($t_{\max} = 10$).

	Overall ECE (%)		Worst-Group ECE (%)		t^* / Hit $t_{\max}$
Method	Raw	TS	Raw	TS	Range / count
Naive	55.5±1.5	7.1±0.4	64.1±2.0	79.8±2.6	10.00–10.00 / 3/3
EWC	56.5±1.3	**7.0±0.4**	65.6±2.4	79.6±3.0	10.00–10.00 / 3/3
LwF	8.3±1.9	21.1±4.3	**16.0±2.7**	**37.5±3.5**	1.77–2.24 / 0/3
DER++	**4.4±0.3**	12.5±2.6	26.3±2.6	39.2±5.9	1.23–1.45 / 0/3

based comparisons vary both data provenance and budget, so they characterize data dependence rather than isolate a calibrator advantage.

Under matched information, the comparison is more favorable to the specialized calibrator. The T-CIL adaptation Hwang et al. (2025) roughly halves overall ECE and substantially reduces worst exact-age ECE relative to replay-plus-validation TS. The group gap nevertheless persists, and a simpler scalar fit on the held-out pool attains lower error, underscoring the importance of fitting data. Appendix Table 25 separately evaluates T-CIL on three independently trained ER learners. That comparison averages across stages and uses a different training protocol, so it is not ranked against this final-stage evaluation.

Vector scaling illustrates why accuracy must accompany calibration metrics. On the same replay-augmented split in Table 3, vector scaling changes top-1 accuracy in opposite directions, raising it for ER and lowering it for DER++, because per-class weighting alters the argmax. Reporting only scalar calibration metrics would hide this operational effect entirely.

Table 4: Overall and worst age-group ECE on Tiny-ImageNet under chronological calibration split (mean±std over three seeds). **Bold** and underline denote best and second-best entries per column.

	Overall ECE (%)		Worst-Group ECE (%, age)	
Learner	Raw	TS	Raw	TS
Naive	61.1±1.2	6.5±0.4	69.1±1.2	69.3±2.7
EWC	62.0±0.8	6.6±0.4	70.0±0.7	70.4±2.7
LwF	23.4±0.8	15.7±1.7	33.0±2.8	33.5±4.9
DER++	**5.5±0.6**	**3.3±0.7**	**20.4±0.5**	**31.4±1.8**

These comparisons also clarify what constitutes matched evidence. Comparing T-CIL with scalar TS on the same replay-plus-validation samples tests the transformation under a shared information budget. Comparing either fit with the held-out pool instead tests a different source of calibration information. A lower error in the latter comparison is useful for protocol design, but cannot by itself establish that scalar scaling is intrinsically better than a class-incremental-specific procedure.

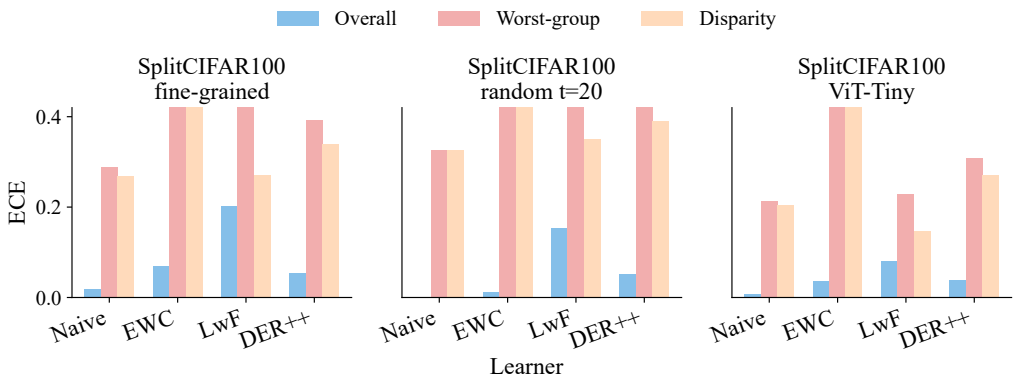


Figure 4: Cross-setting comparison on additional SplitCIFAR-100 configurations under coarse age bins, reporting overall ECE, worst-group ECE, and disparity after global TS. Each bar is a seed mean.

Table 6: Grouping comparison on the secondary SplitCIFAR-100 chronological-split audit under the chronological calibration split. The table reports mean post-scaling overall ECE, worst-group ECE, and disparity for each grouping axis. Bold and underline indicate the grouping axes that expose the highest and second-highest subgroup risk and disparity, respectively.

Grouping	Mean TS overall	Mean TS worst-group	Mean TS disparity
age	0.090	**0.598**	**0.508**
EAQ	0.090	0.205	0.116
confidence quartiles	0.090	0.209	0.119
gap quartiles	0.090	<u>0.579</u>	<u>0.489</u>

4.4 Recurrence Across Datasets and Training Mechanisms

The discrepancy recurs beyond the primary setting. Figure 4 presents supporting three-seed means for a fine-grained semantic split, a 20-experience sequence, and ViT-Tiny. Tables 4 and 5 show larger worst-age than overall ECE on Tiny-ImageNet and ImageNet-100 under coarse age bins. Scaling need not improve either metric: on ImageNet-100, overall ECE increases for LwF and DER++, and worst-age ECE more than doubles for LwF. Appendix A.19 reports the corresponding mismatch for L2P, CL-LoRA, IDER, and X-DER.

Matched training comparisons reveal opposing average trade-offs in Appendix Table 24. Relative to paired DER++, CoCa raises post-scaling overall ECE while lowering worst exact-age ECE on both datasets. Enabling DGC gives the opposite, smaller tendency on Tiny-ImageNet. These comparisons concern complete training paths, with mixed per-seed directions, and do not establish a universal mechanism ranking.

4.5 The Structure of the Residual Risk

The grouping axis is itself a design choice, and it determines which residual errors become visible. Under the secondary chronological split, Table 6 shows larger mean worst-group ECE for coarse schedule age and label-dependent gap quartiles than for confidence or EAQ. This average ordering is not universal: gap groups expose larger errors than age groups for LwF and DER++ in Appendix Table 33. Their worst-group subsets overlap only modestly, as shown in Appendix G.1. Under the primary audit, EAQ-conditional scaling also fails to improve true-age worst-group ECE uniformly in Appendix Table 32. An inference-available proxy therefore complements rather than replaces age-based evaluation.

Unsigned ECE also does not determine bias direction or selective risk. In Appendix Table 19, the worst-ECE age group is underconfident in every seed for Naive, EWC, and DER++, but overconfident in four of five seeds for LwF. For the former three learners, this group never coincides with the worst selective-risk group at 70% coverage. These are distinct diagnostic quantities, further examined in Appendix A.16.

This distinction changes how the audit should inform a decision rule. An underconfident group may contain correct predictions that a confidence threshold unnecessarily rejects, whereas an overconfident group may contribute errors that the same threshold retains. ECE measures confidence–accuracy agreement, not either outcome directly. Signed gaps help interpret the calibration residual, and selective-risk measurements test the consequence for the retained set. Reporting them together avoids treating the largest calibration error as synonymous with the largest decision risk.

Finally, age retains an incremental association with per-class NLL after controlling for class identity, seed, and replay exposure in the multi-stage panel of Appendix Table 26. The association is larger for ER than DER++ and weak for Brier score and signed gap. A separate sequence-length analysis in Appendix Table 27 shows monotonically increasing raw ECE for Naive but non-monotonic behavior for DER++. The 100-experience endpoint is a stress case, not evidence for a smooth scaling law.

5 Discussion

Calibration improvements should be evaluated at the level of the reliability claim. Schedule age supplies an interpretable evaluation axis, while confidence and EAQ provide complementary inference-available views. The size-matched permutation control in Appendix Table 11 shows that the chronological age residual is not explained by group size alone. Thus, reporting groups adds information that an aggregate score and group counts do not provide.

Diagnosing that residual and correcting it are separate questions. Oracle ACTS substantially reduces exact-age error in Table 1, but the coarse-age controls in Appendix A.13 do not give a uniform ranking across learners. The contrast emphasizes the role of calibration support across groups. True age is useful for evaluation without becoming a required inference-time input.

The practical recommendation is to report the group error vector and signed gaps with the fitting-data source, score mapping, and search boundaries. Comparisons should match both calibration data and evaluated checkpoints. Operational metrics remain necessary because calibration error does not identify the least reliable retained group. In the secondary decision simulation, ACTS has the lowest overall selective risk at 50% coverage, whereas global TS has the lowest worst-age risk. At 70%, ACTS leads both criteria but still leaves a large group risk gap, as shown in Appendix Table 18. These distinctions allow applications to choose their own temporal priorities without treating one aggregate score as a universal reliability guarantee Kearns et al. (2018). Appendix K discusses broader impacts.

The evidence has several limits. Naive and EWC behave similarly in the evaluated low-accuracy setting and should not be counted as independent mechanisms. Their error magnitudes do not transfer directly to high-accuracy ACMap. Age and replay exposure are associational variables, and true age requires evaluation labels. Three-seed cross-setting checks offer supporting evidence rather than exhaustive coverage of architectures, schedules, or label spaces. We do not report a matched DATS Serra & Buettner (2025) comparison, whose batch-dependent procedure requires a separate controlled study.

Uncertainty estimates also depend on the audit design. Class-tail bootstrap intervals condition on the fitted calibrator rather than resampling the calibration split, as explained in Appendix A.17. The secondary sensitivity checks in Appendix H concern chronological splits and coarse age bins, not uncertainty in the primary exact-age protocol. These distinctions determine which sources of variation each comparison can assess.

6 Conclusion

Aggregate calibration improvement does not establish reliable confidence across schedule-age groups in class-incremental learning. Our audit exposes residual group error across datasets and accuracy regimes. The direction of group-level change depends on the learner and calibration data, while the score scale can substantially alter its measured magnitude. Reporting group errors, signed gaps, and selective risk alongside aggregate metrics, with explicit fitting-data and search specifications, provides a firmer basis for comparing continual learners and assessing confidence where decisions are made. Age-conditional fits characterize learner–calibrator compatibility, as detailed in Appendix A.15, rather than establishing a universal correction or causal mechanism.

