WHEN DO NOISY-LABEL CORRECTIONS HELP? A COUNTERFACTUAL AUDIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Human annotation is imperfect, and the standard remedy corrects labels before training, yet the corrections it proposes mix repairs with damage that must be separated. Latent-shift noisy-prediction correction (LSNPC) repairs a prediction by moving latent variables rather than the input. Their reliability can be assessed before retraining by scoring each proposed correction. First, we develop an audit with label-free scores inspired by counterfactual properties; label-free refers to scoring and selection, independently of corrector training. We apply LSNPC to training labels and study latent paths that change the decoded label as counterfactual candidates. Second, we find that confidence, one candidate score, is anti-informative where corrections rarely succeed and informative where they usually do. Score informativeness varies with the recovery rate and encoder. In paired held-out diagnostic runs, label-free proximity retains about half of its clean-label-referenced separation over chance. Across our end-to-end study, a rank threshold over the scores reverts the corrections it scores lowest before retraining. At the operating coverage the thresholded stream falls at most 0.1 accuracy points below the original noisy stream in the reported means over corrector seeds. The pattern holds across the three downstream models. At wide coverage it outperforms Co-teaching, the in-training baseline, on the crowd-labelled text streams. Its margin over the original stream is widest on the image streams. At matched coverage, label-free selection performs similarly to random rejection, identifying score discrimination as a target for improvement. This study connects counterfactual properties to selective acceptance of label corrections.

1 INTRODUCTION

Supervised models inherit the errors of human annotators, and label noise is pervasive in practice, since real-world labelling follows procedures and criteria that human annotators apply imperfectly (Frénay & Verleysen, 2014; Song et al., 2023). A recent method in this family, latent-shift noisy-prediction correction (LSNPC), models label noise as a shift in a learned latent representation (Huang et al., 2025). A variational autoencoder (Kingma & Welling, 2014) decodes clean and noisy labels through a shared decoder, and decoding a shifted representation proposes a corrected label. A later work reproduces its empirical effectiveness independently (Hou et al., 2025). LSNPC is post-hoc on a classifier's predictions; we apply it to the training labels themselves, across crowd-labelled text and image streams under both human and injected noise. LSNPC performs well on average, and its proposed corrections mix repairs with damage that must be separated before retraining. Unselective correction, which applies every proposed correction, can repair and damage comparable numbers of examples, even on a crowd-labelled benchmark, which is what makes a per-correction decision before retraining valuable. Before retraining, a practitioner lacks a per-correction judgement of repair or damage, which motivates scoring each proposed correction.

In LSNPC, a correction leaves the input unchanged but moves only the latent variable. An encoder infers a latent $\tilde{\mathbf{z}}$ from the input $\mathbf{x}$ and its noisy label $\tilde{y}$, and the shared decoder g_ϕ reconstructs a label from that latent. A correction replaces $\tilde{\mathbf{z}}$ and decodes again from the same input, and the query stays fixed. When the latent shift changes the decoder's label, we treat it as a counterfactual candidate and measure the properties of that shift (Wachter et al., 2018; Karimi et al., 2023; Mothilal et al., 2020; Ustun et al., 2019). Counterfactual explanations describe an input change that would alter a decision (Wachter et al., 2018). Their properties organise the field, as cost functions (Ustun

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.

et al., 2019), a proximity-diversity trade-off (Mothilal et al., 2020), and a shared vocabulary (Karimi et al., 2023). We keep the properties and change the object they are computed on, from an input that would have been classified differently to a label that has been corrected. Algorithmic recourse then treats such an explanation as a course of action, turning what would have been different into what the affected subject should do (Wachter et al., 2018; Karimi et al., 2021). The quality of a label correction can likewise be understood through counterfactual explanations, which motivates candidate reliability scores. We score each correction on the four counterfactual properties of validity, proximity, noise-robustness, and plausibility, without the clean label (Wachter et al., 2018; Mothilal et al., 2020; Dominguez-Olmedo et al., 2022); Section 4 defines the scores.

Few reviewed counterfactual methods have been evaluated with the users they are meant to inform, and the computational measures used in place of those evaluations are not established substitutes (Keane et al., 2021; Barocas et al., 2020; Karimi et al., 2021). That critique targets advice offered to a person; however, we use the properties as label-free scores (computed without clean labels) on LSNPC's own proposals rather than as recommendations anyone must act on. The frozen backbones used in most settings reduce the fragility of recourse under a changing model (Upadhyay et al., 2021), though not the decay of recourse as the world moves (De Toni et al., 2025). Confidence, the label-free score for validity, is anti-informative where fewer than about a third of the corrections succeed, ranking the least accurate corrections first, and informative above one half. Clean-label-referenced noise-robustness ranks recovered above unrecovered examples; its label-free counterpart is weaker. We use plausibility as a lower-density floor. A setting's recovery rate is its *regime*.

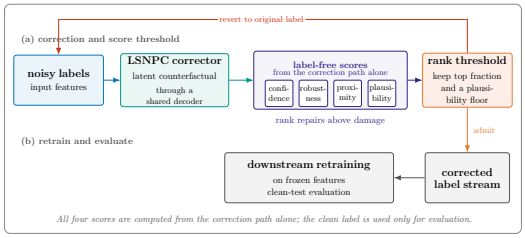


Figure 1: The audit pipeline. Noisy labels enter the LSNPC corrector, which proposes a corrected label for each example. The four scores rank those corrections, and a rank threshold decides which of them are kept for retraining.

We regard this study as an audit, shown in Figure 1. We design thresholds that decide which corrections to keep when the downstream model is retrained, and retraining on the thresholded stream decides whether the rule outperforms the alternatives.

Contributions. First, we score proposed label corrections using counterfactual properties before retraining. Label-free refers to scoring and selection; trusted supervision for corrector fitting and clean-label diagnostics are separate. Second, clean-label diagnostics favour ranked selection over matched random rejection on the repair/damage subset. Third, label-free selection improves downstream accuracy over the original stream in most reported means and outperforms Co-teaching (Han et al., 2018) on the crowd-labelled text streams at wide coverage. Its similar performance to matched random rejection motivates more discriminative label-free scores.

2 Related Work

Learning with noisy labels supplies the correctors we audit. Confident learning estimates the joint distribution of noisy and clean labels and identifies the label errors it infers (Northcutt et al., 2021). Active label cleaning ranks instances by estimated label correctness and spends a human re-annotation budget where it is most valuable (Bernhardt et al., 2022). In-training methods such as Co-Regularization and self-adaptive training instead revise or reweight labels during training (Wei et al., 2020; Huang et al., 2020). Work that judges the correction itself does so inside training. PLS scores the pseudo-label it proposes with a pseudo-loss correlated with its correctness (Albert et al., 2023). TRACE assesses the observed label and the pseudo-target separately, because down-weighting one up-weights the other (Fan & Li, 2026), and Detect-and-Correct repairs the loss of the noisy subset alone (Grinberg et al., 2025). These methods operate on the label set or the training loss as a whole, whereas we decide per correction. Confidence, the validity score, is the maximum softmax probability of the trustworthy-prediction literature (Hendrycks & Gimpel, 2017; Geifman & El-Yaniv, 2017), which ranks inputs, not corrections.

We borrow three reference points from counterfactual explanation (Wachter et al., 2018; Ustun et al., 2019; Karimi et al., 2023), namely the minimal change, the properties that judge it, and the recourse view that turns an explanation into action. That literature has been criticised for assuming a feature change maps to a real-world action (Barocas et al., 2020), and for recourse that a model update invalidates (Upadhyay et al., 2021). A further line of criticism covers recourse that decays as the world changes (De Toni et al., 2025) and recourse whose value depends on when it is followed (Upadhyay et al., 2025). The closest work applies counterfactuals to mistakes. NDCC detects noisy instances and corrects them with counterfactual learning (Qi & Chelmis, 2024), and conceptual counterfactuals explain the mistakes of an already trained model (Abid et al., 2022). Neither method decides per correction, whereas the audit decides each one.

Learning with noisy labels produces the corrections. Work on reliable prediction judges a model's own signals. Our audit connects the two by deciding which corrected labels to keep. We study per-correction scoring before retraining, separating clean-label diagnostics from label-free selection.

3 METHODOLOGY

Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ be an input and $y \in \{1, \dots, C\}$ its true label, with clean dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ and noisy observation $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{y}_i)\}$. We study data-level repair, correcting the noisy labels of $\tilde{\mathcal{D}}$ before retraining rather than adjusting a trained classifier's predictions post hoc.

3.1 NOISE AS LATENT DISPLACEMENT

LSNPC posits that label noise appears as a geometric displacement in a learned latent space. Each input has a clean latent $\mathbf{z} \in \mathcal{Z}$ that generates the correct label y. The observed label $\tilde{y}$ is generated from a perturbed latent $\tilde{\mathbf{z}}$ that a heavy-tailed Student-t transition connects to $\mathbf{z}$. The generative process is defined as follows.

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), & \tilde{\mathbf{z}} \mid \mathbf{z} &\sim t_{\nu_0}\big(\mu_\psi(\mathbf{z}), \operatorname{diag}(\sigma^2_\psi(\mathbf{z}))\big), \\ y \mid \mathbf{z}, \mathbf{x} &\sim \operatorname{Cat}\big(g_\phi(\mathbf{x}, \mathbf{z})\big), & \tilde{y} \mid \tilde{\mathbf{z}}, \mathbf{x} &\sim \operatorname{Cat}\big(g_\phi(\mathbf{x}, \tilde{\mathbf{z}})\big). \end{aligned} \tag{1}$$

where t_{ν_0} is the Student-t with ν_0 degrees of freedom. Its location and scale come from a shift network; the displacement is learned rather than assumed. The shared decoder $g_\phi : \mathcal{X} \times \mathcal{Z} \mapsto [0, 1]^C$ produces a distribution over the C classes. LSNPC targets noisy multilabel classification. We model multiclass labels, replacing a Bernoulli likelihood by a Categorical one.

Following the variational autoencoder (VAE) training scheme, an encoder maps $(\mathbf{x}, \tilde{y})$ to the noisy latent $\tilde{\mathbf{z}}$, and a *shared* decoder g_ϕ decodes a label from either $\mathbf{z}$ or $\tilde{\mathbf{z}}$. Sharing the decoder is the key design choice. It couples the clean- and noisy-latent decoding paths; whether this yields the intended alignment is evaluated empirically. A decoder fit to the noisy stream alone could decode each latent directly to the label it carries and reproduce the noise. A single g_ϕ, by contrast, must assign one label to the $\mathbf{z}$ and the $\tilde{\mathbf{z}}$ of the same class. Correction attempts to invert that displacement by replacing $\tilde{\mathbf{z}}$ with a clean latent $\mathbf{z}$ drawn from the correction-time posterior $q(\mathbf{z} \mid \tilde{\mathbf{z}}, \mathbf{x})$ and decoding through the same g_ϕ. Correction therefore needs no separate repair module, since the encoder and shared decoder attempt to invert the displacement the representation has learned.

3.2 TRAINING OBJECTIVE

The unsupervised learning loss is a cross-entropy that reconstructs the noisy label at the corrected latent together with a Kullback–Leibler (KL) divergence keeping the proposal near the prior,

$$\mathcal{L} = \textstyle\sum_{(\mathbf{x}, \tilde{y}) \in \tilde{\mathcal{D}}} \operatorname{CE}\left(g_\phi(\mathbf{x}, \mathbf{z}), \tilde{y}\right) + \beta \operatorname{KL}\left(q(\tilde{\mathbf{z}}, \mathbf{z} \mid \mathbf{x}, \tilde{y}) \,||\, p(\tilde{\mathbf{z}}, \mathbf{z})\right), \tag{2}$$

using a β-weighted VAE objective (Higgins et al., 2017), with $\beta \leq 1$ for this task. We factorise the proposal family as $q(\mathbf{z}, \tilde{\mathbf{z}} \mid \mathbf{x}, \tilde{y}) = q(\mathbf{z} \mid \tilde{\mathbf{z}}, \mathbf{x})\, q(\tilde{\mathbf{z}} \mid \mathbf{x}, \tilde{y})$, with

$$q(\tilde{\mathbf{z}} \mid \mathbf{x}, \tilde{y}) = t_\nu\big(\mu_\theta(\mathbf{x}, \tilde{y}), \operatorname{diag}(\sigma^2_\theta(\mathbf{x}, \tilde{y}))\big), \quad q(\mathbf{z} \mid \tilde{\mathbf{z}}, \mathbf{x}) = \mathcal{N}\big(\mu_\kappa(\tilde{\mathbf{z}}, \mathbf{x}), \operatorname{diag}(\sigma^2_\kappa(\tilde{\mathbf{z}}, \mathbf{x}))\big).$$

In the posterior, the heavy tails are the inference counterpart of the transition, and the encoder can still place mass on the true-label region even when the label is corrupted. The clean posterior is

Table 1: The four label-free scores, one per counterfactual property.

Score	Counterfactual property	Definition	Orientation
Confidence p^*	validity	max class probability of the corrected class, Equation (3)	high
Noise-robustness r	robustness	label preservation under K perturbations, Equation (4)	high
Proximity $1 - f^*$	proximity	minimal fraction of the correction reaching the class, Equation (5)	high
Plausibility π	plausibility	log-density of the corrected latent, Equation (6)	high (lower floor)

Gaussian, a local recentering of $\tilde{\mathbf{z}}$ toward the clean mode. Up to a constant, the loss of Equation (2) is the negative evidence lower bound, taken in expectation over the sampled latents. The divergence has no analytic form under a Student-t posterior, and is evaluated through a density-ratio surrogate (Schulman, 2020). Given the space limit, we defer the implementation details to Section A.

