BEYOND NOISY LABELS: ROBUST CONTINUAL LEARNING WITH DEPENDENT WEAK SUPERVISORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning from weak supervision must account for shared source errors, reliability drift, and partial corruption. TraceDuo combines temporal source-cluster label estimation, an exactly conserving source-deletion gradient decomposition, component-wise gradient clipping, and certified update routing. Full and adapter-only candidates are frozen before protected-risk and current-domain utility bounds govern commitment. Our analysis characterizes the independent-cluster support and score-separation regimes in which the pseudo-label bound becomes informative, and gives local alignment conditions for beneficial deletion-based component control. Under fresh conditional IID sampling, the V2 certificate bounds the probability of any invalid commitment by 0.05 per stream and limits net risk increase relative to a fixed initial anchor to 0.02 on the simultaneous coverage event. Across the V1 experiments on CIFAR-100 and the binary Amazon, Elliptic, and TON tasks, domain final balanced accuracies are 98.10%, 99.20%, 98.80%, and 99.30%, respectively. Their macro average is 98.85%, compared with 97.55% for RobustAgg+DERpp. A common task–time definition gives 99.17% current-task accuracy and 0.40 percentage points of forgetting. Factorial routing and optimizer comparisons quantify component effects, while a separate numerical check tests no-clipping equivalence. Known-risk V2 certification experiments record 95 erroneous five-window trajectories out of 100,000 with joint correction, versus 11,236 without it, and quantify when 512 observations provide useful endpoint power.

1 INTRODUCTION

Continual learning from weak supervision must account for sources that are neither independent nor stationary. In security, graph, text, and vision pipelines, supervisors may share a checkpoint, data provider, rule template, or annotation convention. Counting these outputs as independent votes can overstate support, while rejecting uncertain updates sacrifices adaptation. The challenge is to learn from new weakly supervised data while controlling the effect of dependent and partially corrupted sources on protected behavior.

TraceDuo uses *source-dependency provenance*: a declared partition into clusters that may share an error mechanism. It deduplicates upstream sources, tracks cluster reliability over time, and constructs a robust trimmed-log posterior. Sufficient cluster support, posterior confidence, and bounded conflict determine admission. Full and adapter-only candidates then use a source-deletion base–residual gradient decomposition; protected-risk upper bounds and current-domain gain lower bounds determine commitment.

Identical full posteriors can have different leave-one-cluster responses. Preserving these responses connects label estimation to update control, retaining source information discarded by aggregation alone. Matched-posterior comparisons isolate this use of provenance, while complete trajectories evaluate estimation, training, and routing together. Our contributions are threefold. First, TraceDuo combines temporal robust aggregation, an exactly conserving deletion decomposition, component clipping, and Full/Partial/Reject routing. Second, the analysis characterizes informative source-support regimes, equal-posterior deletion differences, and conditional local benefit; a separate simultaneous certificate provides repeated commitment and fixed-anchor net-risk protection. Third, four-domain V1 experiments achieve 98.85% macro Final BA versus 97.55% for RobustAgg+DERpp.

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

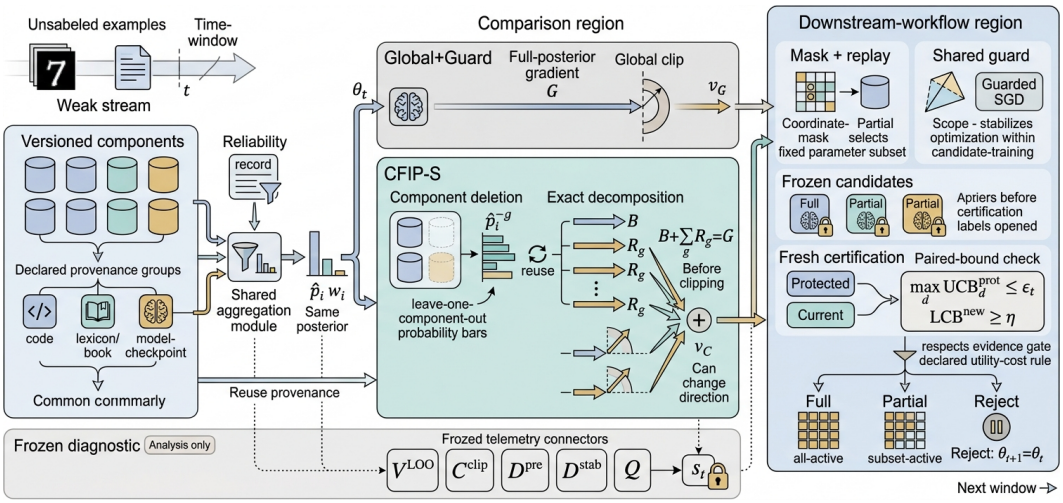


Figure 1: TraceDuo for dependent weak-supervision continual learning. Declared source-dependency clusters first support robust temporal label estimation and then deletion-based update influence control. Full and Partial candidates are trained before certification labels are opened; Reject preserves the incoming model when protected-risk or current-domain criteria are not met.

Factorial controls quantify update and routing effects, while separate V2 experiments measure endpoint power, candidate decisions, and trajectory-level certification failure under known population risks.

2 Related Work

Weak supervision and robust aggregation. Weak-supervision methods estimate latent labels from noisy sources, including models that account for source dependencies and structure selection Dawid & Skene (1979); Bach et al. (2017); Ratner et al. (2017; 2016); Varma et al. (2019); Northcutt et al. (2021). TraceDuo extends this perspective to reliability drift and cluster corruption, retaining deletion responses after aggregation to structure the parameter update.

Continual learning. Replay, distillation, and parameter-importance penalties support retention during adaptation Buzzega et al. (2020); Kirkpatrick et al. (2017); Lopez-Paz & Ranzato (2017); Chaudhry et al. (2019); Aljundi et al. (2019); Rebuffi et al. (2017); Parisi et al. (2019); De Lange et al. (2022). DERpp-soft and EWC-soft controls share soft targets, replay capacity, candidate masks, and certification routing, testing whether the provenance-based update provides additional benefit beyond these regularizers.

Risk certification. Candidate training precedes certification, and explicit endpoint bounds account for repeated decisions and selection, in the spirit of time-uniform confidence control Howard et al. (2021). The guarantee concerns risks on declared populations under the sampling model; prediction-level coverage, undeclared subgroups, and future distribution shifts require different targets.

3 Problem Setup

At window t, weak records x_i arrive with source outputs, identifiers, and a declared partition into G_t dependency clusters. Sources may abstain, share upstream errors, or belong to up to b corrupted clusters. Reliability, certification, analysis, and final testing use separate labelled resources; replay and training anchors use training records only.

Given the incoming model θ_t, Full updates all allowed parameters and Partial updates only the adapter and head. Both candidates are frozen before current certification features or labels are used. Reject preserves the model. V1 compares candidates with the incoming model; V2 additionally protects the fixed initial anchor. The primary empirical estimand is complete learner-level trajectory

performance, distinct from a same-state update or one candidate's improvement. Feasible candidates are ranked by a cost-adjusted gain lower bound.

4 TRACEDUO

Temporal labels and source deletion. Canonical provenance keys deduplicate upstream sources before within-cluster plurality voting; ties lead to abstention. Cluster reliability uses labelled checks, Beta smoothing, and drift-aware temporal averaging. A symmetric-error model maps reliability to class probabilities. Coordinate-wise trimming of $b = 2$ log probabilities at each end produces posterior $\widehat{p}_i$. Admission requires $K_i \geq 6$ active clusters, at least five after each deletion, confidence at least 0.55, and disagreement $\chi_i \leq 0.5$; the weight is $w_i = \max_c \widehat{p}_i(c)(1-\chi_i)$. Removing cluster g and repeating trimming and softmax gives p_i^{-g}, with reliability, admission, and weights fixed. Appendix H specifies these operations.

Conserving updates. For soft-target cross-entropy $\ell_i(\theta; q)$, candidate mask P_m, and an admitted weak batch I with $W = \sum_{i\in I} w_i > 0$, define

$$\bar{p}_i^{-} = \frac{1}{K_i}\sum_{g\in\mathcal{A}_i} p_i^{-g}, \qquad B = \frac{1}{W}\sum_{i\in I} w_i P_m \nabla \ell_i(\theta; \bar{p}_i^{-}),$$
$$R_g = \frac{1}{W}\sum_{i\in I: g\in\mathcal{A}_i} \frac{w_i}{K_i} P_m \left[\nabla \ell_i(\theta; \widehat{p}_i) - \nabla \ell_i(\theta; p_i^{-g})\right], \qquad B + \sum_g R_g = G. \tag{1}$$

Here $G = W^{-1}\sum_i w_i P_m \nabla \ell_i(\theta; \widehat{p}_i)$; absent slots contribute zero. Detach targets and weights from differentiation and use

$$g^{\text{weak}} = \text{clip}_{0.10}(B) + \sum_{g=1}^{G_t} \text{clip}_{0.02/G_t}(R_g), \quad \text{clip}_a(z) = z\min\{1, a/\|z\|_2\}, \quad \text{clip}_a(0) = 0. \tag{2}$$

Global clips G at 0.12. Without weak-gradient clipping, the two paths have equal gradients at matched states. Component bounds concern the specified decomposition; source replacement also affects posteriors, admission, and weights. Each candidate takes 24 AdamW steps with replay and a cumulative 0.05 displacement guard based on a separate training-gradient proxy (Appendix H).

4.1 EXECUTABLE RISK CERTIFICATION

V1 uses local-only protection; V2 adds a fixed initial anchor. Write Θ_0 for the initial model, $\Theta_{t-1} = \theta_t$ for the incoming model, and Θ_t for the retained model after decision t. Certification uses bounded error $\ell(\theta; x, y) = \mathbf{1}\{\arg\max_c f_\theta(x)_c \neq y\}$, with prediction ties resolved by smallest class index and evaluation randomness fixed. For declared populations $Q_1, Q_2, Q_{t,\text{new}}$, sample a uniform class and then an independent class-conditional observation. Thus $R_j(\theta) = \mathbb{E}_{Q_j}\ell(\theta; X, Y)$ equals population macro error; $n = 512$ counts mixture draws, not observations per class. For $j = 1, 2$, the five signed quantities are

$$\begin{aligned} Z^{\text{loc}}_{t,m,j} &= \ell(\theta_{t,m}; X, Y) - \ell(\Theta_{t-1}; X, Y), & D^{\text{loc}}_{t,m,j} &= \mathbb{E}_{Q_j} Z^{\text{loc}}_{t,m,j}, \\ Z^{\text{anc}}_{t,m,j} &= \ell(\theta_{t,m}; X, Y) - \ell(\Theta_0; X, Y), & D^{\text{anc}}_{t,m,j} &= \mathbb{E}_{Q_j} Z^{\text{anc}}_{t,m,j}, \\ Z^{\text{new}}_{t,m} &= \ell(\Theta_{t-1}; X, Y) - \ell(\theta_{t,m}; X, Y), & G^{\text{new}}_{t,m} &= \mathbb{E}_{Q_{t,\text{new}}} Z^{\text{new}}_{t,m}. \end{aligned} \tag{3}$$

Each variable lies in $\{-1, 0, 1\}$. Freeze candidates before certification; use fresh conditionally IID blocks across windows and a predeclared sample size. Within a window, both candidates and references share the same three blocks. Certification observations are not reused in later windows.

Count h paired reference-correct/candidate-wrong outcomes and g outcomes with the reverse pattern. With one-sided Clopper–Pearson bounds L_a, U_a (Appendix B), set

$$U^{\text{risk}} = U_{a_t}(h, n) - L_{a_t}(g, n), \qquad L^{\text{gain}} = L_{a_t}(g, n) - U_{a_t}(h, n), \qquad a_t = \frac{\delta_t}{20}, \quad \delta_t = \frac{0.05}{t(t+1)}. \tag{4}$$

The factor $20 = 2 \times 5 \times 2$ covers candidates, signed endpoints, and primitive tails. Equivalently, $\alpha_{t,m,e} = 2a_t$ gives $\sum_{t,m,e} \alpha_{t,m,e} \leq 0.05$, allowing within-window dependence and selection among the predeclared candidates. V2 feasibility is

$$\max_j U^{\mathrm{loc}}_{t,m,j} \leq \epsilon = 0.02, \qquad \max_j U^{\mathrm{anc}}_{t,m,j} \leq \tau = 0.02, \qquad L^{\mathrm{new}}_{t,m} > \eta_{\mathrm{gain}} = 0.005. \tag{5}$$

Rank feasible candidates by $S_{t,m} = L^{\mathrm{new}}_{t,m} - 0.001 c_m$, with $c_{\mathrm{Full}} = 1$, $c_{\mathrm{Partial}} = 0.2$, and Reject score zero. Select the highest positive score, breaking ties toward Partial; otherwise Reject. The anchor is a risk reference, distinct from the training-gradient proxy.

5 Theoretical Properties

Aggregation regime. Condition on input, true label y, history, and active cluster identities. Let $K > 2b$, $h = K - b$, and H contain h honest clusters. Normalized oracle score gaps $X_g(c) = u^*_g(y) - u^*_g(c)$ are conditionally independent across $g \in H$, with means at least γ. On an event E with $\Pr(E^c) \leq \delta_{\mathrm{est}}$, suppose $\max_{g \in H, c} |u_g(c) - u^*_g(c)| \leq e_s$. Normalization and estimation conditions appear in Appendix I.

Proposition 1 (Conditional label error) *Under these assumptions, counting ties as errors, the trimmed rule obeys*

$$\Pr(\widetilde{y} \neq y) \leq B_{\mathrm{lab}} := \min\left\{1, \delta_{\mathrm{est}} + (C-1)\exp\left(-\frac{[h(\gamma - 2e_s) - b]^2_+}{2h}\right)\right\}. \tag{6}$$

For $C \geq 2$ and $0 \leq \delta_{\mathrm{est}} < 1$, this expression is below one exactly when

$$\gamma - 2e_s > \frac{b}{h} + \sqrt{\frac{2}{h}\log\frac{C-1}{1-\delta_{\mathrm{est}}}}. \tag{7}$$

For CIFAR-100 with $K = 8, b = 2$, even ideal scores give $99e^{-4/3} \approx 26.10$ before truncation, hence $B_{\mathrm{lab}} = 1$. The bound diagnoses the independent-cluster support needed for informativeness; it is separate from final accuracy and posterior calibration.