REFERENCES

Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5497–5512, 2023.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *NeurIPS*, volume 33, pp. 15920–15930, 2020.

Antonio Carta, Lorenzo Pellegrini, Andrea Cossu, Hamed Hemati, and Vincenzo Lomonaco. Avalanche: A pytorch library for deep continual learning. *Journal of Machine Learning Research*, 24(363):1–6, 2023.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.

Jinzong Dong, Zhaohui Jiang, Dong Pan, Zhiwen Chen, Qingyi Guan, Hongbin Zhang, Gui Gui, and Weihua Gui. A survey on confidence calibration of deep learning-based classification models under class imbalance data. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9): 15664–15684, 2025.

Takuma Fukuda, Hiroshi Kera, and Kazuhiko Kawamoto. Adapter merging with centroid prototype mapping for scalable class-incremental learning. In *CVPR*, pp. 4884–4893, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, volume 30, pp. 4878–4887, 2017.

Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *ICLR*, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, volume 70, pp. 1321–1330, 2017.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *CVPR*, pp. 30534–30544, 2025.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *ICML*, volume 80, pp. 1939–1948, 2018.

Seong-Hyeon Hwang, Minsu Kim, and Steven Euijong Whang. T-CIL: temperature scaling using adversarial perturbation for calibration in class-incremental learning. In *CVPR*, pp. 15339–15348, 2025.

Zhong Ji, Jin Li, Qiang Wang, and Zhongfei Zhang. Complementary calibration: Boosting general continual learning with collaborative distillation and self-supervision. *IEEE Transactions on Image Processing*, 32:657–667, 2023.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*, pp. 2564–2572, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, volume 32, pp. 12295–12305, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *NeurIPS*, volume 32, pp. 3787–3798, 2019.

Guannan Lai, Da-Wei Zhou, Xin Yang, and Han-Jia Ye. The lie of the average: How class incremental learning evaluation deceives you? In *ICLR*, 2026.

Lanpei Li, Elia Piccoli, Andrea Cossu, Davide Bacciu, and Vincenzo Lomonaco. Calibration of continual learning models. In *CVPR Workshops*, pp. 4160–4169, 2024.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Wei-Chen Lin, Jiaxiang Chen, Ruomin Huang, and Hu Ding. An effective dynamic gradient calibration method for continual learning. *arXiv preprint arXiv:2407.20956*, 2024.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Zhanwang Liu, Yuting Li, Haoyuan Gao, Yexin Li, Linghe Kong, Lichao Sun, and Weiran Huang. IDER: IDempotent experience replay for reliable continual learning. In *ICLR*, 2026.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *NeurIPS*, volume 34, pp. 15682–15694, 2021.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, pp. 7034–7044, 2020.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, pp. 38–41, 2019.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *CHIL*, pp. 151–159, 2020.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, volume 32, pp. 13969–13980, 2019.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *NeurIPS*, volume 30, pp. 5680–5689, 2017.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. In *AISTATS*, volume 151, pp. 4036–4054, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Giuseppe Serra and Florian Buettner. DATS: Distance-aware temperature scaling for calibrated class-incremental learning. *arXiv preprint arXiv:2509.21161*, 2025.

Pravendra Singh, Vinay Kumar Verma, Pratik Mazumder, Lawrence Carin, and Piyush Rai. Calibrating CNNs for lifelong learning. In *NeurIPS*, volume 33, pp. 15579–15590, 2020.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *AISTATS*, volume 89, pp. 3459–3467, 2019.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pp. 139–149, 2022.

Hao Yu, Xin Yang, Xin Gao, Yihui Feng, Hao Wang, Yan Kang, and Tianrui Li. Overcoming spatial-temporal catastrophic forgetting for federated class-incremental learning. In *ACM MM*, pp. 5280–5288, 2024.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024.

A Additional Experimental Details

A.1 Proper Scores

For calibrated class probabilities $p_c(x_i)$ over C classes and n evaluation samples, we compute

$$\mathrm{NLL} = -\frac{1}{n}\sum_{i=1}^{n}\log p_{y_i}(x_i), \qquad \mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\big(p_c(x_i) - \mathbf{1}[y_i = c]\big)^2. \tag{2}$$

These scores assess the full predictive distribution, complementing top-label ECE.

A.2 Temperature Search Grid

To identify the optimal global temperature t^*, we use a logarithmically spaced grid of 91 temperature values bounded by $t_{\min} = 0.5$ and $t_{\max} = 10.0$. The grid points are defined as $t_i = t_{\min} \cdot (t_{\max}/t_{\min})^{i/90}$ for $i \in \{0, \ldots, 90\}$. This spacing allocates higher resolution near $t = 1$, where optimal temperatures frequently reside, while covering strongly overconfident or underconfident distributions.

Table 7 compares the bounded grid with continuous limited-memory BFGS (L-BFGS) optimization Liu & Nocedal (1989) on a separate twelve-run, three-seed training set. A seed-zero random permutation defines a non-stratified 20/80 calibration/evaluation split, unlike both the primary stratified split and the chronological permutation control. L-BFGS starts at the grid optimum and can exceed its bounds. The LwF specification uses distillation weight 1 and temperature 2, while DER++ uses memory 200 and coefficients $\alpha = 0.1$, $\beta = 0.5$, unlike Table 10. Within this auxiliary comparison, differences in overall and worst-age ECE are below 0.01 even when continuous fits exceed the grid boundary. This does not establish optimizer invariance for the primary checkpoints.

Table 7: Bounded 91-point log-grid and continuous L-BFGS fitting on a separate twelve-run training set with a randomly permuted, non-stratified 20/80 split and coarse age bins. Four learners and three seeds each. L-BFGS starts at the grid optimum and is not bounded by the grid endpoints.

Learner	Grid t^* range	L-BFGS t^* range	Mean $\lvert\Delta\rvert$ overall ECE	Mean $\lvert\Delta\rvert$ worst(age) ECE	Max $\lvert\Delta\rvert$ worst(age) ECE
Naive	10.00–10.00	11.14–11.76	0.0021	0.0023	0.0030
EWC	10.00–10.00	11.70–11.92	0.0028	0.0030	0.0033
LwF	10.00–10.00	12.29–12.31	0.0058	0.0086	0.0096
DER++	2.82–4.97	2.82–4.97	**0.0000**	**0.0000**	**0.0000**

A.3 Age Group Definitions

The primary analysis uses exact schedule-age groups: a model trained over T sequential experiences assigns each evaluation sample from class c the age $a(c) = T - \tau(c)$, where $\tau(c)$ is the experience index at which class c first appears, and each exact age value forms its own group. On the primary SplitCIFAR-100 benchmark with $T = 10$, this yields ten groups of 800 evaluation samples each under the stratified split. Coarser bins, such as $\{0, 1, 2\text{–}3, 4+\}$ experiences ago, appear only in the secondary cross-setting audits and the control analyses, and are labeled as coarse wherever they are used. Schedule age serves exclusively as an exogenous evaluation audit lens and is not assumed to be directly available at inference time.

A.4 Confidence, Expected-Age, and Gap Groupings

Confidence groups partition evaluation samples into quartiles of $\max_c p_c^{\mathrm{raw}}(x)$. The expected-age score $s(x)$ defined in Section 3.3 averages the ages of seen classes under the raw predictive distribution. EAQ denotes its quantile groups, not the scalar score itself. For EAQ-conditional scaling, the 25th, 50th, and 75th percentile boundaries are fitted only on calibration scores and transferred to evaluation, whose groups need not have equal sizes. Gap groups use $\mathrm{gap}(x) = \max_c p_c^{\mathrm{raw}}(x) - \mathbf{1}[\hat{y}^{\mathrm{raw}}(x) = y(x)]$ and therefore require labels. Memberships are fixed before calibration is applied, ensuring that raw and calibrated metrics describe the same evaluation populations.

Holding the raw-output group definitions fixed compares calibration paths on the same examples. Redefining groups after scaling would instead change both their membership and their calibration error. Our comparisons use the former estimand throughout.

A.5 Criterion-Dependent Model Selection

Table 8 lists, for each seed of the primary benchmark, the learner preferred by post-scaling overall ECE and the learner preferred by post-scaling worst exact-age ECE. The two criteria disagree on two of the five seeds, which shows that criterion dependence is a property of individual runs rather than an artifact of cross-seed averaging.

Table 8: Per-seed model selection under two criteria on the primary SplitCIFAR-100 audit. Overall-preferred picks the learner with the lowest post-scaling overall ECE. Worst-age-preferred picks the learner with the lowest post-scaling worst exact-age ECE. The two criteria disagree on two of five seeds.

Seed	Overall-preferred	Worst-age-preferred	Disagreement
7	DER++	LwF	Yes
42	DER++	DER++	No
123	LwF	LwF	No
456	LwF	LwF	No
2024	Naive	LwF	Yes

A.6 Selective-Prediction Metrics

Order the n evaluation samples by decreasing maximum predicted probability, and let $e_{(j)} = \mathbf{1}[\hat{y}_{(j)} \neq y_{(j)}]$. At target coverage q, retain the first $k = \lceil qn \rceil$ samples in this common ordering. The empirical selective risk and area under the risk-coverage curve (AURC) are

$$R(q) = \frac{1}{k}\sum_{j=1}^{k} e_{(j)}, \qquad \mathrm{AURC} = \frac{1}{n}\sum_{\ell=1}^{n}\frac{1}{\ell}\sum_{j=1}^{\ell} e_{(j)}. \tag{3}$$

Excess AURC subtracts the AURC of an ideal ordering with all correct predictions before errors. Group risk restricts the same globally retained set to each group. Groups with no retained samples are excluded. Worst-group risk is the maximum of these conditional error rates. Selective-risk disparity here is worst-group risk minus overall risk, unlike ECE disparity, which is the maximum minus minimum group ECE.

A.7 Compact Selective-Prediction Deltas

Table 9 reports paired selective-prediction deltas under the secondary chronological split. Overall risk changes little, whereas worst-group risk under confidence quartiles increases.