We use LSNPC's unsupervised objective and its trusted-set variant (Huang et al., 2025). Given a trusted clean set of known-correct examples, a semi-supervised variant adds the same two objective terms over those examples. Section E.1 states the variant in full and reports the configuration trained with it. In this work the trusted set can also be assembled from the audit itself, the counterfactual scores selecting the trustworthy corrected labels (Section H.1).

3.3 Correction as a Forward Pass

Correction uses feed-forward inference with Monte Carlo latent sampling. Upon sampling $\tilde{\mathbf{z}} \sim q(\tilde{\mathbf{z}} \mid \mathbf{x}, \tilde{y})$ and $\mathbf{z} \sim q(\mathbf{z} \mid \tilde{\mathbf{z}}, \mathbf{x})$, it returns $\mathcal{C}(\mathbf{x}, \tilde{y}) = \arg\max \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\tilde{\mathbf{z}},\mathbf{x}),\, \tilde{\mathbf{z}}\sim q(\tilde{\mathbf{z}}|\mathbf{x},\tilde{y})}\big[g_\phi(\mathbf{x}, \mathbf{z})\big]$, approximated by Monte Carlo. Each correction is therefore a candidate relabelling, and the remaining choice is which to trust. Reverting a correction restores the original annotation, which makes the audit a selective label-correction problem, analogous to selective prediction. A rule scores each corrected example and admits or reverts it, trading kept corrections against avoided damage. The counterfactual properties provide those scores.

4 Counterfactual Auditing with the Four Properties

A counterfactual explanation of a model prediction is the alteration of the input that would have produced a desired alternative prediction (Wachter et al., 2018). In our case, for a target label y^* different from the current prediction, a counterfactual is a nearby point $\mathbf{x}'$ at which the model predicts y^*. A *latent* counterfactual is the same construction in the learned latent space of Section 3. It is the alteration of the latent variable that would have made the shared decoder produce the desired alternative label. For the same target label y^*, it is a nearby latent $\mathbf{z}'$ that the shared decoder maps to y^*, whereas the input stays unchanged. The LSNPC correction $\mathcal{C}$ draws a noisy latent, proposes a corrected latent, and decodes the corrected label through the shared decoder. A path that changes the decoder's label is a latent counterfactual candidate relative to its starting prediction. An unchanged decoded label is a reconstruction; the proposed target need not be the true label.

The audit treats each such correction as a candidate proposal and attaches to it a per-query reliability score. Let $\mathbf{x}$ be a fixed query with noisy label $\tilde{y}$. The encoder of Section 3 draws a noisy latent $\tilde{\mathbf{z}} \sim q(\tilde{\mathbf{z}} \mid \mathbf{x}, \tilde{y})$, and the correction $\mathcal{C}$ proposes a corrected latent $\mathbf{z}$. The shared decoder decodes the corrected label $\mathcal{C}(\mathbf{x}, \tilde{y})$ from it. The true label y evaluates label-free scores; separate diagnostic variants use it as their target reference. The correction succeeds on $\mathbf{x}$ exactly when $\mathcal{C}(\mathbf{x}, \tilde{y}) = y$, and we write $v(\mathbf{x}) = \mathbb{1}\{\mathcal{C}(\mathbf{x}, \tilde{y}) = y\}$ for the per-query success indicator. The protocol introduces four scores, one per counterfactual property (Table 1). We write $s(\mathbf{x})$ for $s(\mathbf{x}, \tilde{y}; \theta, \phi, \kappa, \omega)$, conditional on the fitted model and Monte Carlo randomness ω. These scores require no clean label at scoring time; their diagnostic variants replace $\mathcal{C}(\mathbf{x}, \tilde{y})$ in Equations (4) and (5) by y.

Validity (confidence). LSNPC's confidence in the corrected class is the maximum class probability after Monte Carlo averaging under the correction posterior.

$$p^*(\mathbf{x}) \;=\; \max_c \mathbb{E}_q[g_\phi(\mathbf{x}, \mathbf{z})]_c \tag{3}$$

Robustness (noise-robustness). Robustness measures whether the correction survives perturbation of the noisy latent. With K perturbations $\tilde{\mathbf{z}}^{(k)} = \tilde{\mathbf{z}} + \boldsymbol{\epsilon}^{(k)}$ drawn as $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ on the noisy

latent, we re-run the correction from each perturbed draw. Writing $\mathbf{z}^{(k)}$ for the corrected latent obtained from $\tilde{\mathbf{z}}^{(k)}$, the score is the label-preservation rate under re-correction.

$$r(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\left\{\arg\max g_\phi(\mathbf{x}, \mathbf{z}^{(k)}) = \mathcal{C}(\mathbf{x}, \tilde{y})\right\}, \tag{4}$$

Proximity (minimality). Minimality is the smallest fraction of the applied latent correction that already decodes to the corrected label. Denoting the correction by $\Delta(\mathbf{x}) = \mathbf{z} - \tilde{\mathbf{z}}$ from the noisy latent to the corrected latent, the minimal fraction is

$$f^*(\mathbf{x}) \;=\; \min\left\{f \in \mathcal{F} : \arg\max\; g_\phi\big(\mathbf{x},\; \tilde{\mathbf{z}} + f\,\Delta(\mathbf{x})\big) = \mathcal{C}(\mathbf{x}, \tilde{y})\right\}, \tag{5}$$

where $\mathcal{F} = \{0.25, 0.5, 0.75, 1\}$, and the score is $1 - f^*(\mathbf{x})$. A correction is minimal when the corrected label is reached with only a fraction of the full correction. If no grid point matches, set $f^* = 1$. Scores lie in $\{0, 0.25, 0.5, 0.75\}$. The grid excludes $f = 0$, even for unchanged decoded labels, and measures progress along this path rather than a globally minimal distance.

Plausibility (LSNPC posterior). Plausibility measures the corrected latent's density under the noisy encoder posterior. The score is the log-density of the corrected latent under the encoder posterior $q(\tilde{\mathbf{z}} \mid \mathbf{x}, \tilde{y})$, the Student-$t$ noisy posterior of Section 3.

$$\pi(\mathbf{x}) \;=\; \log q_\theta^{\text{noisy}}\big(\mathbf{z} \mid \mathbf{x}, \tilde{y}\big) = \log t_\nu\big(\mathbf{z};\; \mu_\theta(\mathbf{x}, \tilde{y}), \mathrm{diag}(\sigma_\theta^2(\mathbf{x}, \tilde{y}))\big), \tag{6}$$

The audit operates on the mislabelled set (mis-set), the queries whose label is wrong. Those with $v(\mathbf{x}) = 1$ are recovered, the rest unrecovered. The mis-set AUROC (area under the receiver operating characteristic curve) of a score s is the fraction of pairs in which a recovered query scores above an unrecovered one. Ties count as a half. Above chance is informative and below it anti-informative. We report rank statistics, because the audit asks how a score orders recovered (WR) against unrecovered (WW) mislabelled examples; damage (RW) is measured separately. The definition and the scale-invariance argument are in Section B. Each score is judged against two reference points, the recovery ceiling $\bar{v}$, the fraction of mislabelled queries the correction recovers, and the never-correct fallback, which accepts no query. The properties are adapted from the counterfactual literature (Wachter et al., 2018; Mothilal et al., 2020; Karimi et al., 2023; Guidotti, 2024), the fallback comes from selective classification (Geifman & El-Yaniv, 2017), and the ceiling is our own reference point.

The same scores can also be inserted into the semi-supervised loop of Section 3.2, admitting corrected examples into an enlarged semi-supervised pool. That training-time transfer uses clean-label validation in Section H; its measured gains remain below the specified five-point criterion. Section B.1 states the admission rule and the plausibility floor in full.

5 Empirical Findings

The datasets fall into two types, image and text, and two noise-source families; Table 2 gives sizes and noise sources for every benchmark, and Section C the full per-setting inventory. The labels are either *real* human label streams released with the dataset, or clean labels corrupted with *injected* symmetric (Sym) or instance-dependent (Idn) noise (Zhang & Sabuncu, 2018; Xia et al., 2020). Each injected stream is generated once per benchmark, with each example's noisy label fixed across clean-set sizes. The clean-set size is itself a further budget, varied in Section E.1 over nested prefixes of the clean reference set. Clean reference labels support evaluation and diagnostic scoring; the label-free scores of Section 4 and end-to-end selection do not use the scored example's clean label. Real-noise rates are the measured fractions of flipped labels in the stream, and injected settings are applied at nominal rates of 20% and 40%, as in (Zhang & Sabuncu, 2018; Xia et al., 2020).

Table 2: Dataset summary

Benchmark	Classes	Size	Noise source
Text			
NoisyAG-News (best/med/worst) (Huang et al., 2024)	4	50k each	real
AG-News (Zhang et al., 2015)	4	120k	Sym/Idn
SST-2 (Socher et al., 2013)	2	67k	Sym/Idn
Medical abstracts (Schopf et al., 2022)	5	14k	Sym/Idn
Image			
CIFAR-10N (aggre/worse) (Wei et al., 2022)	10	50k	real
CIFAR-100N (fine) (Wei et al., 2022)	100	50k	real
dopanim (Herde et al., 2024)	15	10.5k	real
EuroSAT (Helber et al., 2019)	10	27k	Sym/Idn

Each setting trains the LSNPC corrector $\mathcal{C}$ on a noisy label stream, produces its per-query candidate scores on an evaluation slice, and is audited against these evaluation-only clean labels. Throughout, a corrector is one trained instance of LSNPC. Correctors use MPNet for text (Song et al., 2020) and ResNet-50 (He et al., 2016) or Swin (Liu et al., 2021) for image; the end-to-end dopanim model operates on pixels. Downstream quality is top-1 accuracy on the clean test slice over three fixed downstream models, with Co-teaching (Han et al., 2018) as the in-training baseline. Table A1 in Section A gives the full benchmark table with clean-label provenance, covering splits, calibration, noise generators, head architectures, the corrector recipe and the seed and bootstrap conventions. Section B.1 distinguishes diagnostic rank cuts, label-free selection and supervised pool calibration.

The appendix gives score definitions and inventories (Sections B and C), label transitions (Section D), calibration and encoder comparisons (Section E), full retraining results (Section F), and extension and robustness studies (Sections G and H). Additionally, the code is available at `https://github.com/iclr2027anonymous-beep/paper-code.git`.

5.1 Finding 1. The informative signal flips with regime

Score informativeness varies with the recovery regime and encoder. The regime is the recovery rate of a setting, the fraction of initially mislabelled examples whose proposed label is correct. Confidence is anti-informative where corrections rarely succeed and informative where they work, noise-robustness stays useful throughout, and proximity follows the frozen image encoder (Figure 2 and Table C2). The two label-referenced scores are measured here against evaluation-only clean labels. Section 5.3 measures their label-free variants against the correction's own label; Table F11 compares the two.

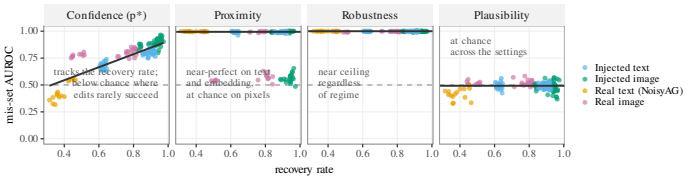


Figure 2: Score informativeness against the setting's recovery rate, one panel per score. Points are settings, coloured by data type and noise source (sky blue for injected text, green for injected image, orange for real text, pink for real image). A grey line is a fit where the score tracks the recovery rate, and the panel median otherwise. The dashed line is chance.

Each correction is scored on the counterfactual properties of Section 4, and mis-set discrimination measures recovered (WR) against unrecovered (WW) examples. The metric is mis-set AUROC, defined in Section B. Whether a score tracks the recovery rate, and in which direction, is a property of the regime. Section D partitions each correction by its label transition, counting repairs against damage: **RR** (both labels right), **WR** (wrong to right, a repair), **RW** (right to wrong, damage) and **WW** (both labels wrong). The transition net, WR minus RW, is positive where repairs outnumber damage.

The confidence family behaves unlike the other three scores. Where the recovery rate is low it ranks recovered examples below chance. Where the recovery rate is high it becomes informative. Its mis-set AUROC falls to 0.23 where the corrector rarely succeeds and rises to 0.75 where it mostly succeeds over the text settings, and from 0.29 to 0.90 over the image ones. It is below chance at the bottom of each range, so a threshold keeping its most confident corrections would favour unrecovered examples within the mislabelled subset. Across the crowd-labelled text streams it rises from 0.33 to 0.56 as the recovery rate rises, and the injected image settings reach 0.88 at the low rate. On the text settings the corrector repairs far more often, and confidence ranks repairs well above chance. Across the 25 settings of that inventory, its mis-set AUROC tracks the recovery rate at $r = +0.98$ over the text settings and $+0.97$ over the image ones.

This association is observed across the tested settings of this corrector. Confidence reports that the corrector is certain, not that it is correct. At low recovery rates, high confidence can attach to unrecovered examples. The clean-label-referenced proximity and noise-robustness scores show a different pattern. On the crowd-labelled text streams proximity's mis-set AUROC runs from 0.85 to 0.88, and across the frozen image settings from 0.75 to 0.86. It is highest where the corrector repairs most of what it touches. Across the frozen Swin settings the same score falls in the 0.50–0.52

Table 3: Margin in accuracy points over the 95th percentile of random rejection at equal coverage, with one standard error over the settings. Bold marks a column's largest margin. Table E8 gives the same margins in full, with the per-setting interquartile ranges.