Local update geometry. Equal full posteriors can have unequal deletion responses under the actual trimmed rule (Appendix A.4). These responses retain additional source information. Let $z = \nabla L_{\mathcal{D}}(\theta)$ and Δ_C, Δ_G be applied displacements for component and global clipping on an L-smooth clean loss. Then

$$\begin{aligned} A_{\mathcal{D}} &= -z^\top(\Delta_C - \Delta_G), \\ L_{\mathcal{D}}(\theta + \Delta_C) - L_{\mathcal{D}}(\theta + \Delta_G) &\leq -A_{\mathcal{D}} + \tfrac{L}{2}(\|\Delta_C\|^2 + \|\Delta_G\|^2). \end{aligned} \tag{8}$$

Benefit requires alignment sufficient to dominate the curvature term. This is a local sufficient condition, and AdamW or the guard can change alignment. Endpoint performance is measured below; directly testing this mechanism requires matched-state gradients and displacements.

Theorem 1 (Valid commitment and fixed-anchor protection) *Let $\mathcal{H}_t$ contain the complete past, current training resources, frozen candidates and references, risk definitions, and predeclared test configuration, but exclude the fresh certification observations. Conditional on $\mathcal{H}_t$, assume each block is IID from its declared population. Keep Q_1, Q_2 fixed over time. Under Equations 4 and 5, for every finite T,*

$$\begin{aligned} \Pr\{\exists t \leq T : \textit{an invalid candidate is committed}\} &\leq \sum_{t=1}^{T} \delta_t \\ &= 0.05\left(1 - \frac{1}{T+1}\right). \end{aligned} \tag{9}$$

Moreover, $\Pr\{\exists t \geq 1, j : R_j(\Theta_t) - R_j(\Theta_0) > 0.02\} \leq 0.05$.

An invalid commitment violates local or anchor tolerance 0.02, or has gain at most 0.005. The bound is per learning stream, rather than conditional on acceptance. Probability budget δ_t differs from permitted risk increment: five local tolerances of 0.02 permit 0.10 net increase. V2 constrains net risk relative to the initial anchor to 0.02 on simultaneous coverage, for fixed declared populations under the stated sampling model. This does not bound the sum of positive successive increases. Proofs and remaining scope details are in Appendices A and I.

6 Experimental Design

V1 evaluates CIFAR-100 and binary Amazon, Elliptic, and TON tasks, with 60 roots per domain and five windows. Every task contains all domain classes and 1,600 class-balanced test records; the sequence is not disjoint-class incremental. Each window uses 6,000 weak records, 128 reliability labels, three certification blocks of 512, and a replay capacity of 512, with eight source clusters and $b = 2$. Initialization, analysis, and final testing use separate labels (Appendix J).

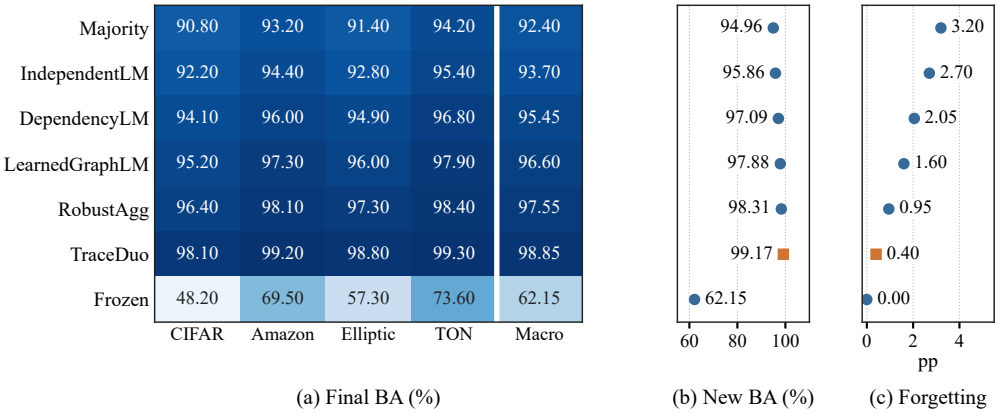
Final BA averages final task accuracies; New BA averages post-decision diagonal accuracies; forgetting averages each earlier task's within-run maximum minus final accuracy. Domain–root pairs are equally weighted. Twenty thousand within-domain root-bootstrap replicates preserve pairing and complete trajectories; intervals are descriptive conditional on the fixed test set. Equations 43–45 give the definitions. All comparisons share V1 task definitions.

V2 separately evaluates the anchored certificate on declared 100-class populations and threshold classifiers with known risks: nine conditions, 100,000 independent five-window trajectories each. It measures certification error and power, separately from V1 neural training and its local-only rule. V1 final accuracy, V2 invalid commitment, and attack success are distinct estimands. Matched-state numerical checks use a separate linear model.

7 Results

7.1 Overall Performance and Temporal Retention

TraceDuo reaches 98.85% macro Final BA, versus 97.55% for RobustAgg and 96.60% for LearnedGraphLM (Figure 2). Domain means are 98.10%, 99.20%, 98.80%, and 99.30%; the four RobustAgg contrasts are 1.70, 1.10, 1.50, and 0.90 pp. New BA is 99.17% with 0.40 pp forgetting, versus 98.31% and 0.95 pp for RobustAgg. Complete domain and macro intervals appear in Figure 17.

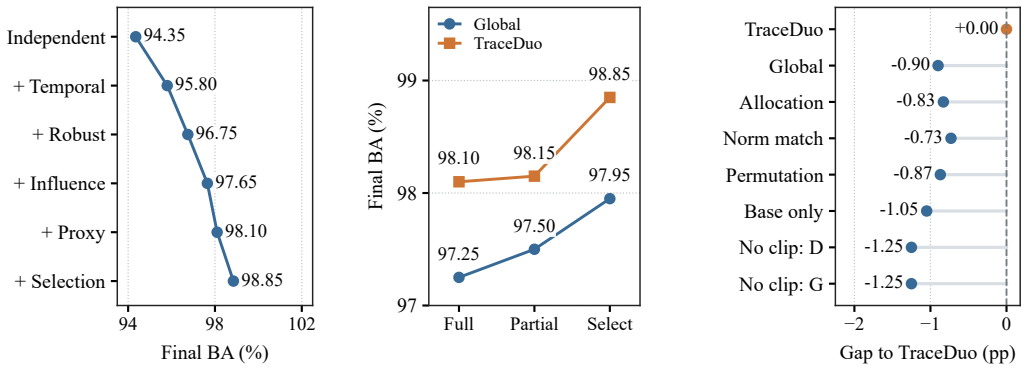


(a) Final BA (%) (b) New BA (%) (c) Forgetting

Figure 2: V1 overall performance. (a) Domain and equally weighted macro Final BA; Amazon, Elliptic, and TON are binary tasks. (b) New BA and (c) forgetting use the common task–time definitions. Shared DERpp suffixes are omitted. All values aggregate 60 roots per domain.

7.2 Components, Routing, and Matched Posteriors

The cumulative sequence rises from 94.35% to 98.85%, with a final 0.75 pp routing increment (Figure 3a). These increments are order-dependent. The factorial control retains an update-rule advantage under Full-only and selection: 0.85 and 0.90 pp, respectively. The routing interaction is $(98.85 - 98.10) - (97.95 - 97.25) = 0.05$ pp. Paired-root descriptive intervals are $[0.8947, 0.9053]$ for TraceDuo–Global selection and $[0.7447, 0.7552]$ for within-TraceDuo routing.



(a) Cumulative (b) Routing (c) Controls

Figure 3: Separating update components and routing. (a) Cumulative components retain the common certificate. (b) Both candidates are trained in each factorial cell; the allowed commitment set changes. (c) Final BA gaps relative to TraceDuo under matched posteriors. Norm matching uses an extra reference candidate; D/G denote decomposed/Global updates without clipping.

Norm-matched Global reaches 98.12%, leaving a 0.73 pp TraceDuo advantage. Matching rescales final displacements after AdamW and the guard, with extra reference-candidate computation; it is a diagnostic, not an equal-cost baseline. Allocation, permutation, and base-only controls reach 98.02%, 97.98%, and 97.80%. Both no-clipping endpoints are 97.60%. A separate linear-model check over 1,152 steps per precision gives maximum parameter differences 1.18×10^{-16} (float64) and 5.22×10^{-8} (float32), with unchanged frozen coordinates (Figure 12). These controls quantify endpoint effects; preferential clipping of negatively aligned responses requires matched-state alignment measurements.

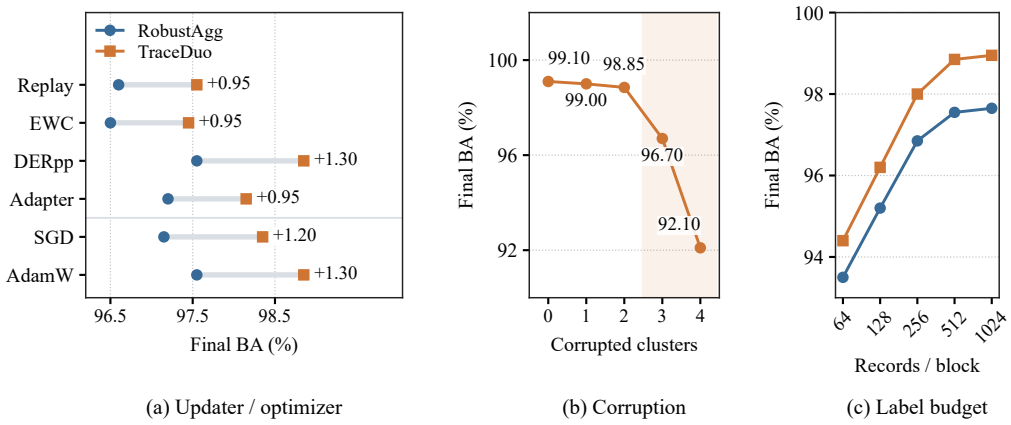
7.3 Transfer, Corruption, and Label Resources

TraceDuo improves over RobustAgg across Replay, EWC-soft, DERpp-soft, and Adapter-only by 0.95, 0.95, 1.30, and 0.95 pp (Figure 4a). Guarded SGD gives a 1.20 pp contrast, versus 1.30 pp for AdamW: a 0.10 pp method-by-optimizer interaction. Adapter-only uses the Partial-only configuration.

With zero through four corrupted clusters, Final BA is 99.10, 99.00, 98.85, 96.70, and 92.10%; three and four exceed the configured budget. Graph perturbations instead quantify partition structure (Figure 16). Increasing certification records from 512 to 1,024 gives 0.10 pp, with paired interval $[0.0946, 0.1054]$. Strong-label uses follow $256 + 5(128 + 3n)$, excluding analysis and final-test labels. A separate frozen candidate rejects $n = 64, 128, 256$ and accepts $n = 512, 1024$ while predictions remain fixed (Figure 15).

7.4 Certification Validity, Power, and Anchor Protection

Joint correction yields 95 erroneous five-window trajectories out of 100,000 at the protection boundary: 0.0950%, with one-sided 95% Monte Carlo upper bound 0.112654%. Without correction, 11,236 trajectories fail: 11.2360%, upper bound 11.401599% (Figure 5b). The corrected five-window theorem bound is 4.1667%. These frequencies describe the evaluated boundary; the theorem supplies conditional coverage.



(a) Updater / optimizer (b) Corruption (c) Label budget

Figure 4: Transfer and sensitivity in V1. (a) Paired RobustAgg (blue) and TraceDuo (orange) means; annotations show pp gains. (b) Actual corruption at configured $b = 2$; shading marks counts beyond the budget. (c) Complete trajectories at each certification block size, distinct from a fixed-candidate power experiment.

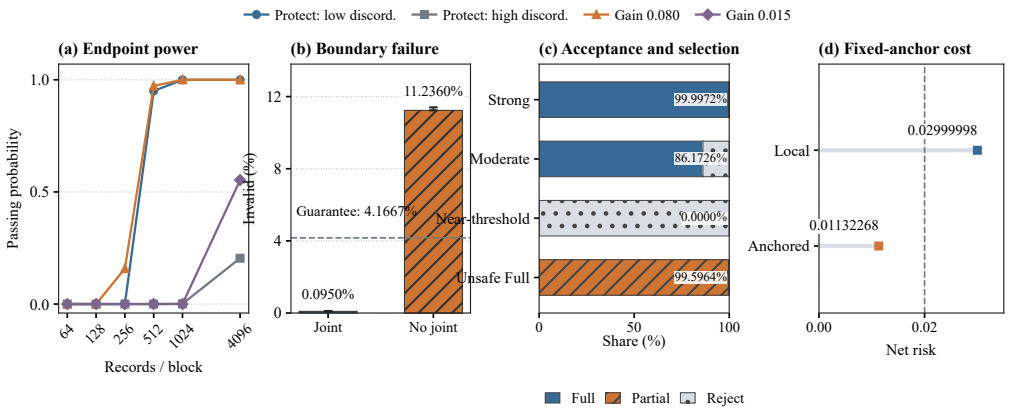


Figure 5: V2 validity, power, and long-term protection. (a) Separate fixed-size endpoint power evaluations. (b) Protection-boundary invalid commitment frequencies with marginal one-sided 95% Monte Carlo upper bounds. (c) Full/Partial/Reject shares; labels give total acceptance. (d) Mean final risk increase under local-only and anchored routing; the dashed line is 0.02. Each trajectory condition uses 100,000 five-window runs; drift uses $n = 4096$, and the other displayed trajectory conditions use $n = 512$.

What 512 observations certify. At window five, $a_5 = 1/12000$ and $U_{a_5}(0, 512) = 0.0181778 < 0.02$; zero observed discordance requires at least 465 observations. For the same true protected difference -0.002, power is 0.950086 at $(p_H, p_G) = (0.0001, 0.0021)$ but 8.35303×10^{-5} at $(0.05, 0.052)$. A true gain of 0.015 has power 1.99290×10^{-8} at $n = 512$, rising to 0.553290 at $n = 4096$. Power depends on discordance as well as the mean margin. With $h = g = 0$, gain LCB remains negative.

Strong, moderate, and near-threshold gain conditions accept 99.9972%, 86.1726%, and 0% of opportunities. Unsafe-Full/safe-Partial accepts 99.5964%, exclusively Partial. In the drift condition, local-only routing accepts 100% with no local/utility violation, yet every trajectory exceeds the initial anchor tolerance. Anchoring reduces acceptance to 40.8710% and mean final net increase from 0.02999998 to 0.01132268. Zero observed anchor violations still has a positive upper bound of 0.002996%.

A checkable decision trace. Figure 6 shows all ten candidate-level bound triples. The sequence is Partial, Full, Partial, Reject, Partial: at window four both candidates pass protection but fail utility with LCB -0.0173689. Figure 11 gives all 50 component endpoints and paired counts. This V2 ledger is separate from the four-domain V1 trajectories.