A.8 Formalizing the Masking Mechanism

Within a nonempty confidence bin B, let acc_g and conf_g be accuracy and mean confidence on each nonempty intersection $B \cap g$, and let $w_g = |B \cap g|/|B|$. Since both quantities are weighted averages,

$$|\mathrm{acc}(B) - \mathrm{conf}(B)| = \left|\sum_g w_g(\mathrm{acc}_g - \mathrm{conf}_g)\right| \leq \sum_g w_g|\mathrm{acc}_g - \mathrm{conf}_g|. \tag{4}$$

The left side is the error entering aggregate ECE. It can vanish when the signed gaps cancel, even though the subgroup errors on the right remain positive. This elementary identity explains why shared confidence bins do not guarantee agreement between aggregate and group-level calibration.

A.9 Class-Incremental Learning Method Hyperparameters

The primary training specification sets ten epochs per experience, SGD with momentum 0.9 and weight decay 10^{-4}, and cosine learning-rate decay. Table 10 gives the method-specific parameters.

Table 9: Selective-prediction changes under the secondary chronological SplitCIFAR-100 audit (20 runs). Values are paired differences, global TS minus raw. Positive changes indicate greater AURC, risk, or risk disparity.

Metric	Δ mean	Paired 95% CI
ΔAURC	+0.0042	[+0.0007, +0.0080]
ΔRisk@50% (overall)	-0.0005	[-0.0034, +0.0023]
ΔRisk@50% (worst confidence quartile)	+0.0606	[+0.0377, +0.0857]
ΔRisk@70% (overall)	-0.0010	[-0.0018, -0.0002]
ΔRisk@70% (worst confidence quartile)	+0.0179	[+0.0062, +0.0317]
ΔDisparity@70% (confidence quartiles)	+0.0189	[+0.0076, +0.0323]

LwF restricts the current-data cross-entropy term to the current task, while DER++ restricts it to seen classes. The primary seeds are 7, 42, 123, 456, and 2024, each determining a random class permutation partitioned into ten ten-class experiences. These parameter choices define the configured training budget, not a claim about optimal hyperparameters for each method.

For SplitCIFAR-100 and Tiny-ImageNet, ResNet-18 uses a 3×3 stride-1 input convolution without initial max-pooling. ImageNet-100 uses a standard ResNet-18 without external pretraining and a 100-way head. Its training transforms are random resized cropping and horizontal flipping; evaluation uses resizing and center cropping with ImageNet normalization. The stratified primary audit and chronological secondary audits use different calibration populations, as specified in Section 3.2.

Table 10: Primary SplitCIFAR-100 training specification, with ten epochs per experience. Distillation temperature is distinct from the fitted calibration temperature.

Learner	Hyperparameter	Value
Naive	Batch size	32
	Learning rate	0.1
EWC	Batch size	32
	Learning rate	0.1
	λ (regularization strength)	1.0
LwF	Batch size	32
	Learning rate	0.1
	α (distillation weight)	2.0
	Distillation temperature	4.0
DER++	Batch size	64
	Learning rate	0.01
	Memory size	5,000
	α (logit distillation)	0.5
	β (label distillation)	0.5

A.10 Calibration-Stream Training and Transformations

The calibration-stream benchmark uses a CIFAR-style ResNet-18, 100 epochs per experience, batch size 128, and a replay memory of 2,000 examples for ER and DER++. Each experience reserves ten validation examples per class outside learner optimization. DER++ uses coefficients $\alpha = \beta = 0.5$. This is a separate training protocol from the low-accuracy primary benchmark.

The cross-setting specifications in Figure 4 vary the class schedule or backbone. The fine-grained schedule retains ten ten-class experiences, but distributes each of the five classes within a CIFAR-100 superclass across different experiences. After shuffling the twenty superclasses into two halves, five rounds select one fine class from each half per experience, rotating superclass assignments between rounds. The twenty-experience setting instead partitions a shuffled class order into five-class experiences. The ViT-Tiny setting uses ten ten-class experiences and a ViT-Tiny/16 with 32×32

inputs and no pretrained weights. These specifications retain a ten-epoch per-experience budget. As in the primary specification, EWC uses $\lambda = 1$, LwF uses distillation weight 2 and temperature 4, and DER++ uses memory 5,000 and coefficients $\alpha = \beta = 0.5$.

For vector scaling, calibrated probabilities are $\mathrm{softmax}(w \odot z + b)$, where w and b are learned class-specific vectors. Matrix scaling uses $\mathrm{softmax}(Wz + b)$, and the Dirichlet variant applies the affine map to $\log \mathrm{softmax}(z)$. These transformations can change the predicted class. Top-label histogram binning replaces maximum confidence by calibration accuracy in its confidence bin and redistributes the remaining probability mass proportionally among other classes. Unlike scalar temperature scaling, this reconstructed distribution can also change its maximum-probability class.

The T-CIL adaptation operates on a frozen learner. It first fits a target temperature on current validation data, constructs class centroids from replay examples, and chooses the perturbation magnitude by bisection so that the temperature fitted on perturbed current-class replay examples approaches the validation target. It then fits the final scalar temperature on perturbed replay examples. Thus, its information budget includes 2,000 replay examples and 100 current validation examples, but it does not simply minimize NLL on their unmodified union.

Paired external learners. On SplitCIFAR-100, the CoCa and paired DER++ runs share a ResNet backbone, replay memory of 200, batch and replay minibatch sizes of 32, 50 epochs per experience, learning rate 0.03, and $\alpha = 0.1$, $\beta = 0.5$. DGC on/off uses the same batch sizes, epochs, learning rate, and coefficients, with memory 500. Each pair shares class order and seed, and differs in its designated training component.

A.11 Computational Resources

The primary experiments use an NVIDIA GeForce RTX 4070 Ti SUPER with 16 GB VRAM, an AMD Ryzen 9 7950X CPU, Python 3.11, Avalanche 0.6.0 Carta et al. (2023), and PyTorch 2.10.0 (CUDA 12.8). The PermutedMNIST auxiliary experiment uses PyTorch 2.13.0 (CUDA 13.0). The primary audit comprises twenty runs, four learners with five seeds each. Auxiliary comparisons state their own replicate counts; partition resampling is not counted as independent learner training.

A.12 Control Experiments for Cross-Setting Robustness

This subsection outlines the control analyses pertinent to the primary protocol. The budget-matched bucket controls in Table 13, the strict size-matched permutation control in Table 11, and the offline joint-training control in Table 12 use a secondary chronological-split audit rather than the stratified primary audit. These settings comprise a 20% calibration split, a 91-point logarithmic temperature grid over $t \in [0.5, 10.0]$, and the standard coarse age bins $\{0, 1, 2\text{–}3, 4+\}$.

Size-matched permutation control. For each of the 20 benchmark runs, the strict permutation control compares the true age-based worst-group temperature-scaled ECE against 100 matched permutations. These permutations preserve the exact bucket sizes of the true age partition while destroying the underlying sequential age structure. Table 11 shows that, under this secondary protocol, the mean true worst-age ECE is 0.598 ± 0.225, whereas the matched-permutation baseline reaches only 0.102 ± 0.059. None of the 20 runs is exceeded by the permutation baseline. This confirms that the age-conditioned failure is not an artifact of bucket size alone.

Offline joint-training control. The offline joint-training control merges all ten experiences before training. In Table 12, calibrated overall ECE is 0.014 ± 0.002, compared with worst age-group ECE of 0.046 ± 0.007 and worst pseudo-age ECE of 0.041 ± 0.009. The paired age-minus-pseudo-age gap is small (0.005 ± 0.013) and changes sign across seeds. The large chronological age discrepancy is therefore substantially attenuated under joint training.

A.13 Boundary Probe: Age-Aware Scaling vs. Budget-Matched Controls

Table 13 compares coarse-age ACTS with budget-matched partition controls under the secondary chronological split. ACTS applies $\tilde{z}(x) = z(x)/t_{g(x)}$, fitting each t_g by calibration NLL within age bin g. Bins with fewer than two calibration samples retain $t_g = 1$. Coarse-age ACTS attains the

Table 11: Strict size-matched permutation control on the secondary SplitCIFAR-100 chronological-split audit across 20 runs. Across all runs, the mean true worst-age TS ECE is 0.598 ± 0.225, whereas the matched-permutation baseline reaches only 0.102 ± 0.059, yielding a zero permutation exceedance rate.

Method	Seed	t^*	Overall TS ECE	Worst age-group TS ECE	Worst permutation TS ECE	Δ (age − perm)
DER++	7	1.27	0.019	0.327	0.043 ± 0.006	0.284
	42	1.40	0.096	0.362	0.109 ± 0.006	0.253
	123	1.36	0.039	0.286	0.057 ± 0.008	0.228
	456	1.45	0.030	0.320	0.053 ± 0.007	0.267
	2024	1.45	0.032	0.305	0.052 ± 0.007	0.253
EWC	7	10.00	0.066	0.808	0.074 ± 0.005	0.734
	42	10.00	0.067	0.816	0.076 ± 0.005	0.740
	123	10.00	0.068	0.818	0.076 ± 0.005	0.743
	456	10.00	0.070	0.834	0.079 ± 0.005	0.755
	2024	10.00	0.067	0.819	0.075 ± 0.005	0.744
LwF	7	2.64	0.175	0.400	0.188 ± 0.007	0.212
	42	3.02	0.205	0.436	0.217 ± 0.007	0.219
	123	4.07	0.244	0.578	0.257 ± 0.007	0.321
	456	2.64	0.167	0.424	0.181 ± 0.007	0.243
	2024	2.39	0.115	0.370	0.127 ± 0.007	0.242
Naive	7	10.00	0.066	0.798	0.074 ± 0.005	0.724
	42	10.00	0.066	0.804	0.075 ± 0.006	0.730
	123	10.00	0.068	0.821	0.076 ± 0.005	0.745
	456	10.00	0.069	0.826	0.078 ± 0.005	0.749
	2024	10.00	0.065	0.809	0.073 ± 0.005	0.736

Table 12: Offline joint-training control evaluated under the secondary SplitCIFAR-100 chronological-split protocol across five seeds.