Threshold score	Text, 10%	Text, 50%	Image, 10%	Image, 50%
Confidence	$+0.23 \pm 0.06$	$+0.92 \pm 0.17$	$+0.84 \pm 0.25$	$+2.77 \pm 0.85$
Proximity	$+0.17 \pm 0.07$	$+1.91 \pm 0.24$	$+1.29 \pm 0.40$	$+4.98 \pm 1.45$
Noise-robustness	$\mathbf{+0.31 \pm 0.07}$	$+1.57 \pm 0.34$	$\mathbf{+1.50 \pm 0.45}$	$+5.25 \pm 1.71$
Plausibility	-0.54 ± 0.16	-0.52 ± 0.08	-0.45 ± 0.10	-0.37 ± 0.09
Proximity $\wedge$ robustness	$+0.20 \pm 0.08$	$\mathbf{+1.96 \pm 0.26}$	$\mathbf{+1.50 \pm 0.45}$	$\mathbf{+5.45 \pm 1.67}$

range; its informativeness follows the image encoder rather than the data type. Under the settings of Section C, proximity separates recovered from unrecovered examples on real text and frozen ResNet-50 image settings.

Clean-label-referenced noise-robustness is the stronger recovery indicator. It ranks recovered above unrecovered examples on every setting of the inventory, from 0.94 to 0.99. On the injected settings with the highest recovery rate, where the corrector mostly succeeds, it changes little between the two regimes and maintains recovery discrimination.

Plausibility's mis-set AUROC ranges from 0.35 to 0.68 across the 25 settings. We use it as a lower-density floor in Section 5.2; this design does not imply reliable damage detection by plausibility alone. On the injected settings it sits close to chance. On the real crowd-labelled text streams, where the human label stream is itself noisy, it falls below chance, because the corrected latent lies in a region of lower density under the encoder.

The unselective accuracy changes at $K{=}32$ are consistent with this pattern. The gain grows with the recoverable damage, with the injected rate on the injected settings and with severity on the real streams, and it shrinks where the corrector rarely succeeds. On the injected settings where the corrector mostly succeeds it sits at $+8.6$ to $+11.0$ accuracy points at the low rate and $+20.4$ to $+26.5$ at the high one on text. Image is higher at both rates, $+14.8$ to $+16.0$ and then $+32.5$ to $+33.5$. On the low-recovery settings of both noise types it stays between -0.1 and $+6.0$ accuracy points, and the noisier real streams follow the injected pattern. The largest mean loss among retained settings is 0.1 accuracy points. The transition net agrees in sign with the accuracy delta on every setting of that inventory (Section D), providing a label-level comparison alongside downstream accuracy.

5.2 Finding 2. Clean-label diagnostics favour ranked selection

This diagnostic study ranks the clean-label-identified WR/RW subset, using clean-label-referenced proximity and noise-robustness. It retains a top fraction and reverts the remainder; WW edits are excluded here, but included end to end. We call the retained fraction its coverage (low or operating 10%, midpoint 50%, wide 90%); a threshold clears the control when it exceeds both the median and the 95th percentile of matched random rejection. Over proximity and noise-robustness, such a threshold outperforms random rejection of a matched fraction. It does so on the eligible settings of the label-noise benchmark study, at the coverage point the audit reports for each data type. A case is eligible where its coverage retains at least 30 corrections, the specified minimum count. The second comparison is against applying every proposed correction. The end-to-end study instead ranks all edits using label-free scores (Section 5.3).

These rank cuts are computed on the diagnostic population, without a separate threshold-calibration split. A setting enters the analysis only where its mislabelled set supports the audit. The power floor is 1,000 mislabelled queries on the primary settings and 200 elsewhere, over the evaluation slices of Section A. Table 3 reports archived margins over random rejection, with the full set of settings and its per-setting interquartile ranges in Table E8 and the tests and accounting behind both in Section E. First, proximity, noise-robustness and their conjunction exceed the control's 95th percentile in the mean of every column of Table 3. Taken one score at a time, no score falls below 70% of its eligible settings except plausibility, which succeeds in about one case in five. That result motivates testing plausibility as a floor alongside stronger ranking scores. Its average case falls below the control's own 95th percentile, at -0.4 accuracy points; its occasional exceedances are local rather than systematic.

Second, the ordering of the threshold scores holds by paired tests rather than assumption. On text, proximity and noise-robustness agree in the share of cases they clear at the midpoint, and noise-robustness shows the higher median AUROC, 0.95 against 0.78 over the text settings. On image the two are complementary at the low coverage, and their conjunction attains the largest mean margin at 50% coverage. Paired within case, confidence and proximity each clear the control strictly more often than plausibility, by 13 and 15 discordant settings at the midpoint against none the other way. The exact two-sided sign tests give $p = 2.4 \times 10^{-4}$ and 6.1×10^{-5}. On image the same contrast holds at 11 against none, while proximity and noise-robustness do not separate. The coverage columns include different sets of settings because eligibility depends on the retained count. Third, the threshold's benefit tracks the regime (Section E). Where unselective correction is largely constructive, the threshold avoids damage while retaining most of the gain from it, whereas where that correction is destructive, the threshold limits the loss. Whether the damage removed compensates for the repairs forfeited is determined by the damage share of the correction pool. Clean labels identify the recovery regime for this diagnostic comparison. Its favourable ranking results motivate the separate label-free test below.

5.3 Finding 3. Label-free selection preserves downstream gains

The ordering of the downstream models is preserved, with two entries below the original stream across unselective and thresholded correction (Figure 3 and Tables F9 and F10).

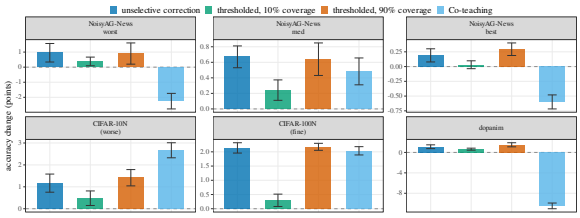


Figure 3: End-to-end accuracy change over the original noisy stream, one subplot per setting with a corrector. Bars give unselective correction, the thresholded stream at the specified 10% operating coverage and at 90% coverage, along with the Co-teaching baseline. Error bars are ± 1 standard error over the five corrector seeds.

The end-to-end study corrects the training label stream, retrains the downstream model from scratch, and evaluates on the clean test split. The MLP comparison has six configuration families, with three thresholded coverage levels; the clean-label reference is unavailable for dopanim. The entries of Table F9 are the original noisy labels (a), unselective correction (b) and thresholded correction under the label-free rule of Section B.1 (c). The fourth is a clean-label reference trained on clean training labels (d), evaluated on the same held-out test split. The remaining two are Co-teaching (Han et al., 2018) as an in-training small-loss baseline (e) and random rejection of a matched fraction (f). On the image streams the thresholded entry also applies its plausibility floor at target coverage. The three fixed downstream models are a three-layer multi-layer perceptron (MLP) on frozen encoder features, a linear head on the same features, and the MLP with the asymmetric loss in place of cross-entropy. Within each head, the relabelling configurations share architecture, initialisation and schedule; Co-teaching changes the training procedure. The corrector is fixed before selection; Section A states what the retained records document about its trusted set. Means are over five corrector seeds, each averaged over five downstream seeds; Sections A and F give the recipe and all three heads.

For the cross-entropy MLP, the margin over the original labels is 1.4–2.2 accuracy points on the image streams at wide coverage. It is positive in the mean on the crowd-labelled text streams, and widest on the real image streams. Under matched measured pool noise, close to half of one real image stream's corrected candidates are damaging, against about one in seven on the other. One of the two entries below the original stream is unselective correction with the asymmetric-loss head on the middle text stream, where a single corrector seed diverges from the other four (Table F10). The other is the linear head at the operating coverage on the dopanim stream, 0.1 accuracy points below and inside the reported standard errors.

For the cross-entropy MLP, thresholded correction agrees with size-matched random rejection to within half an accuracy point at 50% coverage and closely tracks unselective correction at 90%. Across all three heads, the operating configuration falls at most 0.1 accuracy points below the original stream in the reported means. Score saturation is consistent with these small selection margins. On the candidate edits the specified score is saturated. It takes only four distinct values on the text

candidates, with 98% of them at the same top value; the threshold therefore sets how many corrections are admitted rather than which ones. The diagnostic advantage in Table 3 uses clean-label-referenced scores and a WR/RW candidate subset. Both differences limit direct comparison with end-to-end selection. In the separate paired diagnostic runs of Table F11, label-free noise-robustness averages 0.55, near chance. Proximity retains a signal there, at 0.66. Against Co-teaching the comparison splits by data type, and within the image streams by how much damage the corrector does. At 90% coverage, the thresholded stream outperforms it on all three crowd-labelled text streams, by 3.2 accuracy points on the noisiest of them. On the image stream whose corrector damages close to half of its edits the two agree to within an accuracy point. On the image stream whose corrector damages about one in seven of its edits, and is the more accurate of the two, the baseline exceeds every corrected configuration. It moves ahead by 1.2 accuracy points over the best thresholded configuration and 1.5 over unselective correction.

A corrector that is accurate and rarely damaging leaves unselective correction most of the gain the audit could add, and it gains $+1.1$ accuracy points against the $+1.4$ of the best thresholded configuration. The best thresholded configuration remains 4.5 accuracy points below the clean-label reference. An in-training baseline that never consults the corrector's proposals closes more of that gap than the audit does. This empirical reference contextualises the remaining gap; it is not an upper bound on attainable test accuracy. It is thirteen accuracy points at the best thresholded configuration on the noisiest text stream. It is eight accuracy points on a real image setting, against 1.2 accuracy points on the least noisy text stream. These gaps jointly reflect the label stream and downstream learning, and do not isolate proposal quality from selection quality.

The dopanim corrector is a pixel-level Swin at 224 px over the full annotated pool, a separate corrector family from the frozen-feature settings. That contrast aligns with the finding that stronger backbones help this corrector more (Huang et al., 2025). The stream carries no clean training labels; no clean-label reference is available for it and that entry of the table stays blank.

The gain is broadly consistent across downstream model families and losses. Training on corrected labels improves mean clean-test accuracy in most entries under all three downstream models, and at wide coverage their ordering on the noisy stream is reproduced. On one real image stream the linear head starts four accuracy points below the MLP on the noisy stream. Unselective correction lifts it four times as much as the MLP, of which the threshold at the operating coverage recovers about a fifth.

6 Conclusion

We audit data-level label corrections using properties of their latent paths and select proposals before retraining. Score informativeness varies with recovery rate and encoder. Confidence can rank unrecovered examples highest; clean-label diagnostics identify stronger recovery signals in proximity and noise-robustness. In paired held-out diagnostic runs, label-free proximity retains part of that signal, while label-free noise-robustness averages near chance. Retrained on the thresholded stream, downstream models match or exceed the original stream in the reported means except for one operating-coverage entry, which is 0.1 accuracy points lower. At wide coverage, the MLP outperforms Co-teaching on the crowd-labelled text streams. Matched random rejection performs similarly at 50% coverage; improving label-free ranking is therefore a concrete next step.

AI USAGE DISCLOSURE

In this work, generative AI tools were used to assist with manuscript preparation, literature search, and experimental code development.

For manuscript preparation, LLMs were primarily used for grammar checking, improving clarity and readability, refining wording, and assisting with LaTeX formatting, including equations and tables.

LLMs were also used to assist in identifying relevant literature and existing methods. All references included in the manuscript were independently examined and verified by the authors.

During experiments, LLMs assisted with code implementation, debugging, and efficiency optimization. All AI-assisted code was reviewed and tested by the authors.

The final research methodology, experimental design, and scientific conclusions were independently determined and validated by the authors. The authors take full responsibility for the content of this paper, including all AI-assisted material.

REFERENCES

Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 66–88. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/abid22a.html.

Paul Albert, Eric Arazo, Tarun Krishna, Noel E. O'Connor, and Kevin McGuinness. Is your noise correction noisy? PLS: Robustness to label noise with two stage detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 118–127, 2023. arXiv:2210.04578.

Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 80–89, 2020. doi: 10.1145/3351095.3372830.

Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P. Lungren, Aditya Nori, Ben Glocker, Javier Alvarez-Valle, and Ozan Oktay. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13:1161, 2022. doi: 10.1038/s41467-022-28818-3.

Giovanni De Toni, Stefano Teso, Bruno Lepri, and Andrea Passerini. Time can invalidate algorithmic recourse. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 89–107, 2025. doi: 10.1145/3715275.3732008.

Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 5324–5342, 2022. URL https://proceedings.mlr.press/v162/dominguez-olmedo22a.html. arXiv:2112.11313.

Wenxiao Fan and Kan Li. Stop replacing noise with noise: Two-source reliability assessment for label correction and sample reweighting in label-noise learning, 2026.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017. arXiv:1705.08500.

Yuval Grinberg, Nimrod Harel, Jacob Goldberger, and Ofir Lindenbaum. Detect and correct: A selective noise correction method for learning with noisy labels. *Computer Science Research Notes*, 3501(2025):173–181, 2025. doi: 10.24132/CSRN.2025-19. WSCG 2025 Proceedings.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38:2770–2824, 2024. doi: 10.1007/s10618-022-00831-6.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018. arXiv:1804.06872.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242. Sentinel-2 land-use tiles, ten classes; the audit uses the RGB variant. Doubles as the injected-noise image control.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations (ICLR)*, 2017. arXiv:1610.02136.

Marek Herde, Denis Huseljic, Lukas Rauch, and Bernhard Sick. dopanim: A dataset of doppelganger animals with noisy annotations from multiple humans. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2407.20950.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Senyu Hou, Yuru Ren, Gaoxia Jiang, and Wenjian Wang. Noisy multi-label learning through co-occurrence-aware diffusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.52202/085713-4857. URL https://openreview.net/forum?id=zft0zTOFkN.