Candidate-wise resolution. The signed bounds make each routing decision inspectable. At window one, Full has local and anchor UCBs of 0.061905, whereas Partial has 0.012971 for both; both gain LCBs exceed 0.005, so protection selects Partial. At window two, both candidates satisfy all five endpoints, and Full wins by score (0.057079 versus 0.040783). At window three, Full again fails local and anchor protection (0.073319 and 0.062498), while Partial passes protection (0.016419 and 0.016202) with gain LCB 0.024665. At window four, the candidates share passing protection maxima of 0.017400 and 0.016486, but their gain LCB is -0.017369, producing Reject. At window five, Full fails protection (0.062510 local and 0.053806 anchor); Partial passes with UCBs 0.018178 and 0.017745 and gain LCB 0.019677. These comparisons use full-precision bounds; the displayed rounding serves only presentation.

Selection and repeated decisions. The two candidates share each window's observations, so their statistics are dependent. Validity comes from allocating primitive tails across every candidate and signed endpoint before observing the block; the union bound requires no independence among these within-window events. The candidates, thresholds, score, and tie rule are fixed before the certification labels are revealed, and selection occurs only after all simultaneous bounds are formed. Across windows, fresh conditionally IID blocks prevent earlier routing from turning reused labels into an adaptive test set. This construction gives a single stream-level error budget over the full sequence rather than five unadjusted window-level claims.

Failure budget versus risk budget. The Monte Carlo boundary study evaluates how often the routed trajectory commits a population-invalid candidate under the declared known-risk model; its 0.0950% frequency is an empirical operating-point measurement, while Theorem 1 provides the conditional simultaneous bound. The risk tolerance has a different role. Local constraints compare a candidate with the incoming model and can permit successive increments to accumulate. The anchor endpoints compare every candidate with Θ_0, turning the same 0.02 tolerance into a net-risk constraint on the simultaneous coverage event. In the drift condition, this distinction changes the mean final protected-risk increase from 0.02999998 under local-only routing to 0.01132268 with anchoring, together with the measured acceptance reduction from 100% to 40.8710%.

Paired endpoints and macro risk. Each protection block records the four joint outcomes of the incoming or anchor reference and the frozen candidate. Only the two discordant counts h and g enter the signed risk difference. Bounding their marginal probabilities separately and subtracting the one-sided limits yields a valid UCB for harm minus repair; reversing the signs yields the gain LCB. This paired construction uses the observed discordance structure while retaining the bounded 0–1 loss definition. The class-uniform mixture makes the population risk equal to macro error, but the block remains 512 independent mixture draws: it is not treated as 512 observations for every class. Consequently, the certificate targets the declared macro population rather than simultaneous per-class recall. Fixed per-class protection would require stratified supports and an additional allocation over classes.

Finite-sample operating regimes. The exact power curves separate mean improvement from the probability of a discordant pair. Protection with true difference -0.002 is readily detected when turnover is sparse, but a high-turnover pair with the same mean difference has little power at 512 observations. Utility is more demanding: even a true gain of 0.015 rarely clears the window-five LCB at 512, whereas the strong 0.080 gain reaches high power. The zero-discordance calculation explains why 512 can certify protection near the 0.02 threshold, while the modest-gain calculation explains conservative rejection near the 0.005 utility threshold. Figure 5 therefore reports acceptance together with endpoint power and error frequency; a high acceptance rate alone is not used as evidence for statistical validity.

Evidence scope. Figure 6 plots confidence bounds rather than empirical risk estimates: protection decisions use the upper limits and utility uses the lower limit. The complete endpoint view in Figure 11 retains all 50 signed comparisons, paired counts, thresholds, and population values. The V2 experiments use declared known-risk threshold populations to evaluate coverage and routing power; the four-domain V1 experiments separately evaluate trained-model accuracy, forgetting, transfer, and component controls. The V2 guarantee therefore applies prospectively to its frozen candidates, fresh conditional IID blocks, fixed evaluation randomness, and declared protected populations. It does not retroactively certify V1 or extend to an undeclared future distribution. Within that scope, the experiment checks the full decision path: simultaneous candidate–endpoint correction, selection after bound construction, Reject semantics, fixed-anchor accounting, and a positive confidence upper bound even when no invalid trajectory is observed.

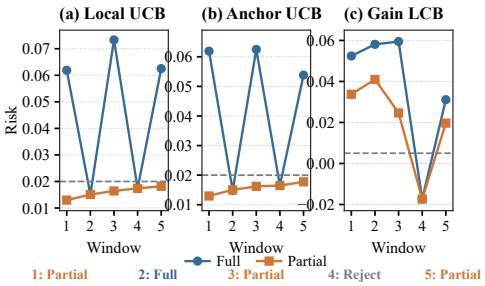


Figure 6: V2 candidate bounds and committed decisions. Full (blue) and Partial (orange) share certification observations. (a,b) Maximum protection UCB across the two groups, relative to incoming and initial models. (c) Current-domain gain LCB. Dashed thresholds are 0.02, 0.02, and 0.005; protection permits equality, utility requires strict improvement. The bottom strip records the committed candidate or Reject.

Conclusion. TraceDuo uses source provenance for temporal weak-label estimation and deletion-based gradient component control under dependent, drifting weak supervision. The theory characterizes informative aggregation regimes and local alignment conditions; under the stated sampling assumptions, V2 certifies commitments while protecting fixed-anchor net risk.

V1 achieves 98.85% macro Final BA versus 97.55% for RobustAgg+DERpp. V2 separately quantifies power, selection, and the acceptance cost of long-term protection. Sharper large-class bounds and matched-state alignment measurements would extend this characterization.

REFERENCES

Rahaf Aljundi, Min Lin, Baptiste Gouinguenet, Lucas Caccia, Tinne Tuytelaars, and Yoshua Bengio. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 273–282. PMLR, 2017.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Matthias De Lange, Tinne Tuytelaars, Joost van de Weijer, et al. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Steven R. Howard, Aaditya Ramdas, Jon D. McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

German I. Parisi, Ronald Kemker, Jose L. Part, Chris Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment*, volume 11, pp. 269–282, 2017. doi: 10.14778/3157794.3157797.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Paroma Varma, Frederic Sala, Sen Wu, Alexander Ratner, and Christopher Ré. Learning dependency structures for weak supervision models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6418–6427. PMLR, 2019.

AI USE STATEMENT

OpenAI ChatGPT was used solely for grammar correction and language polishing of author-written text throughout the manuscript.

REPRODUCIBILITY STATEMENT

The code and supporting materials for this work will be released at `https://anonymous.4open.science/r/F2-7D2B/`.

A PROOFS, OPERATING CONDITIONS, AND IMPLEMENTATION LINKS

A.1 COORDINATE-WISE TRIMMING WITH DIFFERENT RETAINED SETS

We prove Proposition 1 for the class-wise operation in Equation 36, without assuming common retained clusters across classes. Fix the conditioning in Section 5; all probabilities below refer to this conditional law. If fewer than b clusters are corrupted, choose any fixed $h = K - b$ honest clusters as H and treat the remaining b coordinates as unrestricted. This choice must not depend on realized scores. The assumption of independent honest gaps is not inferred from the number of declared clusters or from conditioning on admission.

Put $m = K-2b$. For each class c, write the sorted observed honest coordinates as $u^H_{(1),c}, \dots, u^H_{(h),c}$, and all sorted observed coordinates as $u_{(1),c}, \dots, u_{(K),c}$. For $1 \le j \le m$, interlacing gives

$$u^H_{(j),c} \le u_{(b+j),c} \le u^H_{(b+j),c}. \tag{10}$$

The lower inequality holds because at most b unrestricted values can precede the jth honest value; the upper holds because adding values cannot raise an order statistic with a fixed rank. Ties do not change either inequality. Let $T_c = \sum_{j=1}^{m} u_{(b+j),c}$ and $S^H_c = \sum_{g \in H} u_g(c)$. Since honest coordinates are in $[0,1]$, summing the interlacing inequalities yields

$$S^H_c - b \le \sum_{j=1}^{m} u^H_{(j),c} \le T_c \le \sum_{j=b+1}^{h} u^H_{(j),c} \le S^H_c. \tag{11}$$

In particular, applying the lower bound to y and the upper bound to c gives $T_y - T_c \ge S^H_y - S^H_c - b$. This compares separately sorted lists; it does not assert that the same honest values survive both trims.

On the simultaneous estimation event E,

$$T_y - T_c \ge \sum_{g \in H} X_g(c) - 2he_s - b. \tag{12}$$

This allows estimation error to change every sorting order and retained set. It uses no stability assumption about trimming indices. The common average denominator m is positive, softmax preserves argmax, and the affine score normalization preserves the trimmed class ranking. Hence an error implies $T_c \ge T_y$ for some $c \ne y$.

Each oracle gap lies in $[-1,1]$, so Hoeffding's inequality gives, when $a = h(\gamma - 2e_s) - b > 0$,

$$\Pr\left(\sum_{g \in H} X_g(c) \le b + 2he_s\right) \le \exp\left(-\frac{a^2}{2h}\right). \tag{13}$$

For $a \le 0$, the probability is at most one. Take a union bound over $C-1$ wrong classes and add $\Pr(E^c) \le \delta_{\text{est}}$ to prove Equation 6. We have not conditioned Hoeffding's inequality on E: the inclusion of the error event intersected with E in the oracle tail events is enough. Thus estimation failure may depend on oracle scores, and errors for different wrong classes need not be independent. An unconditional statement requires the assumptions almost surely under the conditioning, or valid averaging of the corresponding conditional bounds.

A.2 When the Bound Is Informative

Taking logarithms of $\delta_{\text{est}} + (C-1)\exp(-[h(\gamma-2e_s)-b]_+^2/(2h)) < 1$ gives Equation 7, including the strict boundary for $C = 2, \delta_{\text{est}} = 0$. If $\delta_{\text{est}} \geq 1$ the expression is trivial. Since $\gamma \leq 1$ and $e_s \geq 0$, a necessary feasibility condition is

$$\frac{(K-2b)^2}{2(K-b)} > \log\frac{C-1}{1-\delta_{\text{est}}}. \tag{14}$$

For a desired bound $\alpha \in (\delta_{\text{est}}, 1)$, a sufficient and equivalent condition on this expression is

$$\gamma - 2e_s \geq \frac{b}{h} + \sqrt{\frac{2}{h}\log\frac{C-1}{\alpha-\delta_{\text{est}}}}. \tag{15}$$

Table 1 evaluates the formula at ideal parameters; these are analytic values, not empirical error rates. For $b = 2, C = 100$, $K = 15$ is the first integer $K > 4$ for which even this ideal expression is nontrivial. The actual protocol has only eight clusters. With $K = 8, b = 2$, only $C \leq 4$ can yield a nontrivial bound, even ideally.

Table 1: Best-case values of Equation 6 with $\gamma = 1$, $e_s = \delta_{\text{est}} = 0$. These are formula evaluations.

K	b	C	Before truncation	B_{lab}
8	2	2	$e^{-4/3} \approx 0.2636$	0.2636
8	2	4	$3e^{-4/3} \approx 0.7908$	0.7908
8	2	100	$99e^{-4/3} \approx 26.10$	1
15	2	100	$99e^{-121/26} \approx 0.9430$	0.9430

A.3 From Reliability Counts to Score Error

The following analysis links reliability counts to the reference estimator under explicit sampling and drift assumptions. Empirical calibration requires the corresponding record-level measurements. Let r^*_{tg} be the current population probability that active cluster g is correct. Assume its $n > 0$ reliability observations are conditionally IID Bernoulli with this parameter and independent of the current training record. Conditional IID here concerns observations for one cluster; different clusters may share reliability labels. For a declared failure allocation α_{tg}, Hoeffding and $|(c+1)/(n+2) - c/n| \leq 1/(n+2)$ imply

$$|\rho^{\text{fresh}}_{tg} - r^*_{tg}| \leq f_{tg} := \min\left\{1, \sqrt{\frac{\log(2/\alpha_{tg})}{2n}} + \frac{1}{n+2}\right\} \tag{16}$$

except on an event of probability at most α_{tg}. Adaptive reuse, nonstationarity within the label pool, and unrepresentative active observations require a different sampling guarantee.

Suppose d_{tg} bounds $|r^*_{tg} - r^*_{t-1,g}|$, and $e^\rho_{t-1,g}$ is a valid previous reliability error radius. On the simultaneous count event, the actual branch in Equation 35 has radius

$$e^\rho_{tg} = \min\left\{1, \begin{cases} e^\rho_{t-1,g} + d_{tg}, & n = 0, \\ f_{tg}, & \text{reset branch}, \\ 0.8(e^\rho_{t-1,g} + d_{tg}) + 0.2 f_{tg}, & \text{EMA branch}. \end{cases}\right\} \tag{17}$$

Initialize a new cluster with radius $1/2$ around $\rho = 1/2$. These branchwise bounds hold on one event even though the reset decision depends on the counts; a deterministic radius may take the maximum over possible branches. A data-dependent radius cannot simply be substituted into a fixed-parameter tail bound without further justification. The threshold 0.15 detects estimator disagreement, not a bound on population drift. Without a drift assumption, use $d_{tg} \leq 1$, which can be uninformative. For a finite horizon, $\sum_{t,g} \alpha_{tg} \leq \delta_{\text{est}}$ gives a simultaneous event by a union bound. An unbounded horizon needs a summable allocation. Instantiating this analysis with 128 labels additionally requires the sampling law, drift bounds, and failure allocation.

For completeness, let $p_c(r)$ be the symmetric-error vector before flooring, $a_c(r) = \max\{\varepsilon, p_c(r)\}$, and $Z(r) = \sum_c a_c(r)$. Then $1 \leq Z(r) \leq 1 + C\varepsilon$ and $q_c(r) = a_c(r)/Z(r) \geq \varepsilon/(1+C\varepsilon)$, proving

the log-score range used in the main text. At a fixed hard vote, $|a_c(r) - a_c(r')| \le |r - r'|$ and $|Z(r) - Z(r')| \le 2|r - r'|$. The logarithm is $1/\varepsilon$-Lipschitz on $[\varepsilon, \infty)$ and 1-Lipschitz on $[1, \infty)$, giving

$$|u_c(r) - u_c(r')| \le \min\left\{1, \frac{\varepsilon^{-1} + 2}{M}|r - r'|\right\}. \tag{18}$$

Thus a deterministic simultaneous reliability radius e_ρ gives $e_s = \min\{1, (\varepsilon^{-1} + 2)e_\rho/M\}$. This conservative global conversion can saturate at one; tighter bounds would need documented interior probability ranges. A positive honest margin, cross-cluster independence, and posterior calibration require separate conditions beyond reliability accuracy and the score-error bound, particularly under the symmetric-error approximation.