Seed	t^*	Overall TS ECE	Worst age-group TS ECE	Worst pseudo-age TS ECE	Gap
7	1.15	0.016	0.038	0.056	-0.017
42	1.11	0.015	0.042	0.032	0.011
123	1.15	0.011	0.053	0.040	0.013
456	1.11	0.017	0.055	0.039	0.016
2024	1.11	0.013	0.042	0.038	0.004
Mean ± std	**1.13 ± 0.02**	**0.014 ± 0.002**	**0.046 ± 0.007**	**0.041 ± 0.009**	**0.005 ± 0.013**

lowest worst-bucket ECE for Naive and EWC, but not for LwF and DER++. This learner-dependent ordering concerns the chronological split, where some age bins lack calibration samples. It does not contradict the gains from exact-age ACTS when every age is represented in the primary stratified split. The table measures calibration, not operational utility. Random controls average twenty partition draws per training seed to reduce partition-sampling noise; these draws are not independent training replicates.

A.14 Calibrator Choice and Learner Rankings under the Chronological Split

Calibrator choice can change which learner an aggregate criterion selects. Table 14 compares global TS, ACTS, and top-label histogram binning under the secondary chronological split. Across five seeds, global TS changes the selected learner relative to raw predictions in 40% of ECE comparisons and 20% of NLL and Brier comparisons. The corresponding rates for ACTS are 20%, 0%, and 0%, while histogram binning leaves the selected learner unchanged for all three scores. Here, top-1 denotes the selected learner, not the predicted class of an individual example.

These ranking comparisons complement the calibration results for global TS in Table 35 and the selective-risk comparisons in Section A.16. They measure different consequences and cannot substitute for a group-calibration comparison of every calibrator. Higher-capacity families such as vector and matrix scaling Guo et al. (2017) and Dirichlet calibration Kull et al. (2019) also differ in regularization requirements. The present results therefore support reporting calibrator-dependent learner rankings, rather than a general ranking of calibration families.

Table 13: Budget-matched partition controls under the secondary chronological split. Each partition uses four temperatures. Random partitions preserve group sizes and use 20 draws per seed over five seeds. Deterministic partitions use one partition per seed. Bold and underline indicate the best and second-best results within each learner, respectively.

Learner	Scheme	Trials	Scaled ECE	Worst-bucket ECE
Naive	random	100	0.045	<u>0.051</u>
	confidence quartiles	5	**0.023**	0.089
	label modulo 4	5	<u>0.038</u>	0.076
	ACTS age bins	5	<u>0.038</u>	**0.041**
EWC	random	100	0.045	<u>0.052</u>
	confidence quartiles	5	**0.027**	0.101
	label modulo 4	5	<u>0.038</u>	0.073
	ACTS age bins	5	<u>0.038</u>	**0.041**
LwF	random	100	0.038	**0.053**
	confidence quartiles	5	**0.028**	0.060
	label modulo 4	5	<u>0.034</u>	<u>0.058</u>
	ACTS age bins	5	0.047	0.077
DER++	random	100	0.046	<u>0.064</u>
	confidence quartiles	5	<u>0.042</u>	0.077
	label modulo 4	5	**0.040**	**0.062**
	ACTS age bins	5	0.049	0.107

Table 14: Top-1 learner change rates under post-hoc calibration over five seeds on SplitCIFAR-100. Lower values indicate greater stability of the induced model ranking. **Bold** and <u>underline</u> indicate best and second-best stability per metric.

Calibrator	ECE	NLL	Brier
Global temperature scaling	0.4	<u>0.2</u>	<u>0.2</u>
ACTS	<u>0.2</u>	**0.0**	**0.0**
Histogram binning	**0.0**	**0.0**	**0.0**

A.15 Compatibility Details: Age-Wise Temperatures and Forgetting Correlations

We use a compatibility suite to compare learner score distributions with bounded calibration fits. Under exact-age ACTS on the primary benchmark, Table 15 shows that fitted temperatures differ across learners. For LwF, the fitted temperatures averaged over five seeds rise with schedule age, from 1.09 for the most recent group to 3.86 for the oldest, a compatibility gradient between the learner's score scale and its age structure. For DER++, the fits averaged over five seeds rise from 0.69 for the most recent group to a plateau near 1.4. For Naive and EWC, the most recent group fits near identity (1.24 and 1.28 on average), while every older group saturates at the upper boundary $t_{\max} = 10$ in all five seeds. Boundary saturation censors the fitted value and does not establish an unbounded optimum. These values describe learner-calibrator compatibility under the declared search protocol rather than a causal mechanism of the learner.

The second result concerns association with catastrophic forgetting under the coarse age grouping of the control analyses. Under global temperature scaling, the Pearson correlation with final average forgetting is 0.713 for age-disparity NLL, 0.517 for worst-age NLL, and 0.645 for worst-age Brier score. For oracle ACTS, the corresponding correlations are approximately zero (-0.007, 0.011, and -0.021, respectively). These finite-benchmark associations do not identify a loss-term mechanism and do not imply that ACTS permanently "solves" age-conditioned calibration. They show that the residual age heterogeneity visible after global temperature scaling is not captured by aggregate calibration metrics alone.

Table 15: Oracle ACTS temperatures under the primary stratified split, with one fit per exact age. Values are mean ± sample standard deviation over five seeds. Values at 10 reach the search boundary.

Age	Naive	EWC	LwF	DER++
0	1.24 ± 0.12	1.28 ± 0.08	1.09 ± 0.02	0.69 ± 0.04
1	10.00 ± 0.00	10.00 ± 0.00	1.35 ± 0.10	1.14 ± 0.18
2	10.00 ± 0.00	10.00 ± 0.00	1.23 ± 0.12	1.17 ± 0.12
3	10.00 ± 0.00	10.00 ± 0.00	1.47 ± 0.14	1.31 ± 0.15
4	10.00 ± 0.00	10.00 ± 0.00	1.49 ± 0.11	1.27 ± 0.05
5	10.00 ± 0.00	10.00 ± 0.00	1.78 ± 0.21	1.38 ± 0.14
6	10.00 ± 0.00	10.00 ± 0.00	2.02 ± 0.32	1.38 ± 0.17
7	10.00 ± 0.00	10.00 ± 0.00	2.25 ± 0.42	1.37 ± 0.14
8	10.00 ± 0.00	10.00 ± 0.00	2.56 ± 0.38	1.31 ± 0.14
9	10.00 ± 0.00	10.00 ± 0.00	3.86 ± 1.29	1.42 ± 0.19

Table 16: Coarse-age controls and mechanism analysis under the secondary chronological split. Temperatures of 1.00 in bins without calibration samples are defaults, not fitted optima. Pearson coefficients measure correlation with final average forgetting (values closer to 0 indicate better decoupling of forgetting from age-structured miscalibration, shown in **bold**).

Quantity	Primary value	Comparison value	Interpretation
ACTS fitted temperatures			
Age 0 ACTS temperature	Naive: 1.00, EWC: 1.00, LwF: 1.00, DER++: 1.00	1.00	default for empty bins, fit otherwise
Age 1 ACTS temperature	Naive: 1.00, EWC: 1.00, LwF: 1.00, DER++: 1.00	1.00	default for empty bins, fit otherwise
Age 2–3 ACTS temperature	Naive: 1.00, EWC: 1.00, LwF: 1.00, DER++: 1.00	1.00	default for empty bins, fit otherwise
Age 4+ ACTS temperature	Naive: 10.00, EWC: 10.00, LwF: 10.00, DER++: 2.09–3.22	varies	fitted on calibration samples
Pearson correlation with final average forgetting			
Corr. with age-disparity NLL	Global TS: 0.713	**ACTS: -0.007**	residual age structure persists after TS
Corr. with worst-age NLL	Global TS: 0.517	**ACTS: 0.011**	forgetting remains coupled to worst-group risk under TS
Corr. with worst-age Brier	Global TS: 0.645	**ACTS: -0.021**	proper-score heterogeneity remains age-structured

The separate chronological-split control uses coarse age bins. Its default temperatures for bins without calibration samples must not be interpreted as fitted optima. Table 16 retains these control results separately.

A.16 Age-Centered Decision Simulation under the Chronological Split

A practical operational question is whether abstention-style decision rules inherently capture what the calibration audit reveals. We evaluate this question with an age-centered decision simulation on the SplitCIFAR-100 learners under the secondary chronological-split protocol. For each individual run, we fit the designated calibrator on the calibration split and evaluate selective prediction across target coverages of 50%, 70%, and 90%. The reported outputs comprise the overall AURC, the excess AURC relative to the ideal risk-coverage curve, the overall selective risk at each coverage, and the worst-age selective risk.

ACTS achieves the lowest mean AURC (0.580) and excess AURC (0.102) in Table 17, followed by raw predictions and global TS. Histogram binning and vector scaling have larger mean AURCs of

Table 17: Age-centered decision simulation on SplitCIFAR-100 across five calibrators (lower is better for all metrics). **Bold** and underline denote best and second-best results per column.

Calibrator	AURC	Excess AURC	Overall Risk@70%	Worst-age Risk@70%	Disparity @70%
ACTS	**0.580**	**0.102**	**0.715**	**0.871**	0.156
Raw	0.621	0.143	0.725	0.877	0.153
Global TS	0.626	0.148	0.724	0.877	0.153
Vector scaling	0.851	0.137	0.896	0.991	0.095
Histogram binning	0.816	0.119	0.854	0.892	**0.037**

Table 18: Detailed age-centered decision simulation summary evaluated on the primary benchmark across 20 runs. Lower values are better for all three reported metrics. **Bold** and underline denote best and second-best entries within each coverage level.