Hongfei Huang, Tingting Liang, Xixi Sun, Zikang Jin, and Yuyu Yin. NoisyAG-News: A benchmark for addressing instance-dependent noise in text classification, 2024.

Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19365–19376, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/e0ab531ec312161511493b002f9be2ee-Abstract.html.

Weipeng Huang, Qin Li, Yang Xiao, Cheng Qiao, Tie Cai, Junwei Liang, Neil J. Hurley, and Guangyuan Piao. Correcting noisy multilabel predictions: Modeling label noise through latent space shifts, 2025.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362, 2021. doi: 10.1145/3442188.3445899.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):95:1–95:29, 2023. doi: 10.1145/3527848.

Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4466–4474, 2021. doi: 10.24963/ijcai.2021/609. Survey Track.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6114.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 607–617, 2020. doi: 10.1145/3351095.3372850. arXiv:1905.07697.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. doi: 10.1613/jair.1.12125.

Wenting Qi and Charalampos Chelmis. Noisy label detection and counterfactual correction. *IEEE Transactions on Artificial Intelligence*, 5(2):763–775, 2024. doi: 10.1109/TAI.2023.3271963.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91, 2021.

Tim Schopf, Daniel Braun, and Florian Matthes. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR)*, pp. 6–15, 2022. doi: 10.1145/3582768.3582795. Introduces the five-class medical abstracts corpus.

John Schulman. Approximating KL divergence. Blog post, http://joschu.net/blog/kl-approx.html, 2020. The k3 estimator, exp(log p - log q) - (log p - log q) - 1.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013. Introduces the Stanford Sentiment Treebank.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. arXiv:2102.13620.

Sohini Upadhyay, Himabindu Lakkaraju, and Krzysztof Z. Gajos. Counterfactual explanations may not be the best algorithmic recourse approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI)*, pp. 446–462, 2025. doi: 10.1145/3708359.3712095.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 10–19, 2019. doi: 10.1145/3287560.3287566. arXiv:1809.06514.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018. arXiv:1711.00399.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018. doi: 10.18653/v1/W18-5446.

Table A1: Datasets, grouped by data type and noise source. The table covers NoisyAG-News (Huang et al., 2024) (over AG-News (Zhang et al., 2015)), SST-2 (Socher et al., 2013) (the 67k binary split follows Wang et al., 2018), medical abstracts (Schopf et al., 2022), CIFAR-10N/100N (Wei et al., 2022), dopanim (Herde et al., 2024) and EuroSAT (Helber et al., 2019).

Type	Dataset	Classes	Size	Noise	Clean label (eval. only)
Text	NoisyAG-News (best / med / worst)	4	50k each	human annotators	curated AG-News label
	AG-News	4	120k	injected SYM / IDN	label before corruption
	SST-2	2	67k	injected SYM / IDN	label before corruption
	Medical-abstracts	5	14k	injected SYM / IDN	label before corruption
Image	CIFAR-10N (aggre / worse)	10	50k	human annotators	curated CIFAR-10 label
	CIFAR-100N (fine)	100	50k	human annotators	curated CIFAR-100 label
	dopanim (Noisy Doppelganger Animals)	15	10.5k	human annotators	verified species label
	EuroSAT	10	27k	injected SYM / IDN	label before corruption

Note: dopanim releases human annotations on the training split only. Its official validation and test splits carry the verified-clean species label with no noisy stream; it supports the end-to-end experiment (correct the training stream, retrain, evaluate on clean test) but not the per-query reliability protocol.
On NoisyAG-News the Best and Worst streams were assembled by selecting human annotations against the curated AG-News label; the audit's evaluation label is not independent of how those streams were built.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13723–13732, 2020. doi: 10.1109/CVPR42600.2020.01374.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. Introduces the CIFAR-10N and CIFAR-100N human annotation collections.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/5607fe8879e4fd269e88387e8cb30b7e-Abstract.html.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pp. 649–657, 2015. Uses the AG's News corpus to build a classification benchmark.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018. URL https://arxiv.org/abs/1805.07836.

A EXPERIMENTAL SETUP DETAILS

This appendix collects the setup detail summarised in Section 5.

The evaluation slices are 2,000 examples for NoisyAG-News and 3,000 for the CIFAR benchmarks. The calibration studies use 2,500-example validation slices. The diagnostic rank cuts and end-to-end selection use the distinct populations and rules in Section B.1. A setting enters the audit only where its evaluation slice clears the power floor of Section 5. Every image output is computed with $K{=}32$ perturbation draws. All aggregates are means over the corrector seeds of that study, with paired bootstrap confidence intervals (2,000 resamples) over seeds for every reported delta. Wilcoxon tests appear only as aggregate tests over the named settings, never per configuration. The downstream models of the end-to-end experiment are three fixed families on the frozen features. A three-layer multi-layer perceptron (MLP) head (256 units per layer, ReLU, dropout 0.1) is trained with cross-entropy, a single-layer linear head follows, and the same MLP uses the asymmetric loss of Ridnik et al. (2021) in place of cross-entropy: a per-class sigmoid focal loss with negative focusing $\gamma_- = 1$, positive focusing $\gamma_+ = 0.25$ and probability margin 0.05 (Section 5.3 fixes the shared schedule). The in-training baseline is Co-teaching with the cross-entropy MLP. The real streams are taken as released. They are the NoisyAG-News crowd variants (Huang et al., 2024) (best, medium and worst),

the CIFAR-10N aggre and worse annotator aggregates and CIFAR-100N fine (Wei et al., 2022), and the single-annotator stream of dopanim (Herde et al., 2024). The injected streams are generated once over the full label array; each example's noisy label is fixed across clean-set sizes. Symmetric noise flips each label uniformly to any other class with the nominal probability. Instance-dependent noise gives each example its own flip probability, set by the ratio of its distance to its own class centroid to its distance to the nearest competing centroid. An example that sits nearer a competing centroid, or further from its own, is the likeliest to be mislabelled. The flipped-to class is drawn over the other classes with probabilities weighted by centroid proximity, and the per-example probabilities are calibrated so their average reproduces the nominal rate. Table A1 lists the noise types used, and Table C2 the rates. The calibration and clean-budget studies use the following training recipe; the end-to-end provenance caveat is stated below. The clean-set variant of Section 3.2 runs as a two-phase schedule within each epoch. Phase A is an unsupervised pass over the noisy examples, and Phase B a clean-conditioned pass over a 2,000-example clean prefix disjoint from the evaluation and validation carves, in which the clean-conditioned posterior is used with probability $\eta = 0.1$. The prefix is nested across sizes, and larger clean sets strictly contain smaller ones. The generative transition of Equation (1) is Student-t with $\nu_0 = 2$ degrees of freedom and the approximate posterior with $\nu = 5$. The Kullback-Leibler (KL) divergence weight is $\beta = 0.5$ and both bounds are importance-weighted over five sampled latents. Parameters are trained with AdamW at learning rate 5×10^{-4} and $\epsilon = 10^{-6}$, under bf16 mixed precision. Text trains for 15 epochs and image for 10. The batch size is 512 over the cached MPNet embeddings (Song et al., 2020) and 256 over the frozen ResNet-50 features (He et al., 2016). Pixel-based correctors use 64. Text corrections correct the 768-dimensional MPNet embedding space directly. ResNet-50 corrections correct its 2,048-dimensional feature space. The frozen Swin (Liu et al., 2021) and trainable ResNet-50 correctors, which operate on pixels, correct a 64-dimensional latent. The trainable ResNet-50 corrector is the same pipeline with the backbone unfrozen. The audit, threshold, calibration and clean-budget runs each use five seeds, the end-to-end retraining five corrector seeds with five downstream seeds each, and the statistical robustness check ten. The paired diagnostic cohort in Table F11 is separate: six archived corrector seeds for each text/CIFAR setting and three for dopanim, totalling 33 setting-seed pairs. Each pair compares score references on the same examples; the aggregate is not the five-seed retraining cohort. The experimental archive records pipeline revision 43cc682. The retained end-to-end result files do not record the corrector's trusted-set size. The campaign run script defaults to 2,000 clean examples, but we could not confirm this setting from the archived runs. We therefore restrict the label-free claim to scoring and selection, which use no clean label of any scored example.

Section C and Section D give the per-setting inventory and the transition figures. Section E reports the calibration runs, the feature-space boundary and the clean-budget ladder, Section F the end-to-end tables, Section G the repair ceiling of the single-proposal rule, and Section H the training-time studies.

B Score Quality Metrics

The main text interprets the audit's metrics in words. For reference, the mis-set AUROC (area under the receiver operating characteristic curve) of a per-query score s (one of the four properties of Section 4) on a query set $\mathcal{Q}$, judged against correction success, is the rank comparison

$$\mathrm{AUROC}(s) \;=\; \frac{1}{|\mathcal{S}_1|\,|\mathcal{S}_0|} \sum_{i \in \mathcal{S}_1} \sum_{j \in \mathcal{S}_0} \Big[\mathbb{1}\{s(x_i) > s(x_j)\} + \tfrac{1}{2}\,\mathbb{1}\{s(x_i) = s(x_j)\} \Big], \tag{7}$$

where, over the mislabelled set $\mathcal{S}$ (Section 4), $\mathcal{S}_1 = \{x \in \mathcal{S} : v(x) = 1\}$ are the correctly recovered queries and $\mathcal{S}_0 = \mathcal{S} \setminus \mathcal{S}_1$ the unrecovered ones. The recovery ceiling $\bar{v}$ and the never-correct fallback against which every score is measured are defined in Section 4. We also report the ranking metric that a selective rule delivers. Both metrics are rank statistics, invariant to any strictly increasing rescaling. This matters because the four scores are not commensurable. Confidence is a probability, proximity a fraction, noise-robustness a rate over K draws, and plausibility a log-density. No score is calibrated before the audit and none becomes a probability when it is thresholded (Section 5.2). A metric that read the scale of a score would assess a calibration the audit never imposes. The two metrics divide the work. Equation (7) asks whether the ranking separates recovered from unrecovered queries at all. AP (average precision) asks how much of the minority class reaches the top of the same ranking, which is the region a threshold rule uses. For a query set $\mathcal{Q}$ with positive class $\mathcal{P} \subseteq \mathcal{Q}$, sort the

queries by decreasing score and let r_k be the position of the k-th positive query. Then

$$\mathrm{AP}(s;\mathcal{P}) \;=\; \frac{1}{|\mathcal{P}|}\sum_{k=1}^{|\mathcal{P}|}\mathrm{prec}_s(r_k),$$

where $\mathrm{prec}_s(r)$ is the fraction of the r highest-scored queries that lie in $\mathcal{P}$ and higher scores indicate the positive class. The paper reports AP as the companion to each AUROC. Within a setting the score audit of Section 4 reports it on the repair channel ($\mathcal{P}$ the recovered examples among the mislabelled, matching the convention of Equation (7)). It reports the same statistic on the damage channel ($\mathcal{P}$ the broken examples among the originally correct). The regime analysis of Section 5.1 reports it one-vs-rest. Each transition group is taken in turn as $\mathcal{P}$ against the other three, and the figures show the per-group AP (and AUROC), averaged over seeds. In every case the value is an average precision for one chosen positive class, not a class mean.

B.1 Score-Thresholded Selection of Corrected Labels

The studies use three distinct selection protocols. In the diagnostic comparison of Section 5.2, clean labels identify the WR/RW candidate subset and define the target of proximity and noise-robustness. Their diagnostic variants replace the emitted-label reference in Equations (4) and (5) by y; confidence and posterior density are unchanged. Rank cuts are computed on this diagnostic population, without a separate validation-calibration split. The conjunction takes the minimum of the within-subset percentile ranks of proximity and noise-robustness. The requested count is the rounded coverage times the WR/RW count, capped by the eligible count. A minimum of 30 retained examples sets eligibility for reporting. These choices require clean labels and serve as a diagnostic comparison, rather than a label-free admission procedure. We do not report a diagnostic conjunction with an additional plausibility floor: the retained analysis code applied the floor mask in its pre-sorting order, and the per-query artifacts needed to recompute it are unavailable. Rules without a floor are unaffected, and the end-to-end rule below has a separate implementation.

End-to-end selection in Section 5.3 instead considers every proposed edit, including wrong-to-wrong edits. Let $\mathcal{E} = \{i : \hat{c}_i \neq \ell_i\}$, where $\hat{c}_i = \mathcal{C}(\mathbf{x}_i, \ell_i)$ and $\ell_i = \tilde{y}_i$. The fixed ranking score is label-free proximity on text and label-free noise-robustness on image. At target coverage $c \in \{0.1, 0.5, 0.9\}$, the rule takes the first $\lceil c|\mathcal{E}|\rceil$ edits in descending score order, preserving input order for ties. Thus a tied score threshold alone does not define the retained count. Image selection also requires $\pi_i \geq \pi_0$. A separate held-out set of proposed edits is $\mathcal{E}_{\mathcal{V}} = \{j \in \mathcal{V} : \mathcal{C}(\mathbf{x}_j, \ell_j) \neq \ell_j\}$. If it contains at least 30 examples, π_0 is its $\alpha = 0.1$ plausibility quantile; otherwise it is the same quantile over $\mathcal{E}$. No clean labels enter either quantile. There is no backfilling after this floor, so realised coverage can be below the target. Text selection has no plausibility floor. Random rejection samples the same realised number of edits uniformly from $\mathcal{E}$.

The plausibility floor is a one-sided exclusion rule: low posterior density may remove an edit, while high density does not certify correctness. The complete audit requires the grid evaluations and K re-corrections in addition to baseline inference.