A.4 Equal Full Posteriors under the Actual Trimmed Rule

Consider binary records with eight distinct, active clusters, $b = 2$, all voting for class 1. Let σ denote the logistic function and set $\rho_g = \sigma(r_g)$, $q_g = (\sigma(r_g), 1 - \sigma(r_g))$. Take

$$\begin{aligned} r^{(A)} &= (0.2, 0.4, 0.6, 0.8, 1.2, 1.4, 1.6, 1.8), \\ r^{(B)} &= (0.2, 0.4, 0.5, 0.9, 1.1, 1.5, 1.6, 1.8). \end{aligned} \tag{19}$$

No probability floor binds. Since $\log q_g(1)$ increases and $\log q_g(2)$ decreases with r_g, trimming two from either end retains ranks 3–6 for both classes. Their aggregate log-odds are the average of those four r_g, which is exactly 1 in both configurations. Thus $\widehat{p}^{(A)} = \widehat{p}^{(B)} = (\sigma(1), 1 - \sigma(1))$. Both satisfy confidence $\sigma(1) > 0.55$, conflict zero, and all active-count requirements, and have the same weight.

Delete cluster 3, preserving the original reliabilities as specified. Re-sorting seven values and retaining ranks 3–5 gives log-odds $(0.8 + 1.2 + 1.4)/3 = 17/15$ in A and $(0.9 + 1.1 + 1.5)/3 = 7/6$ in B. Consequently $p^{-3,(A)} \neq p^{-3,(B)}$ despite identical full posteriors. This is an exact mathematical construction using the reference aggregation and admission rule, not an additional experimental observation. It establishes non-identifiability of deletion responses from the full posterior; a single binary record can still yield collinear gradient components. Rotation needs the non-collinearity condition in Equation 41.

A.5 Conservation, Rotation, and Local Descent

Write $J_i = \partial z_\theta(x_i)/\partial\theta$, so $\nabla\ell_i(\theta; q) = J_i^\top(f_\theta(x_i) - q)$. Soft targets, admission, and weights are stop-gradient quantities. Linearity in q implies $\nabla\ell_i(\theta; \bar{p}_i^-) = K_i^{-1}\sum_{g\in\mathcal{A}_i}\nabla\ell_i(\theta; p_i^{-g})$. Adding the per-record residual sum cancels all deletion gradients and leaves $\nabla\ell_i(\theta; \widehat{p}_i)$. Multiplication by the same mask and w_i/W, followed by summation, proves Equation 1 even with varying active sets. Changing weights or denominators under deletion would invalidate this identity. Subtracting gradients gives Equation 40 directly.

Substitute $G = B + \sum_g R_g$ into $v_C - v_G$ to prove Equation 41. Orthogonal projection onto $G^\perp$ proves the non-collinearity criterion. When it vanishes, the scalar is $G^\top v_C/\|G\|^2$, so zeroing or reversal must be distinguished from positive rescaling. When $G = 0$, the global update is zero, yet unequal component clipping can destroy cancellation and give $v_C \neq 0$. The allocation-only control has $B = 0.8G$ and $R_g = 0.2G/G_t$, hence always has a nonnegative scalar multiple of G after clipping. Arbitrary conserving decompositions can also rotate: deletion is distinguished by its source response interpretation, not by unique access to nonlinear geometry.

If all clipping factors equal one, the weak gradients agree. If all deletion posteriors equal the full posterior, residuals vanish; equality with Global additionally requires the base and global clips to agree on G, since their configured radii are 0.10 and 0.12. Equal weak gradients propagate to equal moments and parameters only from matched states with identical replay, masks, randomness, and downstream operations. A global renaming of residual slots preserves their sum under equal radii, but independent per-record permutations can change the aggregated residual vectors and their clipping.

For an L-smooth clean loss along each displacement segment, integration of the gradient gives $|L_\mathcal{D}(\theta + \Delta) - L_\mathcal{D}(\theta) - z^\top\Delta| \le L\|\Delta\|^2/2$. Subtract the lower bound for Global from the upper

bound for component clipping to obtain Equation 8; substitute $\Delta = -\eta v$ for Equation 42. The alignment condition can be written without a hidden residual label:

$$z^\top R_g = \frac{1}{W} \sum_{i:g\in\mathcal{A}_i} \frac{w_i}{K_i} \langle J_i P_m z, p_i^{-g} - \widehat{p}_i \rangle. \tag{20}$$

Let $\beta = z^\top B$, $r_g = z^\top R_g$, and $\mathcal{N} = \{g : r_g < 0,\ a_g < a_G\}$. Define

$$H = \sum_{g\in\mathcal{N}} (a_G - a_g)(-r_g), \qquad P = [-(a_0 - a_G)\beta]_+ + \sum_{g\notin\mathcal{N}} [-(a_g - a_G) r_g]_+. \tag{21}$$

Then $S \geq H - P$. Hence $H > P$ and $0 < \eta < 2(H-P)/(L(\|v_C\|^2 + \|v_G\|^2))$ suffice for strict local improvement when the denominator is positive. If $L = 0$, any $\eta > 0$ in the stated smooth region suffices when $H > P$. Overclipping useful components, deletion-insensitive corruption, or adverse base alignment can eliminate this guarantee. Applying the sufficient condition requires verifying $H > P$ for the actual component alignments. For norm-matched nonzero *applied* displacements of norm r, a sufficient condition is $A_{\mathcal{D}} > Lr^2$; an angle change alone is insufficient.

A.6 Component Bounds and Source-Replacement Sensitivity

Let $D \sim_g D'$ denote replacing one source cluster's outputs with all other inputs and incoming training state fixed. The replacement may change $\widehat{p}$, every deletion posterior except the one deleting g (when other inputs and reliabilities truly stay fixed), B, other residuals, weights, and the admitted set. Even with frozen admission and weights, non-expansiveness of Euclidean ball projection gives only

$$\|v_C(D) - v_C(D')\| \leq \|B(D) - B(D')\| + \sum_h \|R_h(D) - R_h(D')\|. \tag{22}$$

The bound involves all affected components. The radius of R_g alone is insufficient for source-replacement sensitivity; even changing an isolated clipped component gives a difference bound of $2\kappa_g$, not κ_g. The weak-gradient norm bound 0.12 implies the coarse difference bound 0.24, and the candidate radius 0.05 gives $\|\Delta(D) - \Delta(D')\| \leq 0.10$ for candidates from the same incoming state. Neither bound attributes influence to a source or shrinks with the size of its perturbation.

A useful perturbation-sensitive result would require conditions controlling all affected posteriors and weights, a positive batch-weight denominator, fixed active sets and vote identities or explicit treatment of their changes, and a margin from the confidence/conflict admission thresholds. Without such a margin, an arbitrarily small score change can switch admission and alter the batch normalization discontinuously. Multiple training steps also require stability of model Jacobians, optimizer state, and the guard; final routing introduces further threshold changes. These additional requirements delineate a source-level stability analysis beyond the component bounds proved here.

A.7 Repeated Acceptance

Conditional on $\mathcal{H}_t$ in Theorem 1, the fixed candidate/reference pair maps each IID record into mutually exclusive indicators H, G. Thus $h \sim \mathrm{Bin}(n, p_H)$ and $g \sim \mathrm{Bin}(n, p_G)$ marginally, although they are not independent. The one-sided Clopper–Pearson construction gives $\Pr\{p_H > U_{a_t}(h,n) \mid \mathcal{H}_t\} \leq a_t$ and $\Pr\{p_G < L_{a_t}(g,n) \mid \mathcal{H}_t\} \leq a_t$. Their simultaneous success implies both the relevant upper protection bound and lower gain bound by subtraction. The reverse tails are not used.

Let F_t be failure of any of the 20 assigned primitive bounds. A conditional union bound gives $\Pr(F_t \mid \mathcal{H}_t) \leq 20a_t = \delta_t$, without independence between candidates, groups, or endpoints. On F_t^c, all candidates have valid endpoints simultaneously; selecting among feasible candidates by their displayed scores preserves the population inequalities. Taking expectations over adaptive histories and a union bound over windows,

$$\Pr(\text{an invalid commitment by } T) \leq \sum_{t=1}^{T} \mathbb{E}[\Pr(F_t \mid \mathcal{H}_t)] \leq 0.05 \left(1 - \frac{1}{T+1}\right). \tag{23}$$

On the simultaneous coverage event, an accepted model satisfies the anchor constraint directly. At a rejection, $\Theta_t = \Theta_{t-1}$, preserving it. The base case is $R_j(\Theta_0) - R_j(\Theta_0) = 0$. Induction establishes

the anchor claim for every $t \le T$; continuity of probability as $T \to \infty$ establishes the unlimited-stream bound. This is a sufficient multiple-testing construction, not a new concentration inequality.

For general predictable endpoint allocations, it suffices that $\sum_{t,m,e} \alpha_{t,m,e} \le \delta$ almost surely and that each endpoint has conditional failure probability at most its assigned $\alpha_{t,m,e}$. Here e indexes five endpoints and $\alpha_{t,m,e} = 2a_t$, so the endpoint and primitive allocations agree. Candidates may depend on past certification outcomes because new observations retain the stated conditional law. Reusing a fixed pool affected by those decisions does not satisfy this argument merely because its labels never enter backpropagation.

For local-only protection, telescoping on the same Q_j gives $R_j(\Theta_T) - R_j(\Theta_0) \le \sum_{t:\text{accepted}} \epsilon_t$. This is a risk-increment budget, distinct from $\sum_t \delta_t$. Even with an anchor, the path $0.10, 0.11, 0.10, 0.11, 0.10, 0.11$ satisfies maximum net increase 0.01 but accumulates 0.03 in positive increments. The theorem controls the former quantity only.

B Reference Algorithm and Execution Order

The following reference specification defines the window update. The batch denominator W is fixed across all source deletions within a batch, but not across different batches.

Algorithm 1: TraceDuo window update

Version: V2 certification; the V1 result tables retain local-only routing.

Input: incoming $\theta_t = \Theta_{t-1}$, fixed model Θ_0, buffer $\mathcal{M}_t$, weak pool U_t, reliability pool, source keys/partition, unopened certification blocks.

Output: committed parameters, buffer, reliability state, and audit record.

1. Canonicalize keys; deduplicate; abstain on conflicting duplicates.
2. Compute cluster pluralities and update ρ_{tg} by Eq. 35.
3. Compute $\widehat{p}_i$, conflict, admission, and w_i by Eq. 36.
4. **If** no record is admitted, **return** $(\theta_t, \mathcal{M}_t, \text{Reject})$.
5. Compute active-cluster deletion posteriors; freeze targets, weights, anchor gradients, and candidate sample sequences.
6. **For** $m \in \{\text{Full}, \text{Partial}\}$:
   Clone θ_t; set mask P_m; initialize optimizer moments to zero.
   For $s = 1, \ldots, 24$:
      Read 128 admitted weak records and 32 replay records; fix batch W.
      Compute B, R_g (Eq. 1); clip (Eq. 2).
      Add clipped replay gradient (Eq. 37).
      Update moments; form preconditioned, decayed, masked d_s (Eq. 38).
      Project cumulative displacement and write parameters (Eq. 39).
   Freeze candidate and its parameter fingerprint.
7. Access fresh certification observations only after both candidates are frozen.
8. Allocate $a_t = 0.05/[20t(t+1)]$; compute ten signed endpoints from 20 primitive bounds (Eq. 4), including initial-anchor checks.
9. Discard infeasible candidates; rank positive cost-adjusted scores; ties prefer Partial.
10. **If** none remains, preserve model and buffer and **return Reject**.
11. Otherwise commit the selected candidate and reservoir-update the replay buffer.
12. Record the decision, bounds and margins, optimizer/configuration identifiers, incoming/candidate/initial-anchor fingerprints, block identifiers, class supports, and paired counts.

Adapter and replay state. For backbone output $h \in \mathbb{R}^d$, use $h' = h + ABh$, with $A \in \mathbb{R}^{d\times 8}$, $B \in \mathbb{R}^{8\times d}$, and a linear head. Initially $A = 0$ and $B_{ij} \sim N(0, 0.02^2)$; subsequent windows inherit committed parameters. Full trains backbone, adapter, and head; Partial trains adapter and head. Candidates share the root-determined sequence of admitted weak records, reshuffled at exhaustion. The buffer starts with the 256 initialization samples and one-hot targets. Its capacity is 512; stored soft targets and logits remain frozen while an entry is retained. After commitment, admitted samples in canonical sample-ID order enter reservoir sampling: for the jth cumulative entry draw k uniformly from $1, \ldots, j$, replacing slot k only if $k \le 512$ (unfilled slots are populated first). New weak entries use frozen posteriors and committed-model logits. Certification, analysis, and final-test labels never enter replay. Rejection leaves the buffer and model unchanged but need not undo reliability estimation.

Optimizer and projection details. SGD uses $v_s = 0.9v_{s-1} + g_s$ and direction $d_s = -0.05P_m(v_s + 0.01\theta_s)$, with zero initial momentum and the same guard. AdamW uses Eq. 38, with $\widehat{m}_s = m_s/(1 - 0.9^s)$ and $\widehat{v}_s = v_s/(1 - 0.999^s)$. Sort initialization samples by canonical ID and use the first and last 128 as the two anchor groups. Compute and freeze their average CE gradients at θ_t. To project onto $\mathcal{K}_m$, enumerate the feasible active sets of the two halfspaces in the masked subspace and choose the closest solution; the cone projection followed by radial clipping is the projection onto the cone intersected with the radius-0.05 ball. The reference specification uses float64 for projection before writing to parameter precision. Removing the training proxy removes the halfspaces but retains mask and radius.

B.1 Exact One-sided Bounds and Error Allocation

For a binomial count k among n independent observations, let $\mathrm{B}^{-1}(p; a, b)$ denote the beta quantile. The reference bounds are

$$L_\alpha(k,n) = \begin{cases} 0, & k = 0, \\ \mathrm{B}^{-1}(\alpha; k, n-k+1), & k > 0, \end{cases}$$
$$U_\alpha(k,n) = \begin{cases} 1, & k = n, \\ \mathrm{B}^{-1}(1-\alpha; k+1, n-k), & k < n. \end{cases} \tag{24}$$

Each tail uses a_t, not half of an implicitly two-sided confidence level. Require integer counts $n \geq 1$, $h, g \geq 0$, and $h + g \leq n$; invalid inputs cannot produce a certificate. With $h = 0$, $U_a(0,n) = 1 - a^{1/n} > 0$: observing no new errors is not evidence of zero error probability. Both candidates and both references can share the three blocks within a window; simultaneous coverage accounts for all 20 primitive events without their mutual independence. The local-only V1 rule uses 12 events; V2 recomputes the bounds with the 20-event allocation for anchor protection.

Fresh sampling and fixed looks. After freezing all candidates and evaluation randomness, draw fresh IID class-mixture observations for each declared block. Do not use current certification features to screen candidates before drawing their labels. Past observations cannot be reused under the fresh-sample proof, even if their only influence was through past accept/reject decisions. Finite-pool sampling without replacement, dependent graph nodes, and records from the same device require inference matched to those sampling units. Conditional IID validity depends on the collection process as well as split separation.