Calibrator	Coverage	Overall risk	Worst-age risk	Disparity
ACTS	50%	**0.642**	0.861	0.219
	70%	**0.715**	**0.871**	0.156
	90%	**0.762**	**0.889**	0.126
Raw	50%	0.669	0.858	0.189
	70%	0.725	0.877	0.153
	90%	0.764	0.892	0.128
Global TS	50%	0.668	**0.855**	0.187
	70%	0.724	0.877	0.153
	90%	0.764	0.892	0.128
Vector scaling	50%	0.875	0.992	0.117
	70%	0.896	0.991	0.095
	90%	0.910	0.990	0.080
Histogram binning	50%	0.831	0.873	**0.041**
	70%	0.854	0.892	**0.037**
	90%	0.873	0.906	**0.033**

0.816 and 0.851. This aggregate ordering does not settle worst-age utility: at 50% coverage, ACTS minimizes overall risk but global TS minimizes worst-age risk, as shown in Table 18.

The second finding is central for the main claim of the paper: rigorous coverage control does not eliminate the age-centered worst-case failure mode. Across all five calibrators and all three specific coverage levels, the mean worst-age selective risk remains high. At 70% coverage, the mean overall selective risk is 0.715 for ACTS, 0.725 for raw predictions, 0.724 for global temperature scaling, 0.896 for vector scaling, and 0.854 for histogram binning. Yet, the corresponding mean worst-age selective risk remains 0.871, 0.877, 0.877, 0.991, and 0.892, respectively. This confirms that the retained decision set still contains a specific age-defined subgroup whose decisions remain much less reliable than the aggregate utility summary suggests.

Finally, this decision simulation exposes a complementary failure mode rather than a direct restatement of the age-ECE audit. ACTS consistently traces the lowest overall selective-risk curve, but the worst-age selective-risk curve remains elevated across 50%, 70%, and 90% coverage. This result supports our recommendation that class-incremental evaluation methodologies should report subgroup-aware utility diagnostics alongside standard aggregate summaries.

Table 18 summarizes this secondary simulation. At 70% coverage, ACTS leads both overall and worst-age risk, so that operating point demonstrates residual group risk rather than disagreement in the preferred calibrator.

A.17 Signed Direction, Operational Identity, and Class-Level Tails

Unsigned ECE does not distinguish overconfidence from underconfidence, and the age group with the largest ECE need not be the age group with the largest selective risk. Table 19 makes this distinction explicit for global temperature scaling at 70% coverage. The worst-ECE group is underconfident in

Table 19: Direction and operational identity after global TS under the primary stratified split and exact-age grouping. Counts span five seeds. Identity agreement compares the worst-ECE group with the highest-risk group at 70% coverage. The risk gap is the mean selective-risk difference between these groups. These are diagnostics, not performance rankings.

Learner	Worst-ECE direction	Identity agreement	Mean risk gap
Naive	under 5/5, over 0/5	0/5	0.946
EWC	under 5/5, over 0/5	0/5	0.945
LwF	under 1/5, over 4/5	3/5	0.155
DER++	under 5/5, over 0/5	0/5	0.666

Table 20: Per-class tails after global TS under the primary stratified audit. Each entry reports the 95th percentile ($q_{0.95}$) over 100 classes and five seeds, with a 95% hierarchical bootstrap interval resampling seeds and then classes within each seed. Each class has 80 evaluation examples.

Learner	NLL $q_{0.95}$ [95% CI]	Brier $q_{0.95}$ [95% CI]	$\lvert\text{gap}\rvert$ $q_{0.95}$ [95% CI]
Naive	4.792 [4.782, 4.799]	0.997 [0.996, 0.997]	0.822 [0.772, 0.851]
EWC	4.814 [4.800, 4.831]	0.997 [0.997, 0.997]	0.835 [0.773, 0.850]
LwF	4.745 [4.494, 5.128]	1.096 [1.076, 1.121]	0.311 [0.285, 0.339]
DER++	4.386 [4.208, 4.529]	1.082 [1.053, 1.099]	0.313 [0.290, 0.339]

all seeds for Naive, EWC, and DER++, but LwF exhibits the opposite direction in four of five seeds. The calibration and selective-risk identities never agree for Naive, EWC, or DER++ at this coverage. LwF exhibits agreement in three of five seeds, which shows that the relation is learner dependent rather than universal.

The group summaries can still hide concentrated class-level residuals. Table 20 therefore reports the 95th percentile ($q_{0.95}$) of class tails under global temperature scaling for NLL, Brier score, and absolute signed gap. The intervals use a hierarchical bootstrap that resamples seeds and then classes within each sampled seed. This procedure accounts for variability across seeds and classes. Each class contains 80 test examples, so the class-level estimates remain finite-sample descriptions rather than precise population quantities. These intervals condition on the fitted calibrator and do not quantify uncertainty from calibration-set resampling.

Figure 5 visually complements the tabular data with a trend analysis across variable coverages. The left panel clarifies the broad utility trade-off: ACTS consistently traces the lowest overall selective-risk curve, followed closely by raw predictions and global temperature scaling, while vector scaling and histogram binning perform noticeably worse. The right panel isolates why this broad utility improvement does not invalidate the substantive conclusion. Worst-age selective risk remains elevated across 50%, 70%, and 90% target coverage for every calibrator, even though aggregate utility differs considerably.

A.18 Domain-Incremental Auxiliary Check

As an auxiliary boundary check expanding beyond the core class-incremental setting, we executed a lightweight domain-incremental experiment on PermutedMNIST Goodfellow et al. (2014), a standard benchmark constructed by applying distinct pixel permutations across tasks. We train Naive and EWC with seeds 17, 42, and 123 on a fixed five-domain permutation schedule (benchmark seed 42). Each domain receives one epoch of training using an MLP with hidden width 512, SGD with learning rate 0.01 and momentum 0.9, and batch size 64. EWC uses $\lambda = 0.4$. Within each domain, the first 20% of test samples fit calibration and the remaining 80% support evaluation. This retrospective split differs from the primary class-stratified audit. The four calibration paths are raw predictions, global temperature scaling, vector scaling, and a domain-conditional temperature model conditioning on task/domain index. We emphasize that this particular experiment serves only a narrowly bounded purpose within the appendix. Its role is to test empirically whether the aggregate-versus-group mismatch can manifest in a related non-class-incremental continual regime.

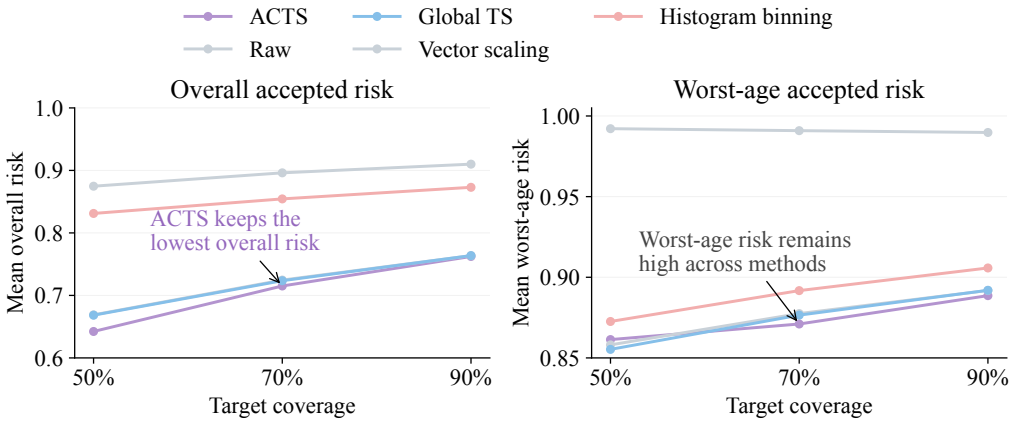


Figure 5: Selective risk under the secondary chronological split with coarse age bins, averaged over four learners and five seeds each. Left: overall risk. Right: worst-age risk within the globally retained set.

Table 21: PermutedMNIST auxiliary comparison. Values are mean ± sample standard deviation over three training seeds with a fixed five-domain permutation schedule. Bold and underline indicate the best and second-best calibration results within each learner, respectively.

Learner	Calibrator	Overall ECE	Worst-domain ECE	ECE disparity	Worst-domain NLL
Naive	Raw	0.040 ± 0.011	0.198 ± 0.032	0.191 ± 0.030	1.436 ± 0.067
	Global TS	**0.034 ± 0.022**	0.187 ± 0.055	0.175 ± 0.062	1.424 ± 0.194
	Vector scaling	0.079 ± 0.020	0.304 ± 0.038	0.278 ± 0.035	1.991 ± 0.265
	Domain-cond. TS	0.035 ± 0.023	**0.144 ± 0.070**	**0.139 ± 0.070**	**1.324 ± 0.235**
EWC	Raw	**0.043 ± 0.023**	0.212 ± 0.073	0.204 ± 0.071	1.538 ± 0.465
	Global TS	0.043 ± 0.026	0.212 ± 0.078	0.204 ± 0.078	1.549 ± 0.464
	Vector scaling	0.083 ± 0.023	0.282 ± 0.058	0.257 ± 0.052	1.848 ± 0.328
	Domain-cond. TS	0.050 ± 0.027	**0.174 ± 0.079**	**0.169 ± 0.079**	**1.394 ± 0.347**

Table 21 shows that global TS modestly improves worst-domain ECE, disparity, and worst-domain NLL for Naive. For EWC, ECE and disparity are unchanged at the displayed precision, while worst-domain NLL increases slightly. Vector scaling increases ECE and NLL. Domain-conditional scaling reduces worst-domain ECE and NLL for both learners, revealing heterogeneity even with a fixed label space. This auxiliary result is limited to two learners, three training seeds, and one permutation schedule.

Figure 6 details the corresponding per-domain temporal trend using calibrated NLL. The curves reveal heterogeneous proper-score errors across domains. Domain-conditional scaling reduces the worst-domain NLL on average for both learners, whereas vector scaling increases it. Domain index alone should not be interpreted as an intrinsic measure of calibration difficulty. This behavior aligns with the supplemental hypothesis that the measurement mismatch may extend beyond the strict class-incremental setting, although this auxiliary experiment does not provide exhaustive evidence for that broader claim.