The training-time pool study in Section H uses a different, supervised calibration: it chooses the proximity threshold on clean-labelled validation examples at precision at least 95%, maximising retained coverage subject to that constraint. When a trusted set $\mathcal{D}_{\text{clean}}$ is supplied, accepted corrected labels form $\mathcal{P}$ and the subsequent semi-supervised pass uses $\mathcal{D}_{\text{clean}} \cup \mathcal{P}$. This optional retraining step is separate from Algorithm B1. Coverage levels, minimum counts and floors are specified protocol choices; we do not claim an externally registered protocol.

C Score Inventory

This appendix reports the score inventory for every benchmark of Table A1, the text benchmarks at each noise type and rate and the image benchmarks at the primary configuration of the study. The main text interprets the patterns. The four scored properties are defined in the main text, in Table 1. The threshold rule runs on those four properties alone.

Algorithm B1: Label-free selection for end-to-end retraining

Input : noisy training examples $(\mathbf{x}_i, \ell_i)_{i=1}^n$; fixed corrector $\mathcal{C}$; modality; target coverage c; held-out noisy examples $\mathcal{V}$; floor quantile $\alpha = 0.1$

Output : relabelled training set $(\mathbf{x}_i, \ell_i^*)_{i=1}^n$

$\ell_i^* \leftarrow \ell_i$ for all i
Compute $\hat{c}_i = \mathcal{C}(\mathbf{x}_i, \ell_i)$ using Monte Carlo inference
$\mathcal{E} \leftarrow \{i : \hat{c}_i \neq \ell_i\}$
if $\mathcal{E} = \varnothing$ **then**
  return $(\mathbf{x}_i, \ell_i^*)_{i=1}^n$
Compute label-free s_i on $\mathcal{E}$: proximity for text, noise-robustness for image
$\mathcal{A} \leftarrow$ first $\lceil c|\mathcal{E}| \rceil$ indices in a stable descending sort by s_i
if *modality is image* **then**
  Compute π_i on $\mathcal{E}$ and on separately inferred held-out edits $\mathcal{E}_\mathcal{V}$
  $\mathcal{H} \leftarrow \mathcal{E}_\mathcal{V}$ if $|\mathcal{E}_\mathcal{V}| \geq 30$, otherwise $\mathcal{H} \leftarrow \mathcal{E}$
  $\pi_0 \leftarrow \text{Quantile}_\alpha(\{\pi_j : j \in \mathcal{H}\})$
  $\mathcal{A} \leftarrow \{i \in \mathcal{A} : \pi_i \geq \pi_0\}$
for $i \in \mathcal{A}$ **do**
  $\ell_i^* \leftarrow \hat{c}_i$
return $(\mathbf{x}_i, \ell_i^*)_{i=1}^n$

D Transition-Group Discrimination

This appendix reports the one-vs-rest discrimination of each score by transition group, the summary condensed in the main text. On AG-News at 0.4 symmetric, robustness ranks unchanged-correct examples far above the other two groups, and places both of those far below chance; it functions as a damage detector. Proximity ranks unchanged-correct examples highest, and its separation of unchanged-wrong from damaged examples is the narrower of its two signals, and plausibility stays close to the prevalence rate throughout. Under the real image noise of CIFAR-10N worse the same pattern is reproduced, and the panels below give the per-group values.

D.1 The Per-Setting Figures for Every Dataset and Noise Combination

Figure D1 shows the two representative settings of the main text, AG-News 0.4 symmetric and CIFAR-10N worse (Figure D1a and Figure D1b). The figures below repeat the same four-score facet layout for the remaining 15 dataset and noise combinations of the study. These are the NoisyAG-News severity ladder, AG-News and SST-2 injected noise at 0.2 and 0.4, Medical-abstracts injected noise, the remaining image settings, and the EuroSAT settings. The three main-text patterns hold in every one of them. Robustness separates the damage group, proximity is the unchanged-correct signal, confidence carries the recovery-dependent signal, and plausibility tracks prevalence.

E Calibration Settings and the Feature-Space Boundary

The calibration study closes with 415 completed runs. They are the γ series (C-a) over nine text and five image settings, the encoder comparison across the clean budgets (C-b), and the dopanim resolution probe (C-c). Three results carry the section.

Table E3 and Figure E17 vary γ alone, at the main configuration's clean budget and with $\eta = 0.1$. On the real-crowd text stream γ re-weights the edit stream rather than changing the corrector. As it runs from 0.1 to 1.0, one seed's edits move from 450 repairs and 147 damaging edits at $+6.2$ accuracy points to 327 repairs and 183 damaging edits at $+4.1$ accuracy points. The recovery rate follows the same direction, from 0.47 to 0.37, and so does confidence, from 0.51 to 0.42. The image settings are unmoved. CIFAR-10N worse holds between $+26.0$ and $+26.3$ accuracy points at every value of γ, its edit counts changing by less than one percent. Across the nine text and five image settings, γ re-weights an edit stream the corrector already produces, and the representation decides whether it exists.

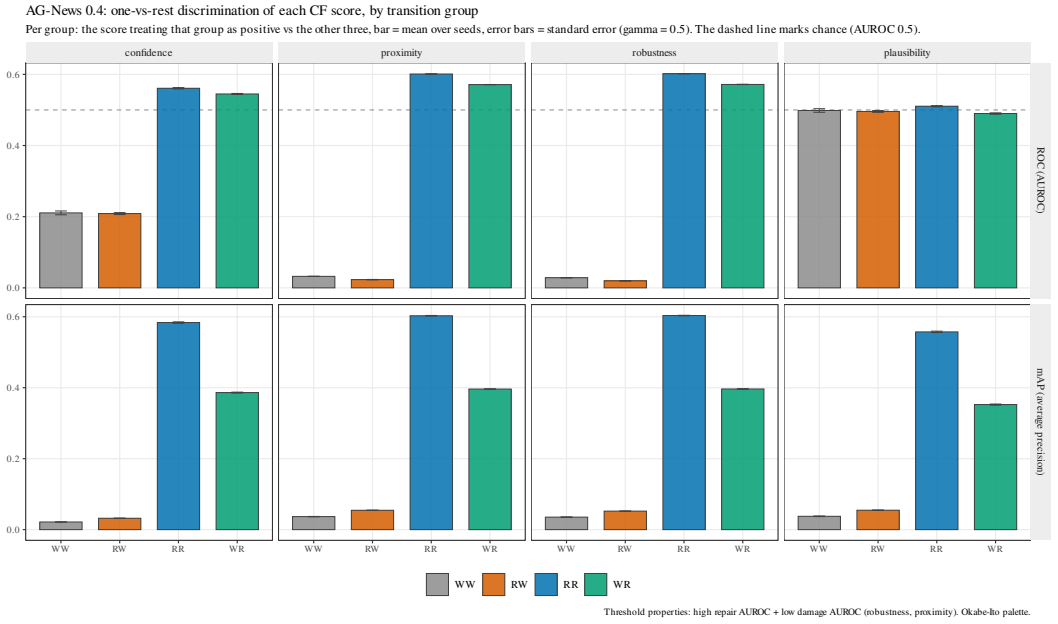
Table C2: Score inventory of the regime-flip study across the five seeds of each configuration ($K = 32$); plausibility uses the posterior of Equation (6). Columns give the recovery rate, the per-score mis-set AUROC and the accuracy delta of unselective correction, mean ± SD over the seeds. The corrector is the main configuration: frozen backbones, fifteen training epochs on text and ten on the image streams, clean-set size 2,000 and $\eta = 0.1$. "Native" entries have no injection. Bold marks the strongest score per entry (ties bolded), and underlining marks confidence below chance. Image entries use the frozen ResNet-50 corrector, except dopanim, which is the 224 px probe of Section E at the low edit rate (γ=0.1) and has three archived seeds.

					Mis-set AUROC				
Dataset	Noise	Type	Rate	Recovery	conf.	min.	rob.	plaus.	Δacc
AG-News	injected	IDN	0.2	78.4	0.69	0.78	**0.96**	0.47	+10.6 ±0.6
AG-News	injected	SYM	0.2	76.5	0.71	0.79	**0.97**	0.44	+11.0 ±0.5
AG-News	injected	IDN	0.3	77.1	0.74	0.76	**0.96**	0.51	+18.6 ±1.0
AG-News	injected	SYM	0.3	78.0	0.71	0.76	**0.96**	0.43	+18.8 ±0.7
AG-News	injected	IDN	0.4	74.8	0.72	0.75	**0.95**	0.49	+26.5 ±0.9
AG-News	injected	SYM	0.4	75.6	0.75	0.76	**0.95**	0.50	+26.3 ±0.8
CIFAR-10N aggre	real	native	–	49.5	0.57	0.81	**0.97**	0.49	+0.9 ±0.4
CIFAR-10N worse	real	native	–	60.4	0.68	0.79	**0.95**	0.51	+22.1 ±0.5
CIFAR-100N fine	real	native	–	25.9	0.59	0.86	**0.97**	0.52	+6.0 ±0.5
EuroSAT	injected	SYM	0.2	89.4	0.88	0.80	**0.95**	0.48	+16.0 ±0.2
EuroSAT	injected	IDN	0.2	85.9	0.85	0.75	**0.97**	0.46	+14.8 ±0.4
EuroSAT	injected	SYM	0.4	90.1	0.90	0.85	**0.97**	0.48	+33.5 ±0.6
EuroSAT	injected	IDN	0.4	86.1	0.84	0.81	**0.95**	0.50	+32.5 ±0.5
dopanim	real	native	–	14.1	0.29	0.54	**0.99**	0.68	+1.9 ±0.5
Medical-abstracts	injected	SYM	0.2	16.8	0.35	0.82	**0.96**	0.47	+2.0 ±0.5
Medical-abstracts	injected	IDN	0.2	19.5	0.35	0.83	**0.96**	0.49	+2.9 ±0.6
Medical-abstracts	injected	SYM	0.4	12.0	0.23	0.79	**0.95**	0.47	+4.1 ±0.6
Medical-abstracts	injected	IDN	0.4	11.8	0.25	0.81	**0.95**	0.47	+3.7 ±0.5
NoisyAG-News best	real	native	–	25.7	0.33	0.87	**0.97**	0.35	-0.1 ±0.2
NoisyAG-News med	real	native	–	42.4	0.45	0.88	**0.97**	0.37	+5.5 ±0.4
NoisyAG-News worst	real	native	–	53.4	0.56	0.85	**0.94**	0.44	+17.2 ±0.9
SST-2	injected	IDN	0.2	59.1	0.55	0.74	**0.95**	0.45	+8.6 ±0.4
SST-2	injected	IDN	0.4	56.1	0.53	0.74	**0.95**	0.46	+20.4 ±0.7
SST-2	injected	SYM	0.2	60.4	0.52	0.72	**0.95**	0.45	+8.9 ±0.3
SST-2	injected	SYM	0.4	56.9	0.51	0.74	**0.94**	0.45	+20.5 ±1.1

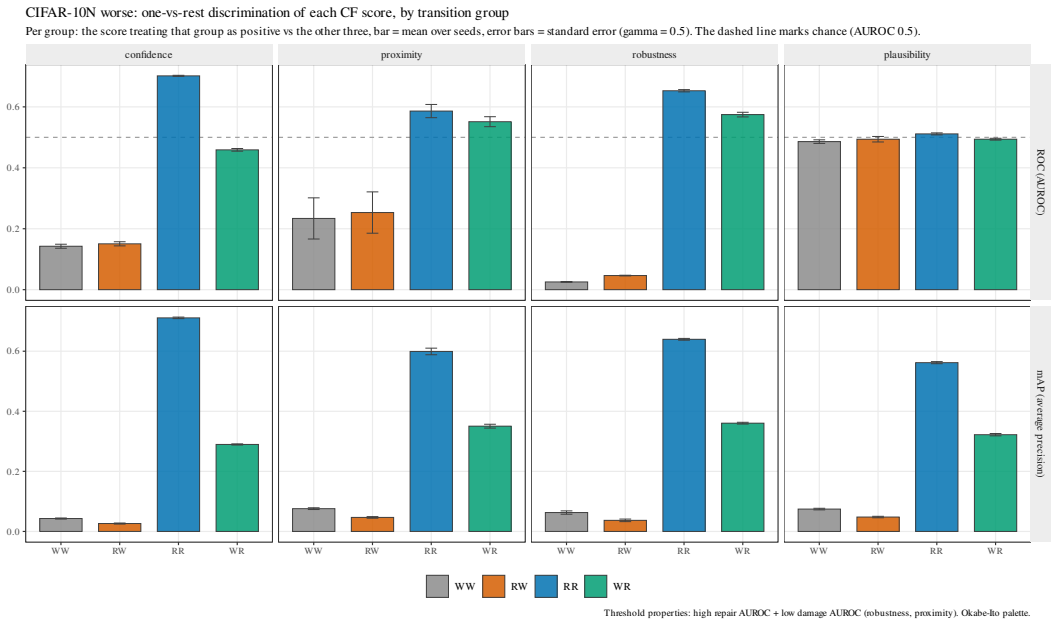
Table E3: The γ (edit rate) series under the main configuration (η=0, clean budget 0, frozen encoders): the recovery rate and accuracy gain (Δacc, points) over the noisy stream, 5 seeds, mean. Bold marks the endpoints discussed in the text.