The reported sample-size curves represent separate fixed-size configurations, not repeated sampling until a candidate passes. For five predeclared nested looks, a sufficient alternative is $a_{t,k} = \delta_t/(20 \times 5)$ and a union bound over looks. That alternative is not used in the reported V2 tables. Arbitrary optional stopping would require a separately specified confidence-sequence or e-process procedure.

Stream and label budgets. The theorem spends 0.0416667 across five windows and at most 0.05 across an unlimited stream. Joint protection across all 240 V1 streams would require an additional allocation, such as $0.05/240$ per stream, with correspondingly recomputed bounds and power. V2 uses the per-stream budget. The two anchor comparisons reuse the existing protected labels, so five windows still use $5(3 \times 512) = 7{,}680$ certification labels. Under the label-usage accounting, $256 + 5(128 + 3 \times 512) = 8{,}576$ strong-label usage slots, excluding analysis and final testing. Fresh certification observations are required by the theorem; usage slots count label accesses, with unique-sample accounting specified separately. Anchor evaluation adds prediction work and tightens multiplicity while retaining the same label usage.

C Certification Protocol V2: Power and Recorded Decisions

This appendix specifies the V2 certification experiments and their recorded decisions, evaluated separately from the four-domain V1 training comparison. The loss, sampling unit, paired counts, error allocation, reference models, and strictness of the decision inequalities define a reproducible endpoint calculation. Endpoint validity and aggregate model accuracy are distinct evaluation targets.

Table 2: Protocol scope. Additional anchor comparisons reuse the two protected blocks but increase the number of statistical events.

Item	Local-only rule	V2 rule
Protected references	Incoming model	Incoming and initial models
Signed endpoints per candidate	3	5
Primitive tails, two candidates	12	20
Five-window error allocation	$\sum_{t\le 5}\delta_t$	$1/24$
Risk guarantee on coverage	Local increments	Local and anchor increments
Evidence reported here	V1 training records	925-3 certification records

C.1 Sampling, Class Support, and Endpoint Records

Under the class-uniform IID mixture, for any fixed candidate/reference pair,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \frac{1}{C}\sum_{c=1}^{C}\mathbb{E}[Z \mid Y = c]. \tag{25}$$

The binomial marginal model applies to the n independent mixture draws. Random class supports n_c may be unequal or zero. Coverage pertains to the declared class-uniform mixture; per-class coverage requires additional endpoints. Validity follows from the mixture sampling law, including its random class supports. In particular, $512/100 = 5.12$ is only the expected support per class in the 100-class configuration.

For a different design with fixed stratified supports n_c, a sufficient alternative is to allocate primitive tails $a_{t,m,e,c}$ across classes, and use, for protection,

$$U^{\mathrm{strat}}_{t,m,e} = \frac{1}{C}\sum_{c=1}^{C}\left[U_{a_{t,m,e,c}}(h_c, n_c) - L_{a_{t,m,e,c}}(g_c, n_c)\right], \quad \alpha_{t,m,e} = 2\sum_{c} a_{t,m,e,c}. \tag{26}$$

The corresponding lower gain bound reverses the signs. Conditional IID sampling within each class and an overall summable allocation suffice; cross-class independence is unnecessary for the union bound. A class with $n_c = 0$ contributes the uninformative interval $[-1, 1]$, rather than borrowing the full block's sample size. Per-class safety would require each class's own inequality. This alternative is not the protocol used for the V2 tables.

Each window record associates the frozen model/reference identifier with the population definition, block identifier, draw size, class supports, paired error counts, tail allocation, bound direction, threshold, margin, score, and final decision. Protected margins are $0.02 - U$; gain margins are $L - 0.005$, with zero allowed only for protection. The integer counts and full-precision bounds, not display rounding, determine feasibility. Candidate freezing, sampling independence, and absence of cross-window reuse are data-collection conditions of the theorem, additional to the counts and model identifiers needed for endpoint calculation.

C.2 Sample Size and Exact Endpoint Power

Figure 7 shows the complete five-window allocation. When $h = g = 0$, the protected bound is $1 - a_t^{1/n}$, so the smallest sample size for that particular observation pattern is

$$n_{0,t} = \left\lceil \frac{\log a_t}{\log(1 - 0.02)} \right\rceil. \tag{27}$$

This minimum applies specifically to zero observed discordance; sufficiently many repairs can allow other count patterns to pass. As t increases, a_t decreases and the required sample size grows. Thus power at fixed $n = 512$ remains window-dependent even with an unlimited-stream error guarantee.

For a fixed endpoint with paired harm/repair probabilities (p_H, p_G), the exact passing probability is

$$\pi(n, p_H, p_G, a) = \sum_{h=0}^{n}\sum_{g=0}^{n-h}\frac{n!\,p_H^h\,p_G^g\,(1 - p_H - p_G)^{n-h-g}}{h!\,g!\,(n-h-g)!}\mathbf{1}\{(h,g) \in \mathcal{A}(n,a)\}, \tag{28}$$

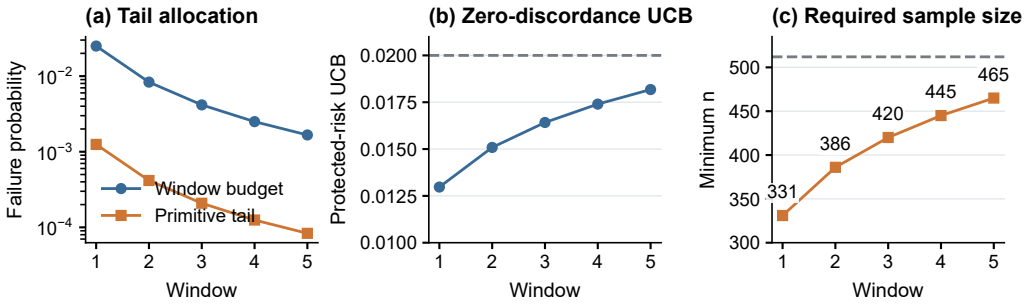


Figure 7: Window-wise certification allocation. (a) Stream-level failure budget and primitive one-sided tail allocation. (b) Protected-risk UCB at $n = 512$ with no observed discordance; the dashed line is 0.02. (c) Minimum sample size for that observation pattern; the dashed line is 512.

where $\mathcal{A}$ is $U_a(h,n) - L_a(g,n) \le 0.02$ for protection or $L_a(g,n) - U_a(h,n) > 0.005$ for gain. This is a single-endpoint probability, not a probability that all five endpoints or a routed candidate pass. The exact calculation conditions on h and sums the binomial tail $g \mid h \sim \mathrm{Bin}(n-h, p_G/(1-p_H))$ above the smallest admissible g, weighted by $h \sim \mathrm{Bin}(n, p_H)$.

Figure 8 uses $a = a_5 = 1/12000$ throughout. Low/high protected turnover has $(p_H, p_G) = (0.0001, 0.0021)/(0.05, 0.052)$, respectively, both with mean difference -0.002. Strong/modest gain has $(0.001, 0.081)/(0.001, 0.016)$, giving gains 0.080/0.015.

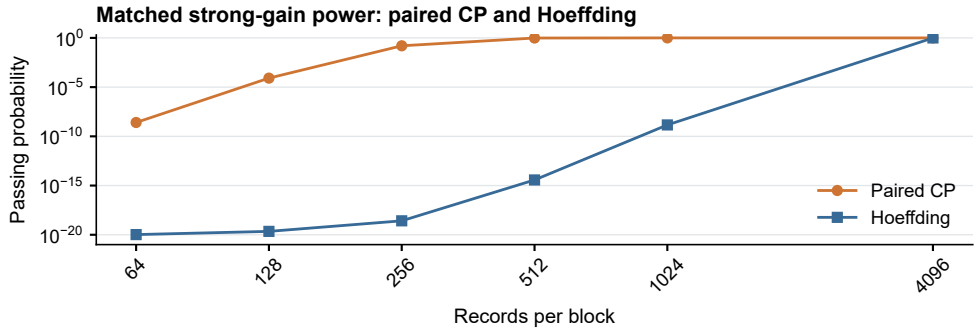


Figure 8: Strong-gain passing probability under paired Clopper–Pearson and Hoeffding bounds. Each point is a separate fixed-size design with the same signed-endpoint failure budget. Values below the plotted range remain available in the internal plotting data.

For scale, a local-only design with five windows, two candidates, and three endpoints equally dividing 0.05 gives the Hoeffding radius $\sqrt{2\log(30/0.05)/512} \approx 0.1581$ for a paired difference in $[-1, 1]$. The V2 comparison must use its own allocation: at window five, the signed endpoint failure budget is $2a_5 = 1/6000$, giving $\sqrt{2\log(6000)/512} \approx 0.184343$. Under the strong-gain pair above, the passing probability is 3.70395×10^{-15} for this Hoeffding rule versus 0.972982 for the paired CP rule. The comparison uses the same signed-endpoint failure budget and demonstrates the role of sparse discordance for these pairs. Optimality across other distributions would require a separate comparison. The power curves evaluate separate predeclared sample sizes; repeated looks require an additional sequential allocation.

C.3 Declared Populations and Trajectory Outcomes

All statistical conditions use $C = 100$ and known-risk threshold classifiers. For population identifier d, independently draw $V \sim \mathrm{Uniform}\{0, \ldots, 99\}$ and $U \sim \mathrm{Uniform}(0,1)$; set $X = (d, V, U)$, $Y = V$, and

$$\widehat{Y}_r(X) = \begin{cases} (V+1) \bmod 100, & U < r, \\ V, & U \ge r. \end{cases} \tag{29}$$

Its population error is exactly r. The two protected anchor risks are 0.10 and 0.08. New-domain identifiers are disjoint, and incoming risk on each not-yet-updated domain is 0.25. Within a block, all models share the same observations; across windows the draws are fresh. The trajectory records use joint multinomial interval counts equivalent to drawing these shared uniforms, preserving dependence among endpoints. Incoming protected risks change after the actual commitment; Reject preserves them. These conditions isolate certification behavior with known risks.

For strong, moderate, near-gain, and utility-boundary conditions, Full/Partial local protected changes are $-0.004/-0.002$ in both groups. Their gains are, respectively, $(0.10, 0.08)$, $(0.06, 0.045)$, $(0.012, 0.010)$, and $(0.0049, 0.0040)$. At the protection boundary, Full/Partial anchor differences are $+0.0201/+0.0250$ in group 1 and $-0.0201/-0.0250$ in group 2, with gains $0.10/0.08$; only group 1 violates protection. In the unsafe-Full condition, Full has anchor differences $+0.03$ and gain 0.12; Partial has local changes -0.002 and gain 0.08. Drift uses local increases $+0.006/+0.004$, gains $0.15/0.13$, and $n = 4096$. Other conditions use $n = 512$. Full/Partial here index the two declared candidate rules; these threshold-classifier experiments isolate statistical routing behavior, with backbone and adapter training evaluated separately in V1.

There are 100,000 independent trajectories per configuration and five decisions per trajectory, totaling 900,000 trajectories and 4,500,000 decisions. The configuration seeds in the displayed condition order are 202709280–202709288. For trajectory r, define

$$E_r = \mathbf{1}\{\exists t \le 5 : \text{an invalid candidate is committed}\},$$
$$A_r = \frac{1}{5}\sum_{t=1}^{5} \mathbf{1}\{\text{commit at } t\}. \tag{30}$$

The reported error frequency is $N^{-1}\sum_r E_r$, whereas acceptance is $N^{-1}\sum_r A_r$. Standard errors for acceptance use trajectories as the independent units, not the correlated window decisions. The upper Monte Carlo bound is $U_{0.05}(\sum_r E_r, N)$ and is marginal per row; it is distinct from the $\delta = 0.05$ budget inside each learning stream. For the local-only drift ablation, the common five-endpoint reference event also counts violations of the omitted anchor condition. Figure 9 separates those events from local/utility violations.

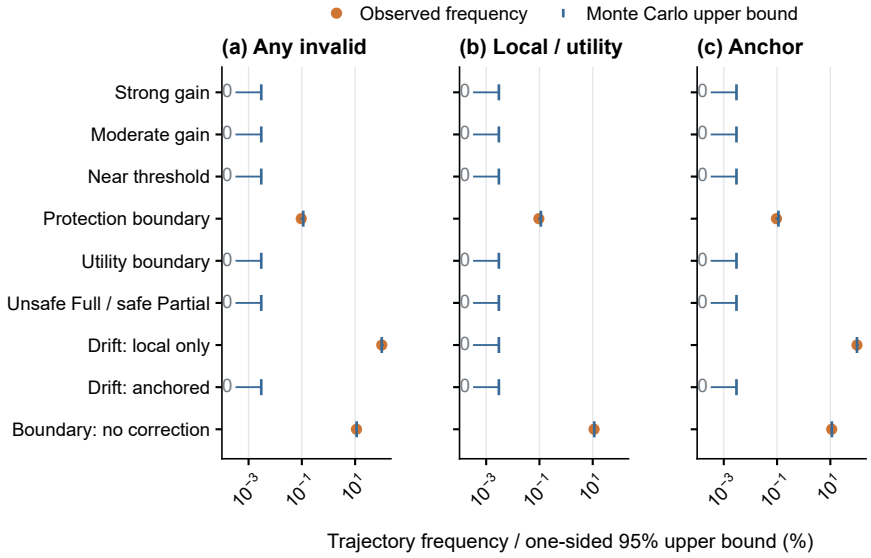


Figure 9: Complete V2 trajectory outcomes. Observed frequencies (circles) and marginal one-sided 95% Monte Carlo upper bounds (caps) across 100,000 independent five-window trajectories per condition. Zero observations retain their positive upper bounds. Local/utility and anchor columns separate the reference-rule components; the local-only drift row violates the omitted anchor constraint. Drift uses $n = 4096$ and other rows use $n = 512$.

For the moderate-gain condition, acceptance decreases across windows from 98.5390% to 94.2060%, 87.7720%, 78.3110%, and 72.0350%, under the decreasing tail allocation. Its overall acceptance standard error is 0.046861 pp. For drift, the local-only mean final anchor increment is

0.02999998, with 99th percentile 0.030; with anchoring these become 0.01132268 and 0.014. The latter accepts 100.000%, 94.350%, 7.115%, 1.891%, and 0.999% of opportunities across the five windows. Its lower overall acceptance is the measured cost of the additional constraint.

C.4 Every Candidate, Endpoint, and Decision in the V2 Ledger

The separate ledger has seed 20270927 and three blocks of 512 observations per window, independently of the V1 main training trajectories. The retained protected risks are $(0.098, 0.078)$, $(0.094, 0.074)$, $(0.092, 0.072)$, $(0.092, 0.072)$, and $(0.090, 0.070)$ after its five decisions. Figure 10 displays the maximum protection bounds, utility bound, and score for both candidates. Figure 11 shows the individual endpoint estimates and bounds rather than only these maxima.