A.19 Extension-Only Baseline Risk Audit

Table 22 reports supplemental runs for X-DER Boschini et al. (2023), L2P Wang et al. (2022), CL-LoRA He et al. (2025), and IDER Liu et al. (2026), each evaluated over three independent seeds per dataset under the secondary chronological split with coarse age bins, distinct from the exact-age primary audit. We report overall accuracy, lowest age-bin accuracy, the raw and post-scaling values of overall ECE, the raw and post-scaling values of age worst-group ECE, and the post-scaling age disparity. This extends the main claim to recent CIL baselines, rather than limiting the evidence to the four primary learners.

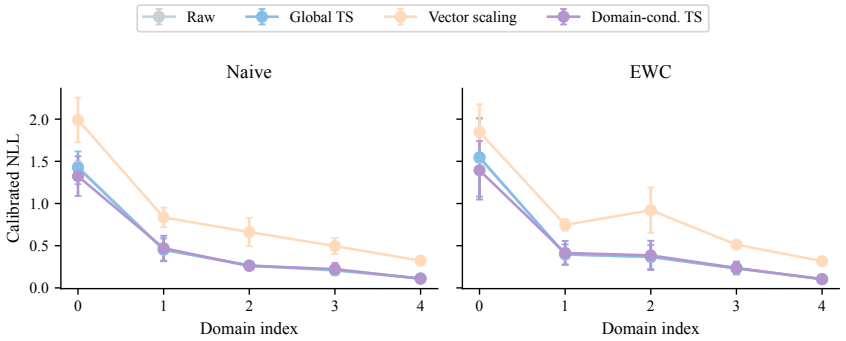


Figure 6: PermutedMNIST domain-incremental auxiliary check. Each panel reports calibrated NLL by domain index. Points and error bars show the mean and sample standard deviation over three training seeds with a fixed permutation schedule. We use this result only as auxiliary appendix evidence that aggregate-vs-group mismatch may extend beyond the main class-incremental setting.

Table 22: Supplemental continual-learning baselines under the secondary chronological split and coarse age bins. Values are mean ± standard deviation over three seeds. Bold and underline indicate the best and second-best results within each dataset, respectively. Disparity reports post-scaling age-group ECE differences.

Dataset	Method	Overall Acc. (%)	Lowest Acc. (%)	Overall ECE (%)		Worst-Group ECE (%)		Disparity TS (%)
				Raw	TS	Raw	TS	
SplitCIFAR-100	X-DER	43.8±1.6	34.6±1.5	23.1±3.8	**4.2±0.3**	28.0±4.5	21.3±4.3	17.1±4.0
	L2P	31.6±0.8	25.0±3.6	13.1±1.6	5.4±0.7	**16.7±2.3**	**19.2±2.1**	**13.8±1.6**
	CL-LoRA	28.9±1.0	21.4±1.0	17.4±1.3	14.9±1.6	23.8±1.7	46.4±1.7	31.5±1.6
	IDER	**46.8±0.8**	**38.4±0.5**	**8.0±0.6**	10.7±2.3	26.9±0.2	30.2±2.1	19.5±0.3
Tiny-ImageNet	X-DER	24.0±1.1	18.7±2.5	11.5±1.6	8.9±2.2	**14.2±2.1**	**11.9±2.6**	**2.9±0.3**
	L2P	**31.9±0.6**	**27.8±0.4**	18.5±0.7	5.8±1.2	21.4±0.3	17.9±1.8	12.1±0.7
	CL-LoRA	30.0±0.9	20.7±1.3	21.5±0.7	20.2±1.6	28.9±0.5	49.3±1.5	29.0±0.7
	IDER	24.0±1.8	6.2±5.8	**7.3±2.7**	**3.9±0.6**	16.4±5.8	13.7±3.8	9.9±4.4
ImageNet-100	X-DER	45.6±0.5	41.8±1.7	**4.4±0.5**	4.5±1.7	**10.0±2.1**	13.6±7.1	9.1±5.8
	L2P	**80.1±0.3**	**77.1±0.6**	11.4±0.1	**2.8±0.6**	13.5±0.5	**6.2±0.5**	**3.4±0.3**
	CL-LoRA	78.0±0.0	71.6±1.3	6.8±0.5	11.8±0.8	11.1±0.1	17.6±0.8	5.8±0.8
	IDER	36.4±1.1	14.6±8.2	12.6±5.2	15.3±2.2	34.6±7.5	36.4±6.5	21.1±4.4

The discrepancy persists across these method–dataset pairs, with varying magnitudes in Table 22. On SplitCIFAR-100, L2P reaches 31.6% accuracy and 5.4% post-scaling overall ECE, yet worst-age ECE remains 19.2%. On Tiny-ImageNet, L2P and CL-LoRA reach 31.9% and 30.0% accuracy, respectively, while retaining larger group than aggregate errors. The same distinction holds at the higher ImageNet-100 accuracies of 80.1% and 78.0%. X-DER exhibits the discrepancy on all three datasets, although its weak Tiny-ImageNet accuracy limits the practical interpretation there.

IDER provides a training-time reliability comparison. In Table 22, its age worst-group ECE remains larger than its post-scaling overall ECE on all three datasets. This supports examining residual group calibration even when a learner is designed to improve reliability during training.

These supplemental configurations are not matched against the primary benchmark. L2P and CL-LoRA use a shared pretrained ViT backbone, whereas the other methods retain their own training settings. The table supports recurrence of the aggregate-versus-group discrepancy, not a ranking across unmatched methods.

B High-Accuracy and External-Family Audit Details

Table 23 summarizes the external-family audits on SplitCIFAR-100 and Tiny-ImageNet. All fits use the 201-point scale-aware grid $[0.005, 100]$. Every global optimum is interior, so no global endpoint extension is triggered; conditional fits can still reach a boundary. The ACMap rows correspond to Table 2. CoCa, paired DER++, and DGC on/off retain their respective training protocols and are not pooled with the primary benchmark.

Table 23: Scale-aware audit of the external learner families on SplitCIFAR-100 and Tiny-ImageNet (three seeds each). All fits use the scale-aware log-grid $[0.005, 100]$. No global fit saturates a grid boundary. Mean $\pm$ sample standard deviation over three seeds. ECE values are percentages.

Dataset	Learner	Acc. (%)	Overall ECE (%, global TS)	Worst exact-age ECE (%, global TS)
SplitCIFAR-100	ACMap	89.5$\pm$0.3	1.94$\pm$0.21	6.55$\pm$1.55
	CoCa	22.8$\pm$2.8	9.57$\pm$0.77	26.80$\pm$2.16
	DER++ (paired)	25.1$\pm$2.1	5.50$\pm$1.48	29.81$\pm$5.89
	DGC off	36.5$\pm$1.2	7.64$\pm$0.25	18.83$\pm$2.15
	DGC on	36.4$\pm$0.4	7.40$\pm$1.31	19.08$\pm$2.22
Tiny-ImageNet	ACMap	85.7$\pm$0.4	1.25$\pm$0.21	4.75$\pm$0.72
	CoCa	12.2$\pm$0.3	5.57$\pm$0.81	36.25$\pm$2.31
	DER++ (paired)	11.8$\pm$0.8	4.05$\pm$1.27	40.62$\pm$2.84
	DGC off	17.9$\pm$0.9	4.39$\pm$0.47	28.20$\pm$1.78
	DGC on	18.0$\pm$0.3	3.26$\pm$0.76	29.39$\pm$1.37

Table 24: Paired deltas for the matched training-mechanism comparisons (three seeds each). Each row reports the change of the intervention relative to its paired reference from the same repository and protocol. Positive ΔECE indicates higher calibration error under the intervention. Mean $\pm$ sample standard deviation over three seeds. Values are percentages.

Dataset	Comparison	ΔAcc. (%)	ΔOverall ECE (%, global TS)	ΔWorst exact-age ECE (%, global TS)
SplitCIFAR-100	CoCa $-$ paired DER++	-2.31$\pm$4.95	+4.07$\pm$1.89	-3.01$\pm$5.70
	DGC on $-$ DGC off	-0.08$\pm$1.51	-0.24$\pm$1.25	+0.25$\pm$2.81
Tiny-ImageNet	CoCa $-$ paired DER++	+0.34$\pm$0.69	+1.52$\pm$1.81	-4.37$\pm$4.42
	DGC on $-$ DGC off	+0.15$\pm$0.67	-1.13$\pm$0.68	+1.19$\pm$1.87

Table 24 reports the paired deltas for the two matched training-mechanism comparisons. The per-seed directions are mixed across metrics, which is why the main text interprets these comparisons as evidence that the audit can resolve opposing trade-offs rather than as a ranking of training mechanisms.

C Calibration Source Details

The held-out pool contains ten validation examples per class from each of ten experiences, excluded from learner optimization, for 1,000 examples. A global temperature is fitted on the 91-point grid $[0.5, 10]$ and evaluated on the same 10,000 test samples as the other calibration-stream rows. Temperature ranges and boundary counts accompany Table 3. This adapts the validation-replay approach of Li et al. (2024): their rule retains all first-experience validation examples and a random 40% of later ones, whereas this pool is class-balanced and fitted at the final stage.

Table 25 evaluates T-CIL on three independently trained ER learners with seeds 1, 2, and 3. The training protocol uses ResNet-32, ten contiguous ten-class tasks, batch size 128, 200 epochs per task, and a class-balanced replay memory of up to 2,000 samples. SGD uses momentum 0.9, weight decay 0.0002, and learning rate 0.1, reduced tenfold at epochs 100 and 150. Each task reserves 100 validation samples outside training, and temperature fitting uses 100 epochs. Adaptive ECE uses ten approximately equal-size confidence bins, whereas standard ECE uses ten equal-width bins. We average ten stages within each seed before computing the mean and sample standard deviation across seeds. Stage-level metrics have 0.01-percentage-point precision. This cross-stage reference is separate from the final-stage calibration-stream comparison.