	NoisyAG-News med		CIFAR-10N worse		CIFAR-10N aggre		CIFAR-100N fine	
γ	Recov.	Δacc	Recov.	Δacc	Recov.	Δacc	Recov.	Δacc
0.10	**0.47**	+6.2	**0.79**	+26.1	**0.74**	-0.5	**0.45**	+3.0
0.25	0.43	+5.6	0.79	+26.3	0.74	-0.5	0.45	+3.0
0.50	0.38	+4.3	0.79	+26.3	0.74	-0.5	0.44	+2.8
0.75	0.38	+4.1	0.78	+26.0	0.75	-0.5	0.45	+3.0
1.00	**0.37**	+4.1	**0.79**	+26.1	**0.75**	-0.4	**0.45**	+3.0

Table E4 compares the two frozen encoders at fixed settings. The Swin corrector produces the larger edit stream and the higher recovery rate on both datasets. At the smallest budget it reaches a recovery rate of 0.87 and +31.7 accuracy points on CIFAR-10N worse against 0.82 and +28.1 for ResNet-50, and on CIFAR-100N fine the gap is wider, 0.54 and +10.0 against 0.46 and +3.4. Enlarging the clean reference raises both encoders on CIFAR-100N fine, from +3.4 to +5.3 accuracy points for ResNet-50, while CIFAR-10N worse stays within an accuracy point of +29. Changing only the encoder therefore moves the gain several times further than the budget does, which places the boundary in the features the corrector uses rather than in the calibration data. Together with the



(a) AG-News 0.4 symmetric.



(b) CIFAR-10N worse.

Figure D1: One-vs-rest discrimination by transition group. Bars are the mean over seeds and error bars the standard error.

threshold study this fixes the image operating point as a frozen-encoder corrector behind the rank threshold.

The 224 px probe settles the specified comparison. At $\gamma = 0.1$ the main configuration's frozen Swin corrector emits an edit stream in every seed, covering between 5% and 35% of the 1,000-example evaluation slice. Left unthresholded, that stream repairs more than it damages. The label-level net is $+66$ examples on average, positive on all five seeds, and clean-test accuracy gains $+6.6$ points on average, also positive on every seed. Clean-label-referenced noise-robustness separates recovered from unrecovered mislabelled examples, reaching AUROC 0.988 to 0.997 on every seed, while proximity on its own sits near chance between 0.487 and 0.585. The separation this appendix reports

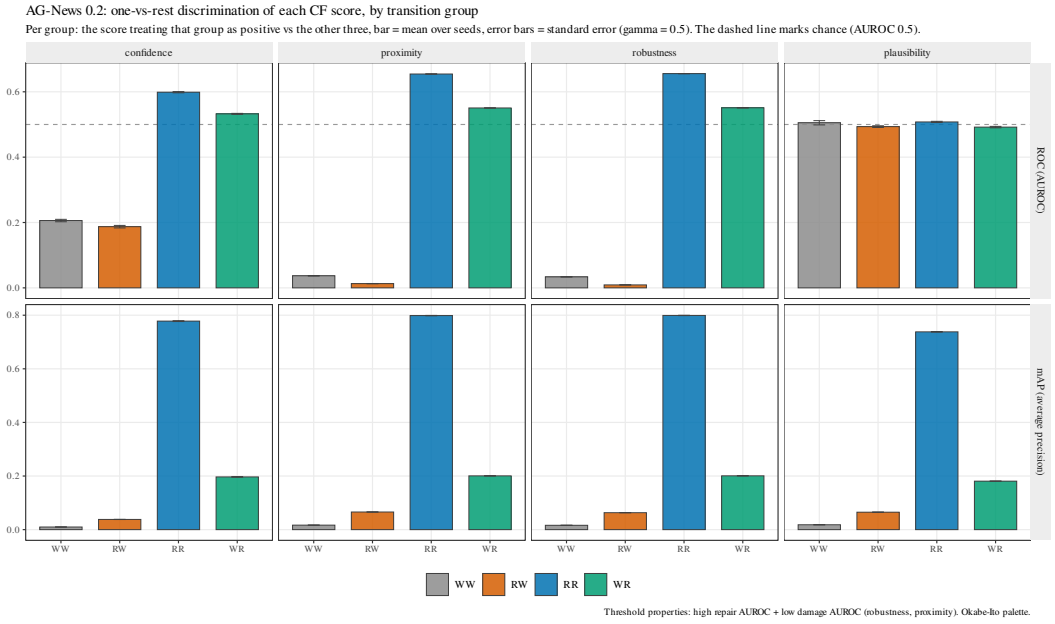


Figure D2: AG-News 0.2.

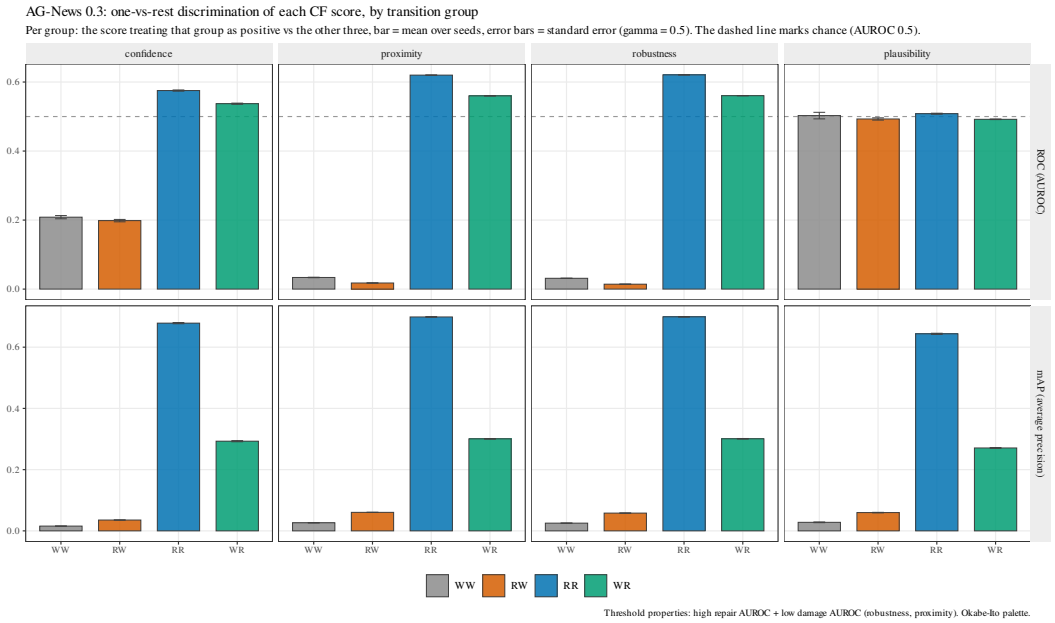


Figure D3: AG-News 0.3.

therefore rests on noise-robustness and on the conjunction that the threshold study builds from the two.

The encoder carries the same boundary on the real-noise image stream. Run in the frozen-feature mode of the same corrector, at 224 pixels and trained unsupervised ($\eta = 0$, no clean reference, five epochs), the ResNet-50 encoder damages far more than it repairs: across the five seeds the corrected stream loses between 6.9 and 8.6 accuracy points against the noisy stream, with all five seeds negative, its recovery rate on the mislabelled set lies between 0.497 and 0.537 as a recovery fraction, and the transition net runs from -257 to -207 examples, near 493 repairs against 718 damaging edits on a 3,000-example slice. The setting the audit reports for dopanim is the pixel-level Swin corrector at 224 pixels, whose unthresholded stream repairs more than it damages. That ordering is the one

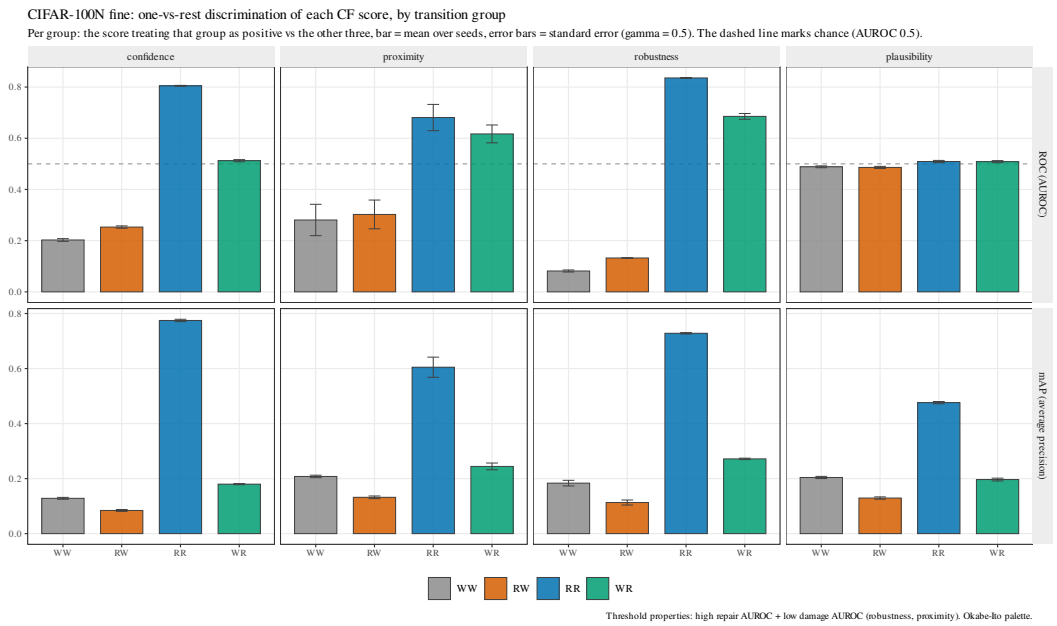


Figure D4: CIFAR-100N fine.

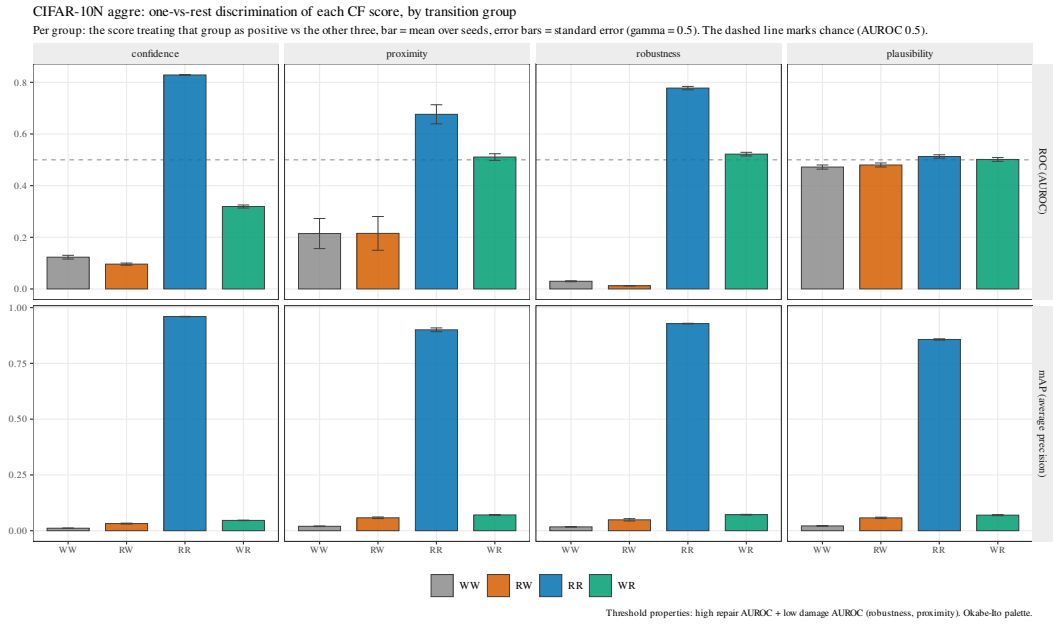


Figure D5: CIFAR-10N aggre.

reported for the LSNPC corrector itself, where a more capable image encoder generally yields larger improvements from correction (Huang et al., 2025).

Two orderings accompany the audit at this operating point. Behind the frozen encoder at 224 pixels, clean-label-referenced noise-robustness carries recovery discrimination while proximity stays near chance. In the separate archived diagnostic cohort of Table F11, label-free noise-robustness has AUROC 0.55 and label-free proximity 0.66, compared with chance at 0.5. Which score sits above chance follows the operating point, which is what the regime account of Section 5.1 describes.

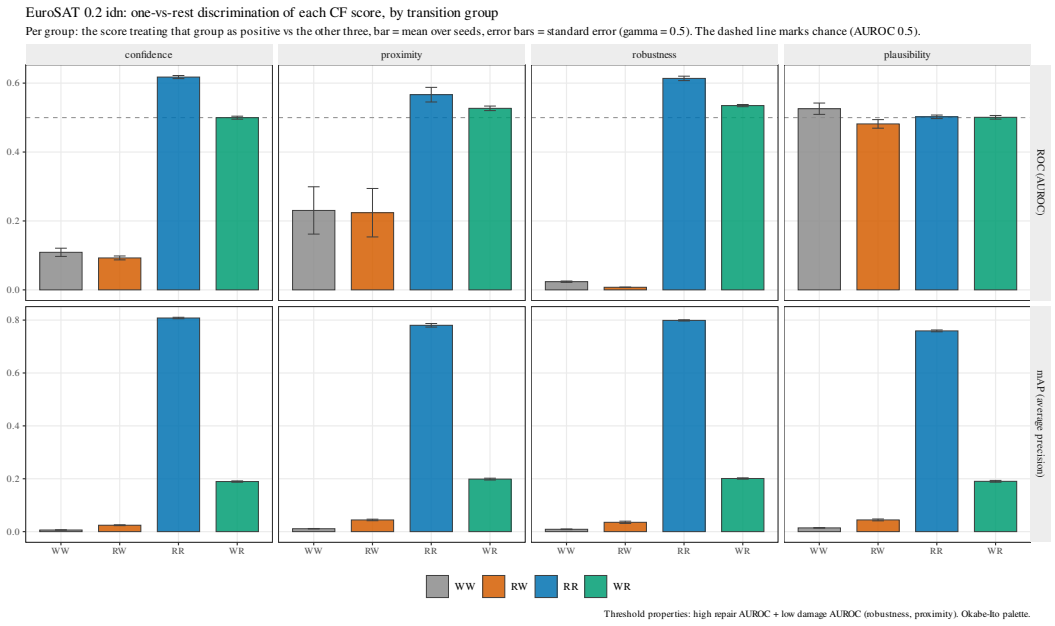


Figure D6: EuroSAT 0.2 IDN.