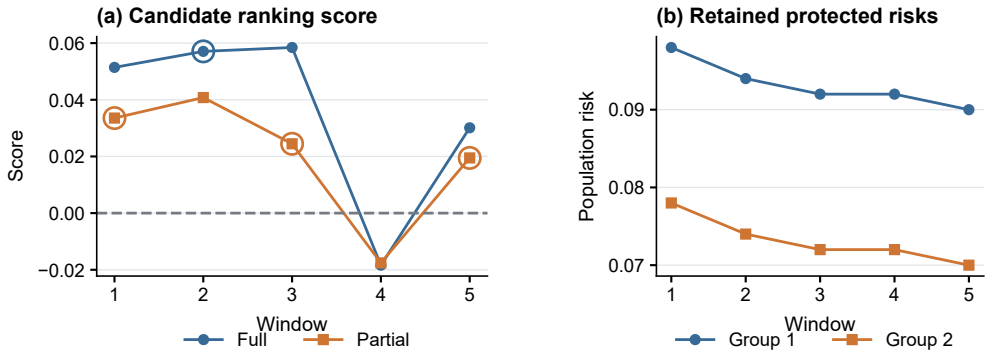


Figure 10: V2 candidate ranking and retained risks. (a) Candidate scores over five windows; open rings identify selected candidates. Feasibility still requires both protection maxima at most 0.02 and gain LCB above 0.005. (b) Retained protected population risks after each decision.

C.5 Canonical Protocol Specification and Evidence Scope

We use the signed endpoint inequalities, full-precision bounds, and block definitions as the canonical protocol specification throughout the analysis. Feasibility in Figure 6 follows directly from these inequalities: Full is infeasible at windows 1, 3, and 5 because protection fails; both candidates are infeasible at window 4 because utility fails. The selected candidates follow the specified feasible-set ranking. Signed inequalities and full-precision endpoint records take precedence over auxiliary summary flags.

The canonical V2 design uses $a_5 = 1/12000$ per stream and block size $n = 512$ in total. The ledger field `descriptor_sha256` identifies threshold-rule descriptions for the known-risk certification populations. These identifiers refer to the V2 candidate rules, separately from the trained V1 checkpoints. Consequently, validity and power are attributed to the corresponding V2 experiments; V1 retains its local-only protocol. Joint allocation across streams, such as $0.05/240$, would define a different calibration with different bounds and power.

The complete endpoint visualization uses $\widehat{D} = (h - g)/n$ for local/anchor protection and $\widehat{G} = (g - h)/n$ for utility. L1/L2 and A1/A2 denote local and anchor group endpoints, respectively; N denotes new-domain gain. All rows have $n = 512$ and use their window's a_t in Figure 7. Margins are $0.02 - U$ (L/A) or $L - 0.005$ (N). The plotted values are rounded to six decimals, while the full-precision bounds determine the decisions. The reported population values are known from the declared threshold rules and are used only for evaluating validity; candidate selection uses the empirical bounds.

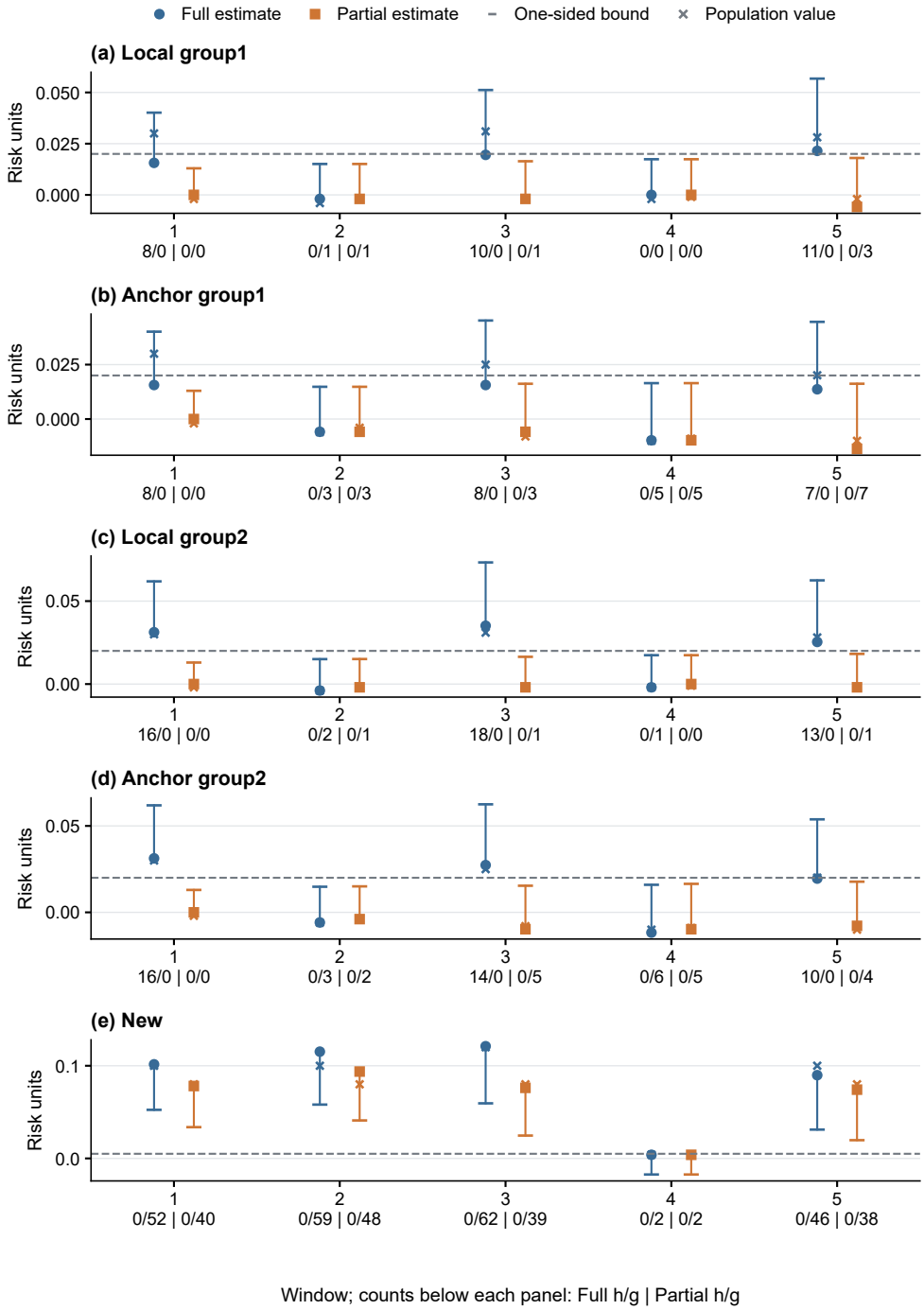


Figure 11: All 50 V2 signed endpoints. Points show empirical paired differences, caps the relevant one-sided UCB or LCB, and crosses the declared population values. Dashed lines mark 0.02 for local/anchor protection and 0.005 for new-domain gain. Protection passes below its line; gain passes strictly above its line. Counts below each window are Full h/g and Partial h/g. Both candidates share observations.

D UPDATE GEOMETRY AND NUMERICAL EQUIVALENCE

Matched-posterior comparisons fix admitted records, targets, weights, optimizer family, budgets, and certification. At a shared state they also fix actual replay entries. Global clips the reconstructed gradient at 0.12; allocation-only uses $B_{\text{alloc}} = 0.8G$ and $R_{\text{alloc},g} = 0.2G/G_t$ before the same component clipping. Independent per-record permutations act on all G_t residual slots, including zero slots, before aggregation. A consistent global relabeling instead transports labels, state, radii, and ordering together and is an identity control. Deduplication likewise makes copies under one canonical key an identity control.

For norm matching, first obtain Global and TraceDuo final candidate displacements from the same incoming model and set

$$\Delta_{\text{NG}} = \frac{\|\Delta_{\text{TD}}\|_2}{\|\Delta_{\text{G}}\|_2}\Delta_{\text{G}} \quad (\|\Delta_{\text{G}}\|_2 > 0). \tag{31}$$

Zero Global displacements remain zero and are recorded separately. The scale does not use certification labels. The reference guard is a homogeneous cone intersected with a ball, so scaling along a feasible ray to a norm within the ball preserves its geometric constraints. Protected-risk feasibility is evaluated separately by the certification endpoints. The extra TraceDuo candidate needed to obtain the norm must count as diagnostic computation.

Stepwise numerical equivalence. The recorded check uses a 16-feature, four-class linear softmax model, 64 weak records, eight slots with six to eight active per record, 32 replay records, 12 numerical seeds, two masks, two optimizers, and 24 steps. The Partial mask updates only the last eight parameter rows; it is not the deep rank-8 adapter. Both paths disable weak-gradient clipping while keeping replay clipping, optimizer state, decay, the training proxy, and the cumulative guard. Each precision contains 1,152 comparisons. Figure 12 reports the resulting numerical discrepancies.

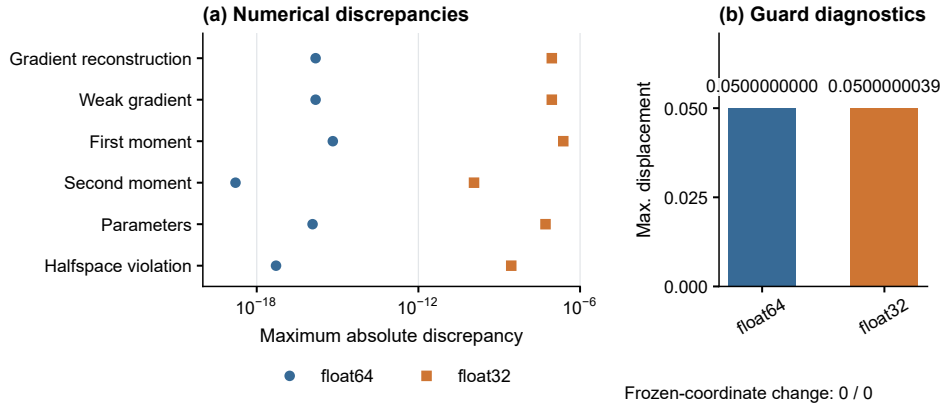


Figure 12: Numerical equivalence check. Nonzero discrepancies are shown on a logarithmic scale for float64 and float32. Frozen-coordinate changes are zero in both precisions and are annotated separately; cumulative displacement is reported on its native scale. The main performance evaluation uses the separate four-domain experiments.

Equal no-clipping Final BA values of 97.60% are a separate reporting-level observation. Stepwise equality follows algebraically only with matched inputs, states, masks, randomness, and subsequent operators. The linear check supports that matched-state statement for the specified linear model. Extending the check to deep-model trajectories requires the same state and operator matching throughout the sequence.

D.1 SELECTIVE ERROR AND MECHANISM REPORTING

The following definitions specify additional record-level diagnostics beyond the current endpoint evaluation. For each domain and window, use an analysis-only labelled pool sampled from incoming weak records before admission. Hold model state, reliability, source outputs, and the aggregation

rule fixed. Let V_i indicate that the active count and all deletion-count requirements hold, and set

$$A_i(\tau, \xi) = V_i \mathbf{1}\{\max_c \widehat{p}_i(c) \geq \tau,\ \chi_i \leq \xi\}. \tag{32}$$

For N analysis records, define

$$\widehat{\pi}_{\tau,\xi} = \frac{1}{N} \sum_i A_i(\tau, \xi), \qquad \widehat{r}_{\tau,\xi} = \frac{\sum_i A_i(\tau, \xi) \mathbf{1}\{\widetilde{y}_i \neq y_i\}}{\sum_i A_i(\tau, \xi)}. \tag{33}$$

When no record is admitted, coverage is zero and error is undefined, not zero. The coverage denominator includes unsupported records, which are automatically rejected. Report error against coverage over confidence thresholds at fixed $\xi = 0.5$, marking the deployed $\tau = 0.55$, and separately vary conflict if those records exist. Admission at thresholds admitting ties must use a fixed prediction tie rule. Threshold sweeps are descriptive, not a license to select training thresholds using analysis or final-test labels.

Because training uses confidence/conflict weights, also report

$$\widehat{r}^{w}_{\tau,\xi} = \frac{\sum_i A_i(\tau, \xi) \bar{w}_i \mathbf{1}\{\widetilde{y}_i \neq y_i\}}{\sum_i A_i(\tau, \xi) \bar{w}_i}, \qquad \bar{w}_i = \max_c \widehat{p}_i(c)(1 - \chi_i), \tag{34}$$

when the denominator is positive. This weighted error is a different estimand from selective error and complements calibration diagnostics. A top-label calibration diagnostic compares empirical correctness with mean confidence within declared bins, with bin counts reported. These conditional quantities require per-record confidence and correctness measurements beyond overall pseudo-label accuracy or final-model BA.

Report results by domain/window with admitted and total counts before forming any explicitly weighted aggregate. Uncertainty should preserve complete-root blocks and pairing; pointwise intervals across thresholds are not simultaneous bands. Source-level support additionally requires non-abstention and correctness counts n_{tg}, c_{tg}, the actual EMA/reset branches, and justified sampling and drift assumptions to instantiate Appendix A.3. Instantiating these diagnostics requires a joined per-record table of true labels, full weak posteriors, conflict, and source support. Those measurements are outside the current endpoint evaluation, which reports label budgets, final-model predictions, and an independent candidate ledger. The definitions here specify the additional analysis; error–coverage curves and calibrated weak-label error remain unmeasured.

Assessing the predicted mechanism. At matched incoming states, a held-out clean analysis loss would provide z, separately from training and certification. Together with B, R_g and the clipping factors, it identifies signed component alignments, $H - P$, and $\|Q_G d\|$. Complete applied displacements identify $A_{\mathcal{D}}$ after replay, AdamW, and the guard; norm-matched displacements distinguish magnitude from alignment. Positive S is only a first-order diagnostic unless the curvature remainder is controlled, and it does not imply positive $A_{\mathcal{D}}$. Reporting both positive and nonpositive alignment cases would assess the success and failure regimes of the local result. These matched-state alignment measurements remain unmeasured for the main V1 trajectories; the current comparisons quantify endpoint effects.

E Supplementary Protocol, Metrics, and Candidate Records

E.1 Experimental Records and Protocol Scope

The V1 figures summarize the experiments under a common protocol. The record hierarchy comprises prediction arrays, class-count tables, run metrics, task–time records, and summaries. The log reports 13,440,000 prediction entries for seven main methods, 10,080 run summaries for 42 configurations, and 50,400 task–time entries. The manifest organizes records by domain, root, configuration, and seed. The independent candidate ledger is analyzed separately from the main comparison.