D Age and Replay Exposure: Association Analysis

Table 26 analyzes the calibration-stream learners across all ten learning stages. Each learner contributes $5 \times (10 + 20 + \cdots + 100) = 2{,}750$ stage–class–seed records: every seen class is measured after each stage, for each of five seeds. The two-learner panel therefore has 5,500 records, not just final-stage class summaries. Controlling for class identity, seed, and replay exposure leaves a larger incremental age association with NLL for ER than for DER++, and weak associations with Brier score and signed gap. The design matrix has condition number 134 and full rank. Repeated class measurements support an associational description, not a causal age effect.

Table 25: T-CIL on three independently trained ER learners under the protocol in Section C. Entries average ten stages within each seed. The final row reports mean and sample standard deviation across seeds. Stage-level ECE has precision 0.01 percentage points.

Training seed	ECE (%)	Adaptive ECE (%)
1	6.94	6.93
2	6.96	6.79
3	7.75	7.74
Mean ± SD	7.22 ± 0.46	7.15 ± 0.51

Table 26: Exposure-adjusted association analysis on the class-level panel of the calibration-stream learners (2,750 class-seed records per learner). Partial R^2 values quantify the incremental variance explained by replay exposure and by schedule age after controlling for class identity, seed, and the other temporal variable. These are associational statistics and do not identify causal effects.

Learner	Metric	Partial R^2, exposure	Partial R^2, age
ER	NLL	0.562	0.175
	Brier	0.764	0.006
	Signed gap	0.707	0.016
DER++	NLL	0.475	0.073
	Brier	0.528	0.019
	Signed gap	0.441	0.019

E Raw Calibration Across Sequence Lengths

Table 27 reports raw calibration for $T \in \{1, 2, 5, 10, 20, 50, 100\}$ experiences on SplitCIFAR-100. Mean raw ECE increases monotonically for Naive up to $T = 100$, where accuracy falls to about 1%. Its NLL is not monotonic. DER++ shows non-monotonic ECE with variation across seeds. The endpoint is a stress case rather than a smooth scaling law. These runs receive no post-hoc scaling.

Table 27: Raw calibration across sequence lengths T on SplitCIFAR-100 (three seeds per cell, 42 runs in total). All values are computed on uncalibrated outputs. No temperature scaling is applied. The T=100 row for Naive is a near-collapse stress endpoint and is retained in full. Mean ± sample standard deviation over seeds. ECE values are percentages.

Learner	T	Acc. (%)	Overall ECE (%)	NLL	Brier
Naive	1	74.7±0.2	11.84±0.19	1.133 ± 0.005	0.376 ± 0.000
	2	41.3±0.5	43.21±0.86	7.703 ± 0.286	0.975 ± 0.012
	5	17.2±0.5	69.14±0.60	16.103 ± 0.592	1.467 ± 0.011
	10	9.0±0.2	78.51±0.04	17.894 ± 0.259	1.644 ± 0.001
	20	4.5±0.2	86.98±2.54	20.245 ± 3.445	1.788 ± 0.035
	50	1.9±0.1	92.64±1.56	22.439 ± 4.242	1.883 ± 0.023
	100	1.0±0.0	99.00±0.00	14.577 ± 0.023	1.980 ± 0.000
DER++	1	74.7±0.0	12.11±0.13	1.124 ± 0.003	0.376 ± 0.001
	2	42.0±12.7	9.15±11.30	2.558 ± 0.520	0.738 ± 0.091
	5	37.3±4.7	11.11±6.34	2.854 ± 0.264	0.784 ± 0.051
	10	30.6±4.1	14.43±2.50	3.272 ± 0.198	0.861 ± 0.034
	20	17.5±4.1	35.41±20.90	4.497 ± 0.867	1.115 ± 0.197
	50	9.8±0.9	25.61±5.01	4.647 ± 0.130	1.055 ± 0.038
	100	5.6±0.9	15.21±6.00	4.841 ± 0.114	1.024 ± 0.026

F Per-Seed Results

Table 28 details the five-seed primary stratified SplitCIFAR-100 audit with exact ages. Naive and EWC reach the upper temperature boundary in every run. LwF has interior optima from 1.71 to 1.83, and DER++ from 1.15 to 1.31. Worst exact-age ECE remains larger than post-scaling overall ECE in every run.

Table 28: Per-seed results of the primary SplitCIFAR-100 audit under exact-age grouping. ECE values are percentages. t^* is the fitted global temperature. † marks saturation at the upper grid boundary $t_{\max} = 10$.

			Overall ECE (%)		Worst exact-age ECE (%)		
Learner	**Seed**	**Acc. (%)**	**Raw**	**Global TS**	**Raw**	**Global TS**	t^*
Naive	7	8.5	66.81	4.55	76.22	79.30	$10.00^\dagger$
	42	8.6	68.15	4.54	77.40	80.03	$10.00^\dagger$
	123	8.8	67.25	4.52	78.43	81.14	$10.00^\dagger$
	456	8.6	67.32	4.45	77.21	80.75	$10.00^\dagger$
	2024	8.5	68.54	4.28	78.84	79.41	$10.00^\dagger$
EWC	7	8.6	66.81	4.52	76.39	80.17	$10.00^\dagger$
	42	8.6	67.26	4.52	76.71	79.78	$10.00^\dagger$
	123	8.9	67.73	4.46	78.05	81.52	$10.00^\dagger$
	456	8.8	68.35	4.63	78.11	82.11	$10.00^\dagger$
	2024	8.6	67.58	4.50	79.24	80.51	$10.00^\dagger$
LwF	7	26.0	26.85	2.79	48.90	24.57	1.77
	42	25.3	26.94	3.22	45.61	23.83	1.77
	123	28.6	29.81	3.30	52.02	24.93	1.83
	456	24.7	25.75	3.68	45.55	23.96	1.71
	2024	25.5	28.57	4.91	48.22	24.87	1.71
DER++	7	33.8	8.74	2.73	18.15	25.80	1.15
	42	35.7	8.44	2.06	18.82	23.31	1.15
	123	30.4	16.06	4.40	24.33	28.49	1.27
	456	30.3	18.82	5.46	29.53	27.64	1.31
	2024	28.5	20.62	7.07	32.59	25.23	1.31

F.1 Seed-wise Paired Deltas

To verify that the overall-versus-group mismatch is not an artifact of cross-seed aggregation, Table 29 isolates the per-seed difference (Δ) between the worst-group ECE and the overall ECE following scaling. These per-run deltas make it possible to check whether the aggregate pattern persists at the level of individual runs rather than only after averaging.

F.2 Exact-Age Direction and EAQ-Conditional Scaling

Tables 30 and 31 report signed gaps and ECE separately for all four learners. The worst group must be identified from ECE, not from the absolute mean signed gap, since opposing bin-level errors can cancel in the latter. These are means within each age group; the mean of per-seed maxima in the main table need not equal the maximum of these means.

G Full Grouping Comparison

Table 33 separates learners under the secondary chronological split with coarse age bins. Age zero, the most recent cohort, has the largest post-scaling age-group ECE for every learner. Confidence grouping instead identifies the highest-confidence quartile, while gap grouping identifies the lowest quartile. Gap-based worst ECE exceeds age-based worst ECE for LwF and DER++, even though age has the larger four-learner average. Table 6 additionally includes EAQ groups. These secondary identities should not be substituted for the primary exact-age results in Section 4.1.

G.1 Grouping Overlap Quantification

To support the principal claim in Section 4.5 that distinct grouping axes isolate largely non-overlapping failure sets, we compute pairwise Jaccard indices between the worst-group sample sets identified by each axis. For each run, we identify the subgroup with the highest post-scaling ECE under each stratification scheme and compute $J(A, B) = |A \cap B|/|A \cup B|$.

Table 29: Per-seed post-scaling ECE under the primary stratified split and exact-age grouping. The gap is worst exact-age ECE minus overall ECE, in percentage points.

Learner	Seed	Overall (%)	Worst (%)	Gap (pp)
Naive	7	4.55	79.30	74.76
	42	4.54	80.03	75.50
	123	4.52	81.14	76.62
	456	4.45	80.75	76.30
	2024	4.28	79.41	75.13
EWC	7	4.52	80.17	75.65
	42	4.52	79.78	75.26
	123	4.46	81.52	77.06
	456	4.63	82.11	77.48
	2024	4.50	80.51	76.01
LwF	7	2.79	24.57	21.78
	42	3.22	23.83	20.61
	123	3.30	24.93	21.63
	456	3.68	23.96	20.28
	2024	4.91	24.87	19.96
DER++	7	2.73	25.80	23.07
	42	2.06	23.31	21.25
	123	4.40	28.49	24.10
	456	5.46	27.64	22.18
	2024	7.07	25.23	18.16

Table 30: Signed confidence gaps by exact age under the primary stratified audit. Five-seed means in percent. Positive values denote overconfidence and negative values underconfidence. Raw and global-TS outputs use the same groups.

Age	Naive raw	Naive TS	EWC raw	EWC TS	LwF raw	LwF TS	DER++ raw	DER++ TS
0	4.39	-80.13	3.90	-80.82	2.73	-22.45	-14.63	-26.06
1	75.21	4.00	74.97	3.97	11.64	-13.86	14.95	4.21
2	72.75	3.75	72.98	3.79	9.65	-16.60	12.16	1.60
3	74.69	4.00	74.54	4.06	20.73	-4.83	16.22	5.81
4	75.79	3.99	75.91	4.06	25.10	0.25	18.60	7.99
5	76.80	4.13	76.87	4.11	35.50	10.90	20.89	10.51
6	73.86	3.91	74.07	3.94	38.44	14.61	19.44	9.03
7	74.20	3.92	74.28	3.94	42.54	18.38	20.41	9.88
8	74.32	3.91	74.39	3.90	43.06	18.97	16.04	5.66
9	74.16	3.87	73.55	3.81	46.44	22.42	21.16	10.80

Table 34 reports modest overlap under the secondary chronological audit. The largest mean Jaccard index is 0.306 for age versus gap, followed by 0.209 for age versus confidence and 0.184 for confidence versus gap. These sample sets are therefore not interchangeable. Their partial overlap supports multi-axis reporting without establishing a common causal source of error.