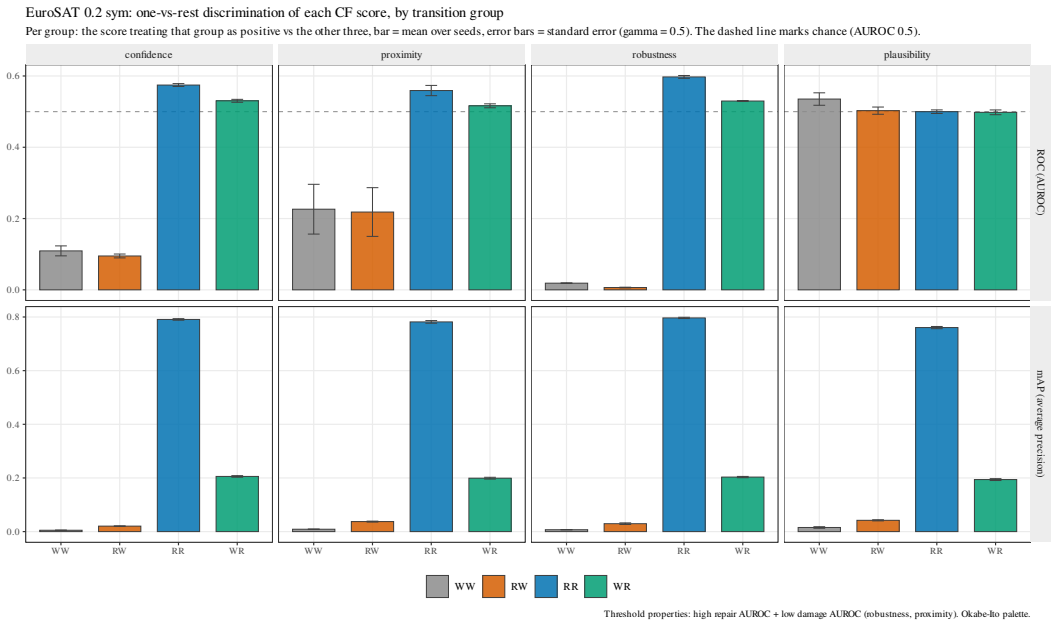


Figure D7: EuroSAT 0.2 SYM.

E.1 THE CLEAN-BUDGET LADDER

The semi-supervised variant trains LSNPC on the noisy stream together with a trusted set of clean labels. The objective keeps the form of Equation (2) and adds the same two terms over the trusted examples. The cross-entropy reconstructs the clean label at the corrected latent, and the clean-latent proposal conditions on that label. A single label-conditioned encoder produces both posteriors; no dedicated clean-encoder branch is needed. The two-phase schedule of Section A interleaves the unsupervised and supervised passes within each epoch. The trusted set can also be assembled from the audit itself, the counterfactual scores admitting the trustworthy corrected examples into the semi-supervised pool using the supervised calibration described in Section B.1. That training-time transfer is tested in Section H.1. On the trusted examples this variant conditions the clean-latent

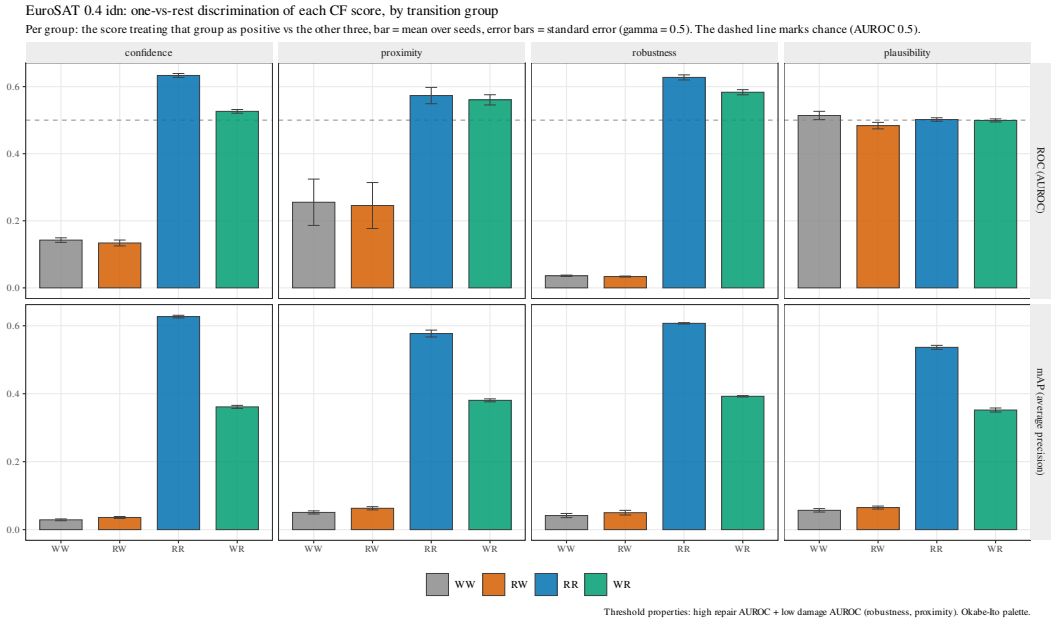


Figure D8: EuroSAT 0.4 IDN.

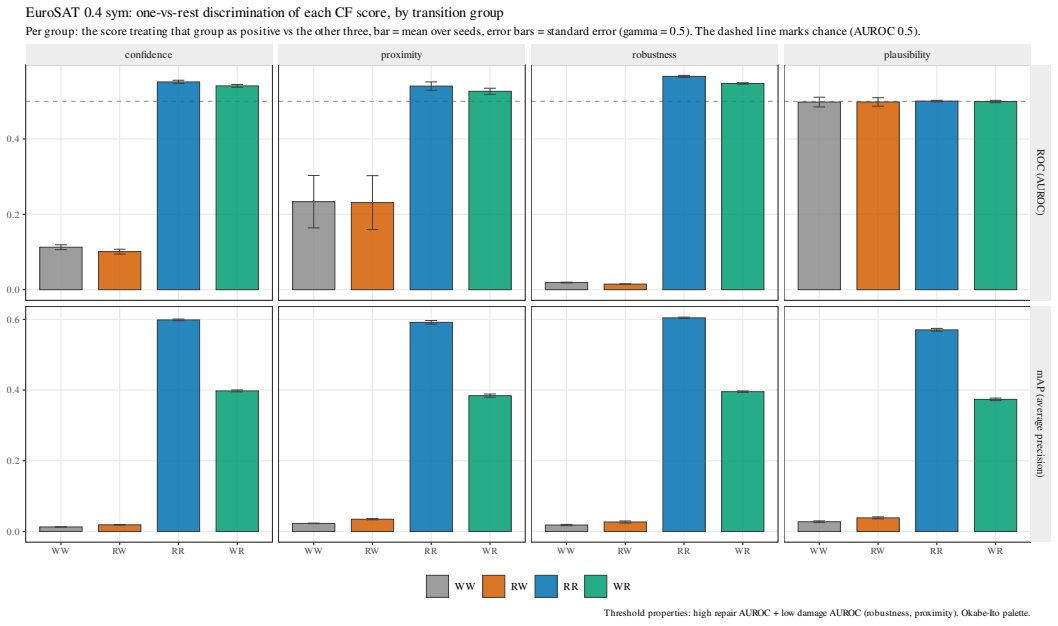


Figure D9: EuroSAT 0.4 SYM.

proposal on the clean label, as the mixture

$$q(\mathbf{z} \mid \mathbf{x}, y, \tilde{\mathbf{z}}) = \eta\, q(\mathbf{z} \mid \mathbf{x}, y) + (1 - \eta)\, q(\mathbf{z} \mid \tilde{\mathbf{z}}, \mathbf{x}),$$

where η is the probability that a trusted example conditions its proposal on its own label and the rest of the objective is unchanged. The study below holds η at the main configuration recipe of Section A and varies the clean-set size alone.

The clean-budget study isolates the cause with a nested design. The clean reference set of each budget is a prefix of the larger one, which makes any change in quality attributable to the budget itself. Table E5 (text) and Table E6 (image) hold the corrector settings fixed at the edit-rate coefficient $\gamma = 0.5$ and $\eta = 0.1$, and vary the budget over $\{5\%, 20\%, 100\%\}$ of 2,000 examples.

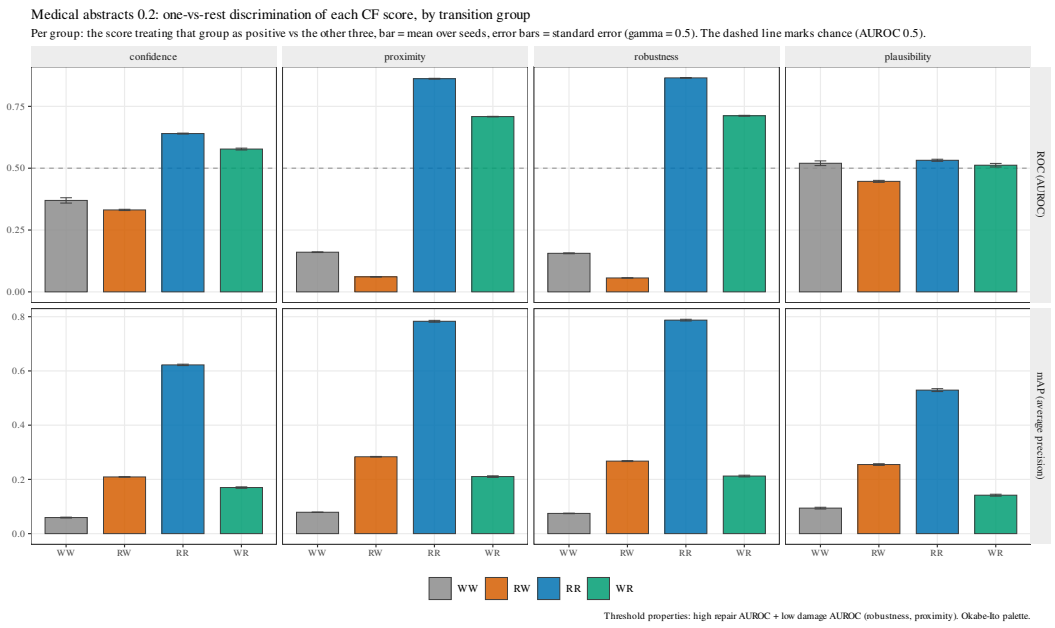


Figure D10: Medical-abstracts 0.2.

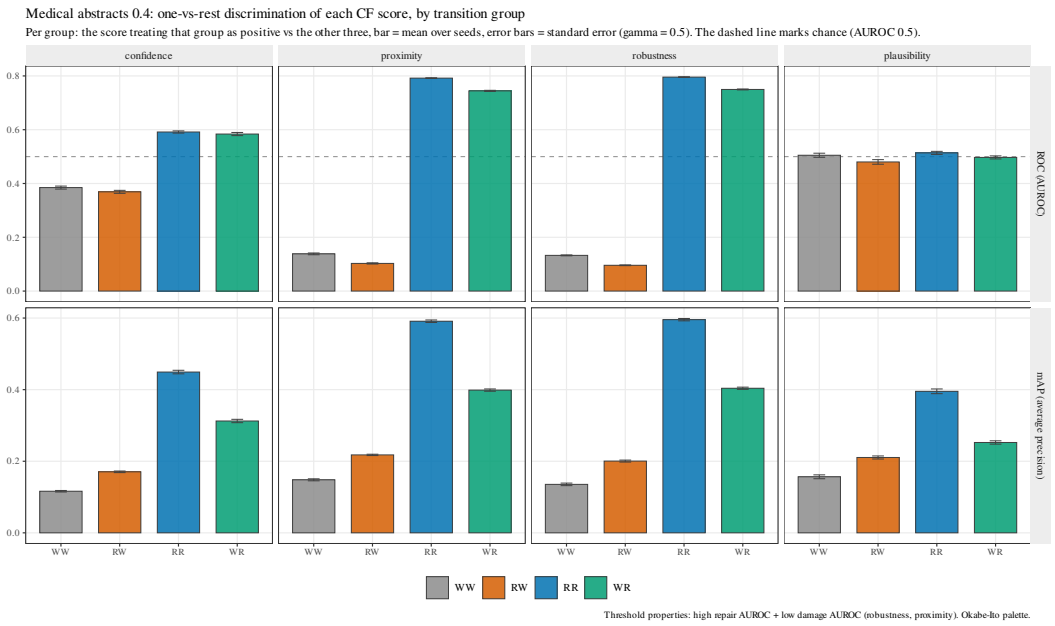


Figure D11: Medical-abstracts 0.4.