The binary tasks and 97.55% RobustAgg baseline define the V1 comparison. Reuse of 98.85% in a main, ablation, transfer, or sensitivity analysis denotes the same configuration. Accuracy and temporal metrics are evaluated under these task definitions; matched attack and clean-acceptance aggregates are separate estimands.

Table 3: V1 record layout and reference resource schedule.

Field	Value
Domains / classes	CIFAR-100 / 100; Amazon, Elliptic, TON / 2 each
Roots / windows	60 per domain, 31001–31060 / 5
Task / final test records	1,600 per balanced task / 8,000 per root
Weak / reliability per window	6,000 / 128
Initialization / replay capacity	256 / 512
Certification / analysis	Three blocks of 512 / 256 per window
Clusters / configured corruption	8 / 2
Candidate steps / batches	24 / weak 128, replay 32
Primary optimizer	Complete-step AdamW; cumulative radius 0.05
Bootstrap	20,000 whole-root replicates; seed 20270926

E.2 Unified Task–Time Metrics

Figure 13 gives the mean trajectory. The computation uses within-root task histories before averaging; a maximum of mean curves is not substituted for the mean of within-root maxima. The log reports four forgetting contributions of 0.50, 0.45, 0.40, and 0.25 pp. Their mean is 0.40 pp; the five final tasks average to 98.85%, and the post-commitment diagonal averages to 99.17%. Display rounding does not define the calculation.

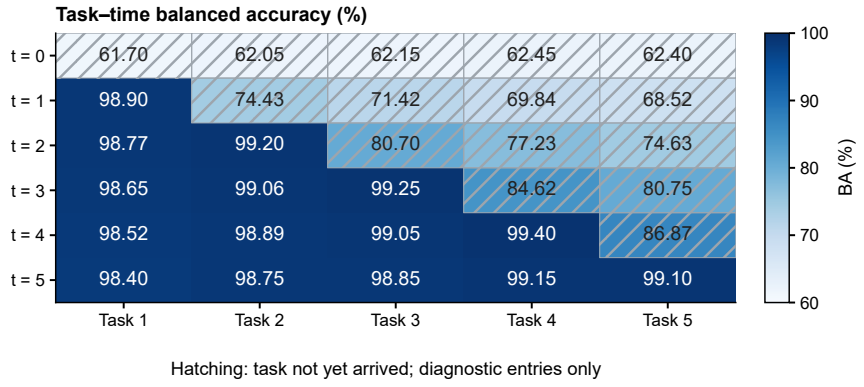


Figure 13: TraceDuo task–time means. Cells report percentages. Outlined cells contribute to New BA, and the final row determines Final BA. Entries for future tasks before arrival are diagnostic and do not enter the temporal metrics at those times.

For each root, the sum of the first four final-task errors cannot exceed the sum over all five. Since each history maximum is at most 100, dividing by four gives $F \le (5/4)(100 - \mathrm{FinalBA})$ after averaging. If forgetting and Final BA instead used exactly the same tasks and normalized weights, the bound would be $100 - \mathrm{FinalBA} = 1.15$ pp. The defined V1 metric uses the former bound, 1.4375 pp, which the observed forgetting satisfies.

The domain-level Final BA and aggregate temporal metrics are reported once in Figure 2; Figure 13 supplies the complete trajectory needed to audit their definitions.

E.3 Independent Candidate Ledger and Fixed-candidate Budget

The candidate ledger is independent of the main comparison. Figure 14 reports both frozen candidates, their diagnostic test BA, bounds, and decisions. Test labels do not enter routing. At window one, Full-only would Reject while selection commits Partial. Only this first divergence is identified; the remaining selected-branch candidates cannot stand in for a retrained Full-only branch.

Figure 15 uses the signed decision inequalities and full-precision endpoints as the canonical specification. The first three rows fail protection, while the final two satisfy both constraints. This fixed-candidate experiment isolates the effect of certification sample size; main-trajectory clean acceptance and attack success require separate evaluations.

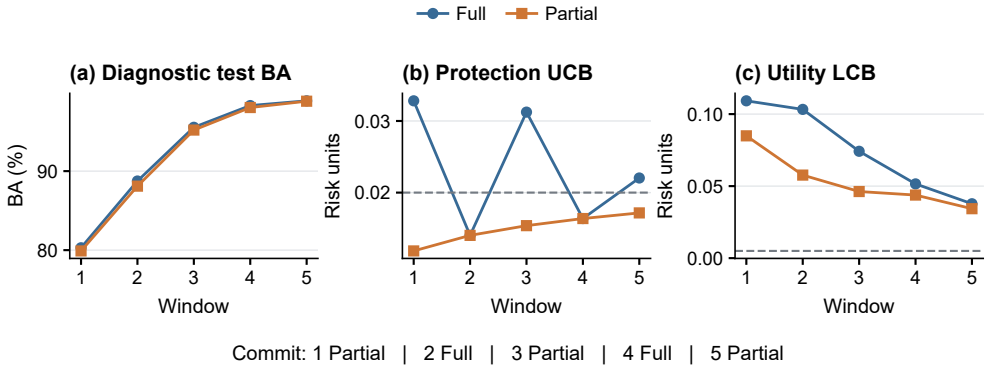


Figure 14: Independent V1 candidate ledger. Candidate test BA is shown separately from the signed protection UCB and utility LCB used for routing. The decision strip identifies the committed candidate.

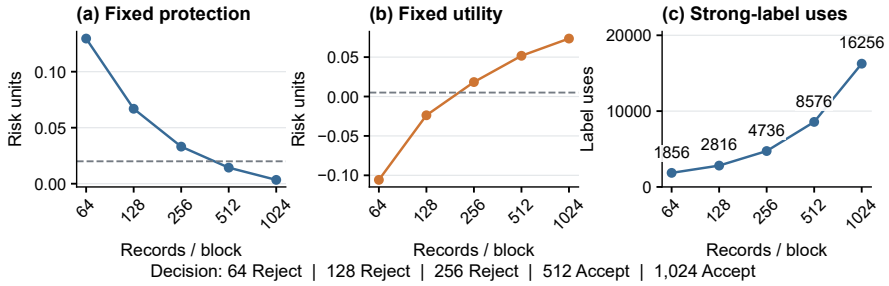


Figure 15: Fixed-candidate certification budget. Protection UCB and utility LCB are evaluated for compatible nested record counts. The dashed lines give the signed acceptance thresholds; the candidate changes from Reject to Accept at 512 records per block.

E.4 From Graphs to Partitions

The reference rule takes connected components of an undirected graph after deduplication and self-loop removal. The structural record has nodes $0, \ldots, 47$, eight true groups $\{6g, \ldots, 6g+5\}$, and a sequential five-edge path within each group, for 40 original edges. True corrupted groups are $g = 0, 1$. At 20% missing edges remove $(6g+2, 6g+3)$ for every group; at 40% also remove $(6g, 6g+1)$. For 20% false edges add $(0,6), (6,12), (12,18), (18,24), (24,30), (30,36), (36,42), (0,42)$; at 40% additionally add $(0,12), (0,18), (0,24), (0,30), (0,36), (6,18), (6,24), (6,30)$. The false-edge denominator is 40 original edges, not the number of possible nonedges. Figure 16 reports the induced partitions. Splitting also affects the independence assumption: applying the reliability theorem requires justified independent-cluster support beyond component counts.

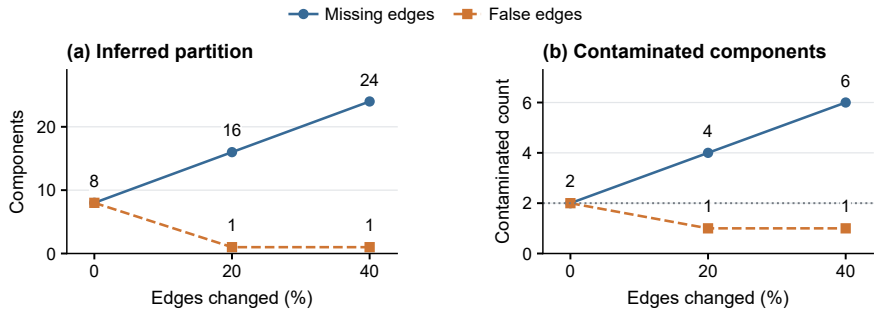


Figure 16: Graph perturbations and induced partitions. Bars show inferred components and contaminated components; annotations give the number of changed edges. The horizontal line marks the configured corruption budget $b = 2$.

F Complete V1 Uncertainty and Paired Comparisons

The core V1 point estimates, component analyses, transfer comparisons, and sensitivity results appear once in Figures 2–4. This appendix reports the complementary uncertainty and paired evidence. All balanced-accuracy values are percentages and differences are percentage points. Intervals are descriptive 2.5th–97.5th percentiles over the recorded roots, with the scope defined in Appendix E.1.

For a pair of methods, compute $D_{d,r} = \mathrm{BA}^{A}_{d,r} - \mathrm{BA}^{B}_{d,r}$ before resampling. A replicate is $\Delta^* = \frac{1}{4}\sum_{d=1}^{4}\frac{1}{60}\sum_{r\in\mathcal{R}^*_d} D_{d,r}$. Positive-pair counts and intervals summarize the paired records conditional on the shared test set; dataset-population uncertainty requires a separate sampling analysis.

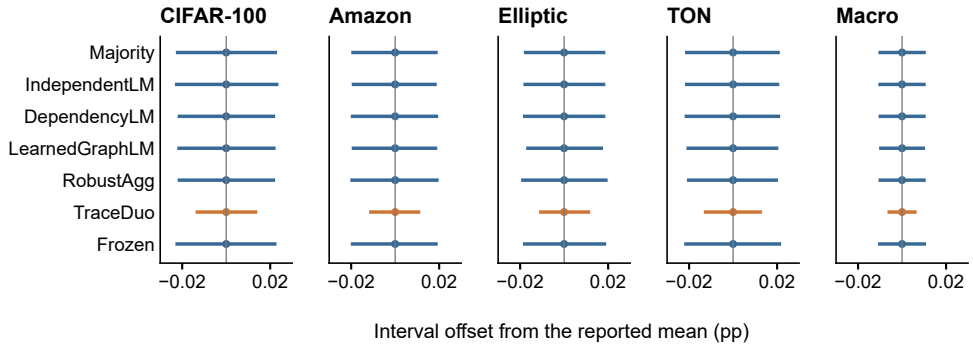


Figure 17: Domain-level and macro uncertainty for the V1 comparison. Intervals summarize variability across 60 recorded roots conditional on the fixed test records. Panels separate the four domain means from the macro aggregate so their narrower ranges remain visible.

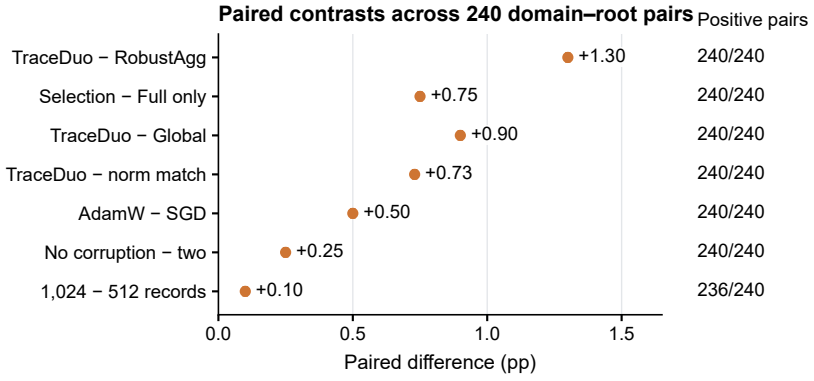


Figure 18: Paired V1 contrasts across 240 domain–root pairs. Points show mean differences and horizontal segments the descriptive whole-root bootstrap intervals. Labels report positive-pair counts; all contrasts preserve pairing across methods and windows.

G Scope and Limitations

Label-model theory. The pseudo-label bound is informative under sufficient independent-cluster support, honest-score separation, and bounded estimation error. Section 5 characterizes the current large-class regime; Appendix A.3 specifies the additional count, sampling, and drift conditions needed to instantiate the reliability bound. Posterior calibration and post-admission error require their own record-level diagnostics, as defined in Appendix D.1.

Mechanism analysis and empirical scope. The local geometry result depends on clean-gradient alignment after the relevant update operations. Direct empirical validation requires matched-state alignment measurements; the V1 evaluation quantifies endpoint contrasts. Component clipping bounds the specified gradients and guarded displacement; source-replacement sensitivity additionally involves changes to label estimation, admission, and other components (Appendix A.6). Root-bootstrap intervals describe recorded-root variability conditional on the shared test set and protocol; population uncertainty and training-run coverage require additional analyses.

Certification scope. The guarantee applies to declared populations under fresh conditional IID sampling, fixed evaluation randomness, and a predeclared fixed-size look. The failure budget is per learning stream; additional streams, per-class targets, or optional stopping require corresponding statistical procedures. The anchor controls the maximum net risk increase relative to the initial model, whereas cumulative positive increments and future distribution shifts require different targets. V2 evaluates known-risk threshold classifiers and its anchor rule; V1 evaluates training trajectories under local-only protection. Extending the V2 guarantee to trained V1 checkpoints requires evaluating the additional endpoints under the V2 protocol.

H Detailed Source Processing and Candidate Construction

The following specification expands the source processing and optimizer operations of Section 4; Algorithm 1 fixes their order.

Sources, clusters, and reliability. A source returns a hard label, a probability vector, or an abstention. Finite, nonnegative vectors with positive mass are normalized and converted to their unique argmax; ties abstain. Provenance keys hash the checkpoint digest, training-data digest, rule-template identifier, and implementation version. Identical keys cast one vote; conflicting outputs under one key abstain, and a key assigned to multiple clusters is an input error. Within a cluster, plurality over distinct non-abstaining keys gives v_{ig}; ties and empty votes give $\perp$. Cluster identifiers hash sorted member keys. A changed member set starts a new reliability history. Graph inputs become connected components after removing duplicate edges and self-loops (Appendix E.4).

Let n_{tg} and c_{tg} count non-abstentions and correct outputs on the reliability pool. Reliability is a scalar per cluster, initialized to $1/2$:

$$\rho_{tg}^{\mathrm{fresh}} = \frac{c_{tg}+1}{n_{tg}+2}, \quad \rho_{tg} = \begin{cases} \rho_{t-1,g}, & n_{tg}=0, \\ \rho_{tg}^{\mathrm{fresh}}, & |\rho_{tg}^{\mathrm{fresh}} - \rho_{t-1,g}| > 0.15, \\ 0.8\rho_{t-1,g} + 0.2\rho_{tg}^{\mathrm{fresh}}, & \text{otherwise.} \end{cases} \tag{35}$$

For an active cluster, assign probability ρ_{tg} to v_{ig} and $(1-\rho_{tg})/(C-1)$ to each other class, floor coordinates at 10^{-4}, and renormalize to obtain q_{ig}. This symmetric-error model is an explicit assumption; the reliability analysis additionally requires score separation and conditional independence across honest clusters.