G.2 Proper-Score Summary Across Groupings

Table 35 summarizes the secondary chronological audit. Worst-age NLL and Brier exceed their aggregate values for each learner, but the gaps are small for Naive and EWC despite their large ECE discrepancies. Proper scores and ECE therefore quantify different aspects of residual error. DER++ has the lowest overall ECE and worst-age proper scores among these learners, while its gap-defined groups expose larger errors than its age groups.

Table 31: ECE by exact age under the primary stratified audit. Five-seed means in percent, using the same groups as Table 30. These are groupwise means, not the mean of per-seed maxima.

Age	Naive raw	Naive TS	EWC raw	EWC TS	LwF raw	LwF TS	DER++ raw	DER++ TS
0	5.04	80.13	4.39	80.82	4.51	22.62	14.63	26.09
1	75.21	4.00	74.97	3.97	11.97	14.37	15.50	7.24
2	72.75	3.75	72.98	3.79	10.01	17.52	13.23	6.11
3	74.69	4.00	74.54	4.06	21.05	10.04	17.02	7.97
4	75.79	3.99	75.91	4.06	25.50	9.07	18.60	8.82
5	76.80	4.13	76.87	4.11	35.55	13.55	20.90	11.40
6	73.86	3.91	74.07	3.94	38.52	15.66	19.48	9.98
7	74.20	3.92	74.28	3.94	42.56	18.58	20.46	10.58
8	74.32	3.91	74.39	3.90	43.07	18.99	16.10	6.38
9	74.16	3.87	73.55	3.81	46.44	22.44	21.27	11.21

Table 32: EAQ-conditional scaling on the primary stratified audit. Values are five-seed mean ECE in percent. Bold indicates the better-calibrated fit within each learner. True-age groups are used only for evaluation of the EAQ fit.

Learner	Fit	Overall	Worst exact age
Naive	Global TS	4.47	80.13
	EAQ-conditional TS	**2.90**	**73.74**
EWC	Global TS	4.53	80.82
	EAQ-conditional TS	**2.85**	**74.68**
LwF	Global TS	3.58	**24.43**
	EAQ-conditional TS	**2.82**	27.58
DER++	Global TS	**4.34**	**26.09**
	EAQ-conditional TS	4.56	32.73

H Sensitivity to Binning, Calibration Fraction, and Age Bins

To assess calibration design choices under the secondary chronological-split protocol, we vary the ECE bin count, the calibration split fraction, and the age-bin definition while holding the rest of the protocol fixed. The resulting analysis shows that the aggregate post-scaling ECE is comparatively stable, whereas the worst-group statistic is more sensitive to how age structure is resolved. This pattern is consistent with the main argument of the paper: subgroup-sensitive conclusions are not an artifact of a single arbitrary preprocessing choice, but they do depend on whether the audit preserves the temporal resolution needed to expose the failure mode.

Across the twenty secondary chronological runs, changing ECE resolution from 10 to 20 bins has little effect in Table 36. Varying the calibration fraction from 10% to 30% also retains a large aggregate-versus-group gap in Table 37. Age resolution has the strongest effect in Table 38: merging age zero with older groups reduces the worst-group statistic. The default and finer partitions both isolate age zero, explaining their similar maxima. The finer partition still merges ages $4+$ and is not the ten-group primary exact-age audit. These checks establish persistence under the tested secondary variations, not sensitivity guarantees for the primary protocol.

I Protocol Specification

The core audit protocol was fixed in advance of the final analysis. The specification fixes the logarithmic temperature grid ($t \in [0.5, 10.0]$ with 91 points) and the scale-aware grid ($t \in [0.005, 100]$ with 201 points plus the boundary-triggered extension rule), the subgroup definitions of the core audit (exact schedule age, confidence quartiles, gap quartiles, and the EAQ proxy), the evaluation hierarchy in which ECE remains primary while NLL and Brier serve as secondary corroborating metrics, and the requirement to report boundary saturation whenever a fitted optimum touches a grid endpoint.

Table 33: Grouping comparison: per-learner breakdown on SplitCIFAR-100 chronological calibration split (5 seeds per learner).

Learner	Grouping	TS overall	TS worst	TS disparity	Worst bin
Naive	age	0.067	0.812	0.745	age 0
	confidence quartiles	0.067	0.232	0.165	Q4
	gap quartiles	0.067	0.392	0.326	Q1
EWC	age	0.068	0.819	0.751	age 0
	confidence quartiles	0.068	0.235	0.167	Q4
	gap quartiles	0.068	0.396	0.328	Q1
LwF	age	0.181	0.442	0.260	age 0
	confidence quartiles	0.181	0.297	0.115	Q4
	gap quartiles	0.181	0.864	0.683	Q1
DER++	age	0.043	0.320	0.277	age 0
	confidence quartiles	0.043	0.073	0.029	Q4
	gap quartiles	0.043	0.661	0.618	Q1

Table 34: Pairwise Jaccard indices of worst-ECE subsets after global TS under the secondary chronological SplitCIFAR-100 audit, using coarse age bins and fixed raw-output quartiles (mean across twenty runs).

Axis pair	Jaccard index
Age vs. Confidence	0.209
Age vs. Gap	0.306
Confidence vs. Gap	0.184

J Licenses for Existing Assets

The following external assets are used in this work. We respect the terms and licenses of each.

- **CIFAR-100** Krizhevsky (2009): Distributed from the University of Toronto dataset page (https://www.cs.toronto.edu/~kriz/cifar.html).
- **Tiny-ImageNet**: Derived from ImageNet and released for academic research use only (https://cs231n.stanford.edu/).
- **ImageNet-100**: The fixed subset used in this study is derived from ILSVRC-2012 and remains subject to the ImageNet Terms of Access (https://image-net.org/download.php).
- **PermutedMNIST** Goodfellow et al. (2014): Constructed locally from MNIST by applying fixed task-specific pixel permutations.
- **Avalanche** Carta et al. (2023): MIT License.
- **PyTorch**: BSD-style License (https://github.com/pytorch/pytorch/blob/main/LICENSE).
- **timm** (ViT backbones): Apache 2.0 License.

K Broader Impacts

Our contribution is a measurement protocol rather than a deployed system. It can help practitioners detect subgroup-specific calibration failures before using confidence for abstention or human referral. Its main societal risk is overinterpretation: neither a small aggregate calibration error nor a successful subgroup audit certifies deployment safety. Evaluation groups may also omit important populations outside the academic benchmarks studied here. The protocol should therefore complement application-specific assessments rather than replace them.

Table 35: Global-TS proper scores under the secondary chronological SplitCIFAR-100 audit with coarse age bins (five-seed means). **Bold** marks the axis exposing the largest worst-group score within each learner.

Learner	Grouping	Overall ECE	Worst ECE	Overall NLL	Worst NLL	Overall Brier	Worst Brier
Naive	age	0.0668	**0.8118**	4.5153	**4.7554**	0.9826	0.9952
	confidence quartiles	0.0668	0.2319	4.5153	4.6737	0.9826	0.9919
	gap quartiles	0.0668	0.3924	4.5153	4.7474	0.9826	**0.9960**
EWC	age	0.0676	**0.8191**	4.5027	**4.7479**	0.9824	0.9951
	confidence quartiles	0.0676	0.2347	4.5027	4.6686	0.9824	0.9920
	gap quartiles	0.0676	0.3960	4.5027	4.7340	0.9824	**0.9959**
LwF	age	0.1813	0.4416	3.0062	3.3711	0.8807	0.9413
	confidence quartiles	0.1813	0.2966	3.0062	**3.5661**	0.8807	0.9525
	gap quartiles	0.1813	**0.8645**	3.0062	3.4514	0.8807	**0.9718**
DER++	age	0.0433	0.3200	2.7691	3.1560	0.8012	0.8733
	confidence quartiles	0.0433	0.0726	2.7691	3.6426	0.8012	0.9501
	gap quartiles	0.0433	**0.6615**	2.7691	**3.6874**	0.8012	**1.0971**

Table 36: ECE-bin sensitivity under the secondary chronological SplitCIFAR-100 audit with coarse age groups (mean±std over twenty runs).

Bins	Overall TS ECE	Worst-Group TS ECE (age)
10	0.0896±0.0611	0.5981±0.2308
15	0.0898±0.0610	0.5981±0.2307
20	0.0899±0.0609	0.5982±0.2306

K.1 AI USE STATEMENT

In this work, we used generative AI tools to implement methods, assist with translation, create or modify scientific figures, draft parts of the paper, and edit the paper to improve readability. All AI-assisted outputs were reviewed and verified by the authors: generated text was checked against the experimental results and revised where it did not match them, and generated code was tested for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

L REPRODUCIBILITY STATEMENT

All evaluated datasets are public academic benchmarks. Section 3 specifies the grouping definitions, temperature grids, and calibration data sources. This appendix provides primary per-seed results, additional protocol details, and auxiliary comparisons. An anonymized implementation is available at https://anonymous.4open.science/status/Auditing-Group-Aware-Risk-in-Class-Incremental-Learning.

Table 37: Calibration-fraction sensitivity under the secondary chronological SplitCIFAR-100 audit with coarse age groups (mean±std over twenty runs).

Fraction	Overall TS ECE	Worst-Group TS ECE (age)
10%	0.0897±0.0666	0.6201±0.2131
20%	0.0898±0.0610	0.5981±0.2307
30%	0.0995±0.0581	0.5894±0.2361

Table 38: Age-bin sensitivity under the secondary chronological SplitCIFAR-100 audit (mean±std over twenty runs). The finer partition is not the primary exact-age partition.

Definition	Bins	Overall TS ECE	Worst-Group TS ECE
Coarse	$\{0\text{–}1, 2+\}$	0.0898±0.0610	0.3300±0.1031
Default	$\{0, 1, 2\text{–}3, 4+\}$	0.0898±0.0610	0.5981±0.2307
Fine	$\{0, 1, 2, 3, 4+\}$	0.0898±0.0610	0.5981±0.2307