Aggregation, admission, and deletion. Let $\mathcal{A}_i = \{g : v_{ig} \neq \perp\}$ and $K_i = |\mathcal{A}_i|$. For class c, sort active-cluster $\log q_{ig}(c)$ into $\ell_{i,(1),c} \leq \cdots \leq \ell_{i,(K_i),c}$. With $b = 2$,

$$s_i(c) = \frac{1}{K_i - 2b} \sum_{j=b+1}^{K_i-b} \ell_{i,(j),c}, \quad \widehat{p}_i = \mathrm{softmax}(s_i), \quad \chi_i = \frac{1}{K_i} \sum_{g \in \mathcal{A}_i} \mathbf{1}\{v_{ig} \neq \arg\max_c \widehat{p}_i(c)\}. \tag{36}$$

Admit a record only if $K_i \geq 6$, every deletion leaves at least five active clusters, $\max_c \widehat{p}_i(c) \geq 0.55$, and $\chi_i \leq 0.5$. The confidence condition excludes posterior ties among admitted records. Set $w_i = \max_c \widehat{p}_i(c)(1-\chi_i)$ if admitted, and zero otherwise. An entirely unadmitted window returns Reject. For $g \in \mathcal{A}_i$, remove g, re-sort, trim two values from each end, average with denominator $K_i - 1 - 2b$, and apply softmax to obtain p_i^{-g}. Reliabilities, cluster probabilities, admission, weights, and the batch weight denominator remain fixed; aggregation and softmax normalizers are recomputed. Deletion posteriors are not separately readmitted.

Replay combines frozen soft labels and stored logits:

$$\mathcal{L}_{\text{replay}} = \frac{1}{|J|} \sum_{i \in J} \left[0.5\, \text{CE}(q_i^{\text{buffer}}, f_\theta(x_i)) + \frac{0.5}{C} \|z_\theta(x_i) - z_i^{\text{stored}}\|_2^2 \right],$$
$$g = g^{\text{weak}} + \text{clip}_{0.05}(P_m \nabla \mathcal{L}_{\text{replay}}). \tag{37}$$

Complete optimizer direction and cumulative guard. Each candidate takes 24 steps with weak/replay batch sizes 128/32, starting from θ_t with zero moments. AdamW uses $\eta = 10^{-3}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and decay 0.01:

$$m_s = 0.9 m_{s-1} + 0.1 g_s, \quad v_s = 0.999 v_{s-1} + 0.001 g_s^2, \quad d_s = -\eta P_m \left[\frac{\widehat{m}_s}{\sqrt{\widehat{v}_s} + 10^{-8}} + 0.01 \theta_s \right]. \tag{38}$$

Hats denote step-index bias correction. Two disjoint 128-record initialization training anchor groups define fixed incoming-model CE gradients h_1, h_2. They form a training proxy, separate from certification:

$$\mathcal{K}_m = \{z : P_m z = z,\ h_1^\top z \leq 0,\ h_2^\top z \leq 0\}, \qquad \theta_{s+1} = \theta_t + \text{clip}_{0.05}\left(\Pi_{\mathcal{K}_m}(\theta_s - \theta_t + d_s)\right). \tag{39}$$

Thus $\|\theta_{s+1} - \theta_t\|_2 \leq 0.05$ bounds cumulative candidate displacement. The guard follows preconditioning, decoupled decay, and masking; parameters are written directly, without a second optimizer update. Protected-risk certification uses the separate statistical endpoints below; the gradient proxy constrains training geometry. Appendix B specifies the adapter, replay state, SGD, and projection.

Empty or undersized certification blocks (fewer than 32 records), missing frozen candidates, or an entirely unadmitted window trigger Reject. The assigned error budget is not recycled. Rejection preserves model and replay state; updated reliability estimates may be retained. Routing costs are penalties rather than measured runtimes. Source replacement can change the posterior, other deletion responses, admission, and weights; clipping bounds specified components, with the broader sensitivity conditions in Appendix A.6.

I Additional Theory and Interpretation

Score normalization and assumptions. Write $\varepsilon = 10^{-4}$ and $M = \log((1 + C\varepsilon)/\varepsilon)$. The floored and renormalized probabilities in Section 4 satisfy $\log q_g(c) \in [-M, 0]$. We normalize *coordinates*, $u_g(c) = 1 + \log q_g(c)/M \in [0, 1]$; differences lie in $[-1, 1]$. This common affine transformation preserves class rankings after trimming. It is used only for analysis: the implemented softmax still takes the original trimmed log probabilities, since rescaling would change confidence and admission.

For a fixed record, condition on its input, true label y, history, and active cluster identities. Let $K > 2b$, $h = K - b$, and choose a fixed set H of h honest clusters. For every $c \neq y$, suppose the oracle gaps $X_g(c) = u_g^*(y) - u_g^*(c)$, $g \in H$, are independent under this conditioning, with $\mathbb{E}X_g(c) \geq \gamma$. Assume a simultaneous event E satisfies $\Pr(E^c) \leq \delta_{\text{est}}$ and $\max_{g \in H, c} |u_g(c) - u_g^*(c)| \leq e_s$ on E. The oracle uses current population reliabilities in the same score model; score separation and conditional independence are additional assumptions beyond the declared grouping and symmetric-error model.

Correctness is distinct from calibration and admission. Equation 6 bounds an argmax error, not posterior calibration. For the confidence/conflict admission event $A_{\tau,\xi}$, the relevant quantities are coverage $\pi_{\tau,\xi} = \Pr(A_{\tau,\xi})$ and selective error $r_{\tau,\xi} = \Pr(\widetilde{y} \neq y \mid A_{\tau,\xi})$. If a uniform or validly averaged bound B applies before admission and $\pi_{\tau,\xi} > 0$, only $r_{\tau,\xi} \leq \min\{1, B/\pi_{\tau,\xi}\}$ follows

without extra assumptions. Thus a confidence threshold of 0.55 does not guarantee an error rate at most 0.45. Appendix D.1 defines the record-level measurements required for error–coverage analysis, a separate estimand from final BA.

Why retain deletion responses? For logit Jacobian J_i, soft-target cross-entropy has gradient $J_i^\top(f_\theta(x_i) - q)$. With targets and weights frozen, this proves $G = B + \sum_g R_g$ and identifies each response as

$$R_g = \frac{1}{W} \sum_{i:g\in\mathcal{A}_i} \frac{w_i}{K_i} P_m J_i^\top (p_i^{-g} - \widehat{p}_i). \tag{40}$$

Unlike an arbitrary conserving allocation, these components retain how the actual trimmed rule changes when a source is removed. Appendix A.4 constructs two admitted $K = 8, b = 2$ configurations with exactly equal full posteriors and unequal deletion posteriors under that rule. This establishes additional source-response information; the alignment conditions below characterize when it is useful.

Let $v_C = a_0 B + \sum_g a_g R_g$, $v_G = a_G G$, with clipping factors from Equation 2. For $G \neq 0$, put $Q_G = I - GG^\top/\|G\|^2$. Then

$$d := v_C - v_G = (a_0 - a_G)B + \sum_g (a_g - a_G) R_g, \qquad Q_G v_C = Q_G d. \tag{41}$$

A nonzero right-hand side is necessary and sufficient for a non-collinear update. If it is zero, v_C is a scalar multiple of G; that scalar can be positive, zero, or negative. Unequal clipping factors alone do not imply rotation. In particular, the proportional allocation control remains on the ray of G, whereas cross-record deletion responses can change direction.

A conditional mechanism for local benefit. Let $z = \nabla L_{\mathcal{D}}(\theta)$ for a clean target distribution and $S = z^\top d$. For plain weak-only steps $\Delta_C = -\eta v_C$ and $\Delta_G = -\eta v_G$ on an L-smooth loss,

$$L_{\mathcal{D}}(\theta + \Delta_C) - L_{\mathcal{D}}(\theta + \Delta_G) \leq -\eta S + \frac{L\eta^2}{2}(\|v_C\|^2 + \|v_G\|^2). \tag{42}$$

Thus clipping helps locally when it preferentially suppresses components with $z^\top R_g < 0$ and this gain exceeds losses from clipping useful components and the curvature remainder. For example, if $a_0 = a_G$, only a nonempty set of negatively aligned residuals has $a_g < a_G$, and all other factors equal a_G, then $S > 0$ and a sufficiently small positive step improves the comparison. Equation 40 makes this condition about deletion-induced target changes. A corrupted source need not produce a negatively aligned residual; helpful responses can also be clipped.

For the actual AdamW, replay, and cumulative guard, use the *applied* displacements instead: Equation 8 gives the applied-displacement comparison. Preconditioning and projection need not preserve S or its sign. These are local comparison tools with explicit success and failure conditions; the reported endpoint differences quantify final performance, while evaluating S or $A_{\mathcal{D}}$ requires matched-state measurements.

Component bounds and commitment have different scopes. Clipping bounds the specified components, not a source's total causal effect through posteriors, other residuals, weights, and admission (Section 4). A source-specific bound on $\sup_{D\sim_g D'} \|\Delta(D) - \Delta(D')\|$ additionally requires stability of all quantities affected by source replacement (Appendix A.6).

Here a violation means $D^{\mathrm{loc}}_{t,m,j} > 0.02$ or $D^{\mathrm{anc}}_{t,m,j} > 0.02$ for at least one protected group, or $G^{\mathrm{new}}_{t,m} \leq 0.005$, for the candidate actually committed. Rejected and unselected candidates do not count as erroneous commitments. Appendix A.7 proves conditional coverage of the explicit endpoints and the induction through Reject. The failure budget is per learning stream: $1/24$ for five windows and 0.05 over an unlimited number of windows. It is not $\Pr(\text{violation} \mid \text{acceptance})$.

Probability and risk budgets are different. δ_t controls coverage failure; ϵ_t permits actual risk increase. Local-only protection implies at most $\sum_{t=1}^{T} \epsilon_t$ net increase on fixed populations: five tolerances of 0.02 allow 0.10, even on perfect coverage. V2 directly constrains the initial-anchor relative net increase to 0.02. This does not bound the sum of positive successive increases, nor does

it protect undeclared groups or arbitrary future distributions. An alternative local-only design could enforce $\sum_t \epsilon_t \leq 0.02$ on the same fixed populations; this is a different risk-budget rule and is not the V2 protocol evaluated here.

J Detailed Experimental Protocol

Protocol and experimental scope. We evaluate TraceDuo under protocol `TRACEDUO-REAL-EXPERIMENT-V1` (V1): CIFAR-100 and binary tasks on Amazon, Elliptic, and TON, with 60 evaluation roots (31001–31060) per domain and five windows. Each task contains all classes of its domain, with 1,600 class-balanced test records; five tasks give 8,000 test records per run. Thus the sequence is not a disjoint-class incremental protocol. All method comparisons use the same V1 task definitions and baselines.

The reference resource schedule is 6,000 weak records per window, 256 initialization labels, 128 reliability labels per window, two protected certification blocks and one current-domain block of 512 records each, 256 analysis-only labels per window, and a replay capacity of 512. The declared partition has eight source clusters and corruption budget $b = 2$. Training uses the reference choices in Section 4. Analysis and final-test labels cannot affect candidates or routing.

Certification experiments (V2). The V2 certification experiments separately evaluate the five-endpoint anchor rule in Section 4.1, with nine configurations of 100,000 independent five-window trajectories each, an endpoint ledger, and exact power tables. They use specified 100-class populations and threshold classifiers whose population risks are known (Appendix C). This permits direct measurement of erroneous commitment. V1 accuracy and budget tables use local-only certification. The fixed-anchor guarantee pertains specifically to V2 and its additional endpoints.

One matrix, three temporal summaries. Let $A_{d,r,t,k}$ be task-k balanced accuracy, in percent, after the window-t commitment for domain d and root r. Rejection retains the incoming model. Task accuracy is $100C_d^{-1}\sum_c \mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FN}_c)$. Equal class sets and equal per-class task sizes make the final task average equal to BA on the pooled final test records:

$$\mathrm{FinalBA} = \frac{1}{240}\sum_{d,r}\frac{1}{5}\sum_{k=1}^{5} A_{d,r,5,k}, \tag{43}$$

$$\mathrm{NewBA} = \frac{1}{240}\sum_{d,r}\frac{1}{5}\sum_{t=1}^{5} A_{d,r,t,t}, \tag{44}$$

$$F = \frac{1}{240}\sum_{d,r}\frac{1}{4}\sum_{k=1}^{4}\left[\max_{u\in\{k,\ldots,5\}} A_{d,r,u,k} - A_{d,r,5,k}\right]. \tag{45}$$

The history maximum is taken within each run before averaging and includes the final time, ensuring nonnegative forgetting. New BA averages the post-decision diagonal. Appendix E.2 gives the task–time matrix.

Roots, pairing, and uncertainty. The reference root namespace covers initialization, within-window training and reliability sampling, weak-source outputs and corruption, minibatch ordering and augmentation, candidate randomness, replay, and certification sampling. Task order and the final test set are fixed. Within each domain, compared methods share external randomness. Same-state controls also fix actual replay entries; full trajectories share sampling rules and random streams but may acquire different buffer contents after different commitments.

We summarize the experiments using 20,000 bootstrap replicates (seed 20270926). Each replicate samples 60 complete roots with replacement within each domain, retains method pairing, and averages domains equally. Five windows remain one trajectory, not five independent replications. We report 2.5th–97.5th percentile intervals as descriptive variability across recorded roots, conditional on the shared test set and fixed protocol. Population-level uncertainty and calibrated coverage for independent training runs require corresponding sampling and calibration analyses.

Comparators and controlled factors. The main methods are Majority+DERpp, IndependentLM+DERpp, DependencyLM+DERpp, LearnedGraphLM+DERpp, RobustAgg+DERpp, TraceDuo, and Frozen. Controls cover cumulative components, a 2×3 update-rule/routing comparison, matched posteriors, optimizer and updater transfer, corruption, and certification budget. Every routing-comparison cell trains both candidates and changes only the allowed commitment set. V1 evaluates accuracy, forgetting, and resource contrasts. Matched attack-success, clean-acceptance, and protected-recall-loss aggregates are outside this evaluation's estimands.

Compatibility of temporal metrics. All domain means obey $a_d \geq 4(98.85) - 300 = 95.40\%$ for the reported macro average. For the declared task weights,

$$0 \leq F \leq \frac{5}{4}(100 - \mathrm{FinalBA}); \qquad 0.40 \leq 1.4375 \text{ pp} \quad \text{for TraceDuo.} \tag{46}$$