On text the budget moves performance by a few accuracy points, and the noise rate sets the sign (Figure E18). NoisyAG-News med, the real-crowd stream at a measured 19.8% flip, gains monotonically, from $+4.3$ to $+6.4$ accuracy points, and its the recovery rate rises with it. The four injected Medical-abstracts settings divide by rate rather than by budget. At 0.2 noise the corrector is near -15 accuracy points at every budget, while at 0.4 it gains near $+4$ and stays close to budget-invariant. Where the stream carries more to correct, correction improves the model. The budget decides by how much rather than whether. On image both datasets gain with the budget, CIFAR-10N worse from $+26.2$ to $+28.1$ accuracy points and CIFAR-100N fine from $+2.2$ to $+3.5$. A twentyfold budget reduction therefore costs about two accuracy points on CIFAR-10N worse and rather less on CIFAR-100N fine. The frozen ResNet-50 corrector is close to budget-invariant while still improving

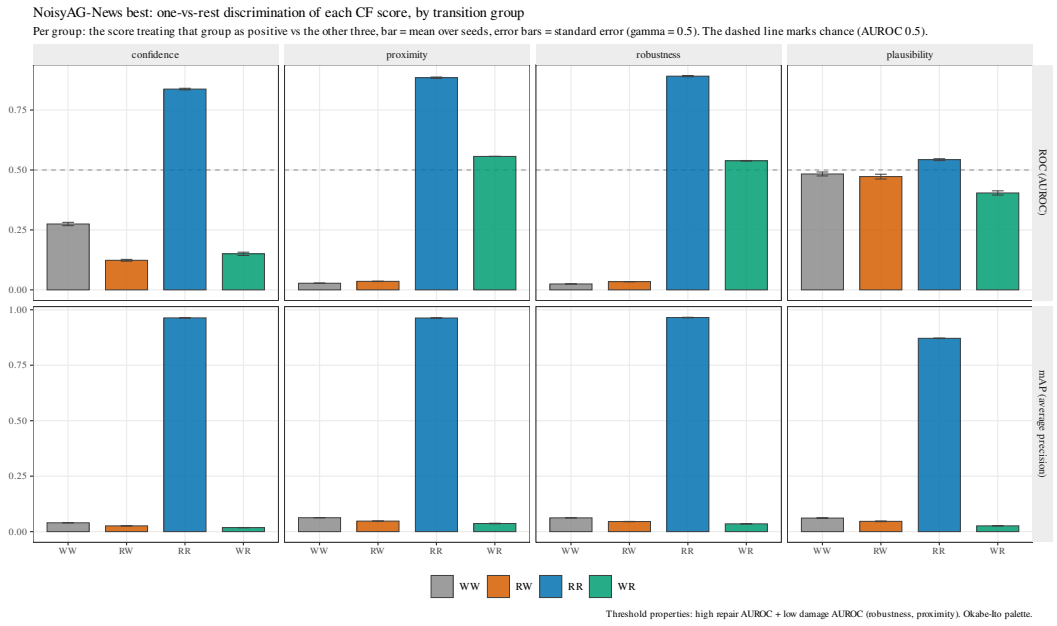


Figure D12: NoisyAG-News best.

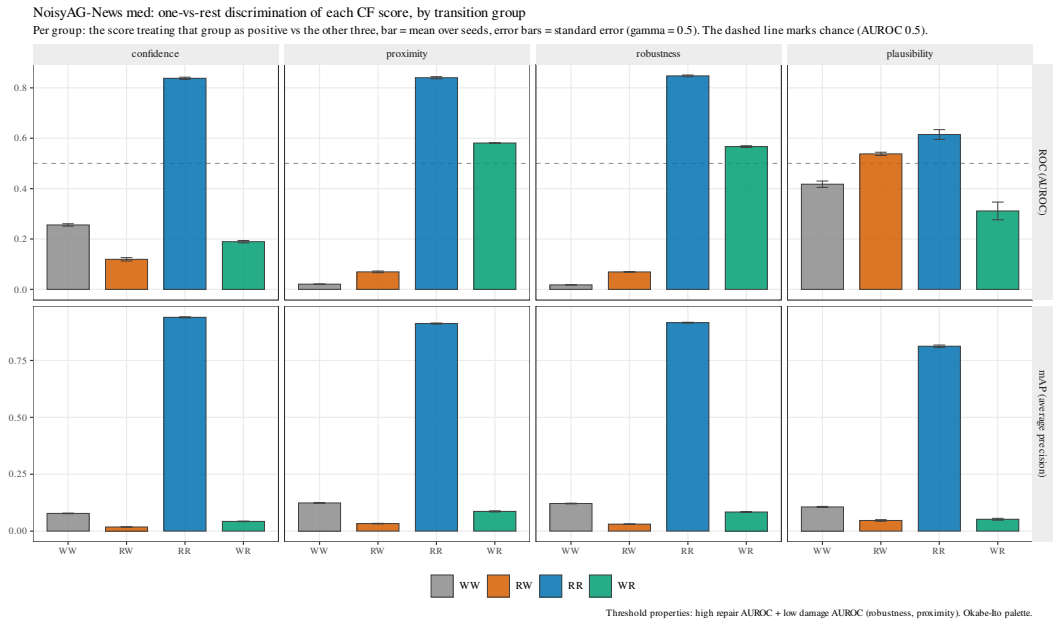


Figure D13: NoisyAG-News med.

at the full reference. The response is the same on both data types. The clean reference adds an accuracy point or two, and the noise in the stream decides the rest.

Table E7 lists the same comparison per headline setting at 10% and 50% coverage.

Table E8 reports each margin with the spread behind it: the interquartile range of the per-setting margins and one standard error over the settings quoted. The setting is the unit of inference, since its seeds share a clean split, an encoder and a noise realisation. A column quotes between five and seventeen settings. A case is one setting with one corrector seed; cases below the 30-correction floor drop out of their column's denominator.

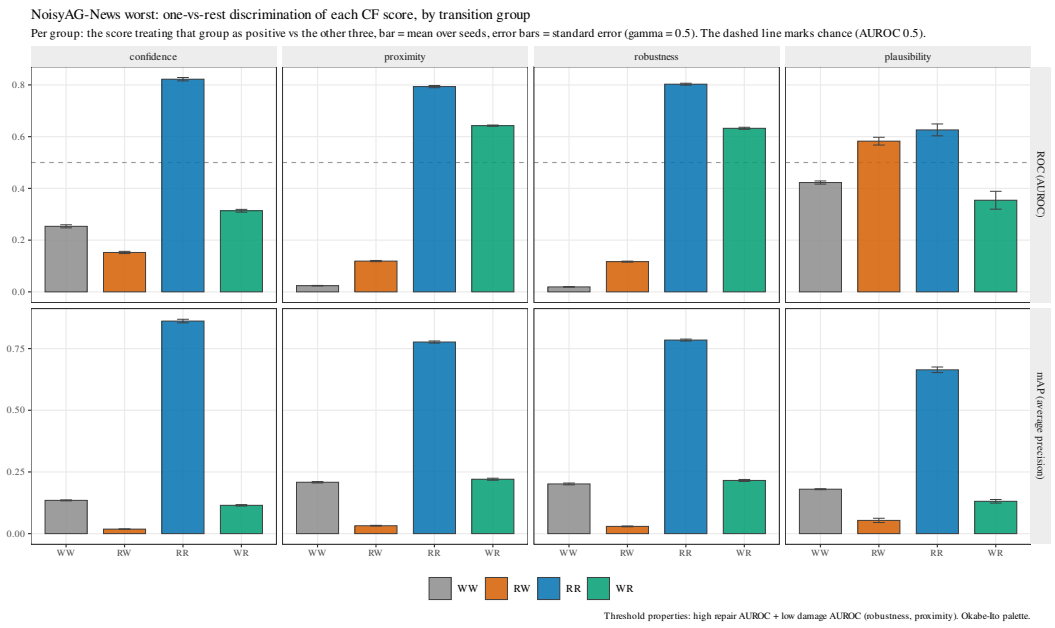


Figure D14: NoisyAG-News worst.

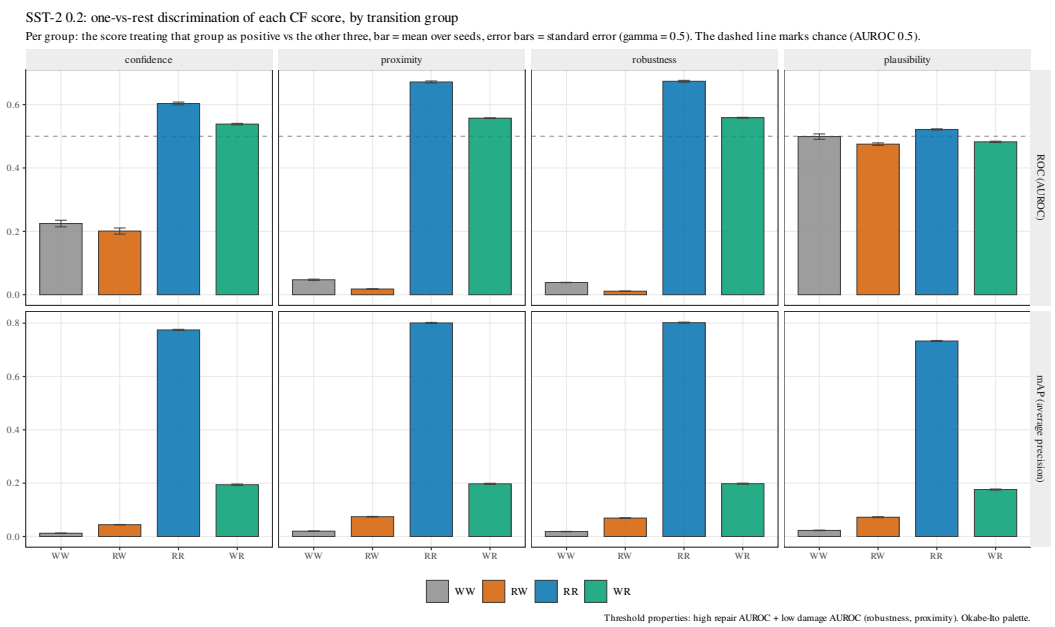


Figure D15: SST-2 0.2.

F Full End-to-End Comparison

This appendix reports the end-to-end results in two tables. Table F9 gives the compact comparison for the cross-entropy MLP, and Table F10 the complete one across all three downstream heads at the three reported coverage levels.

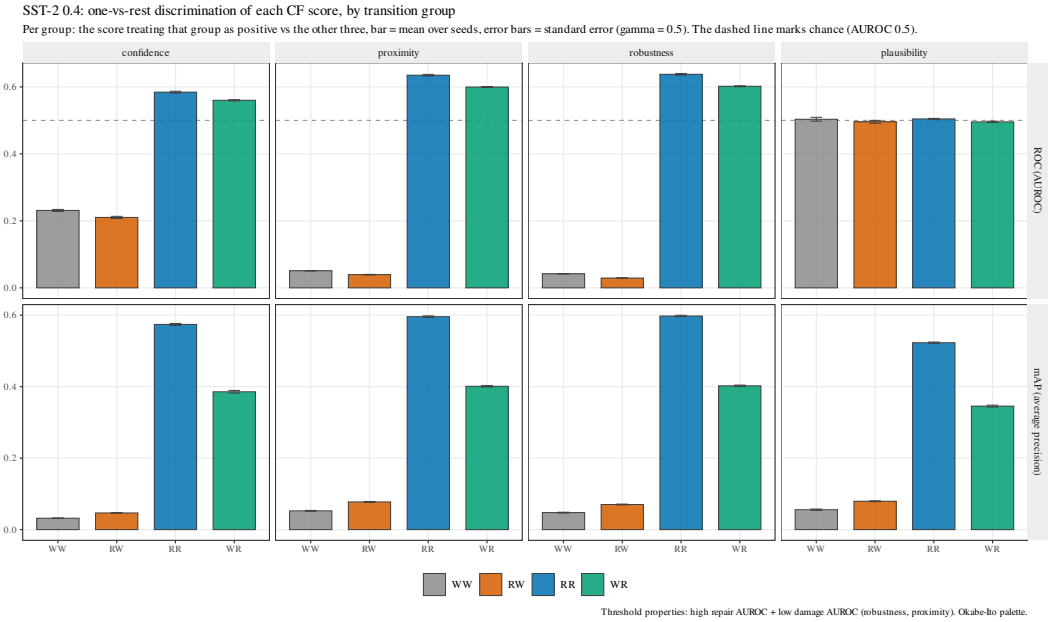


Figure D16: SST-2 0.4.

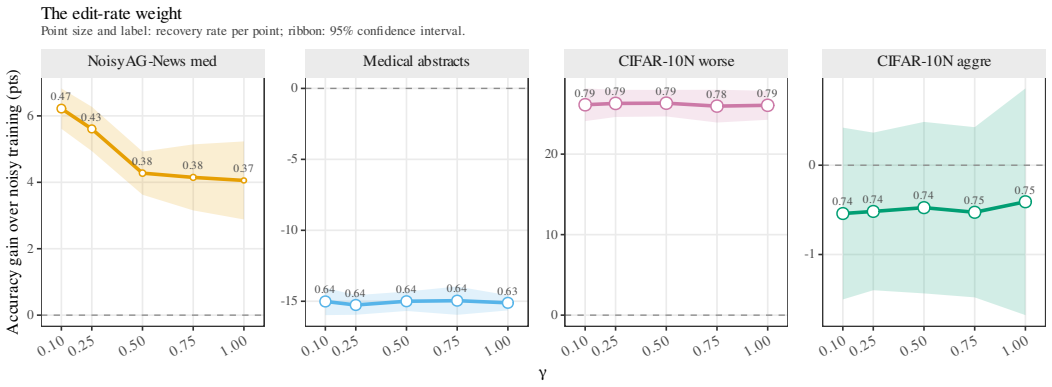
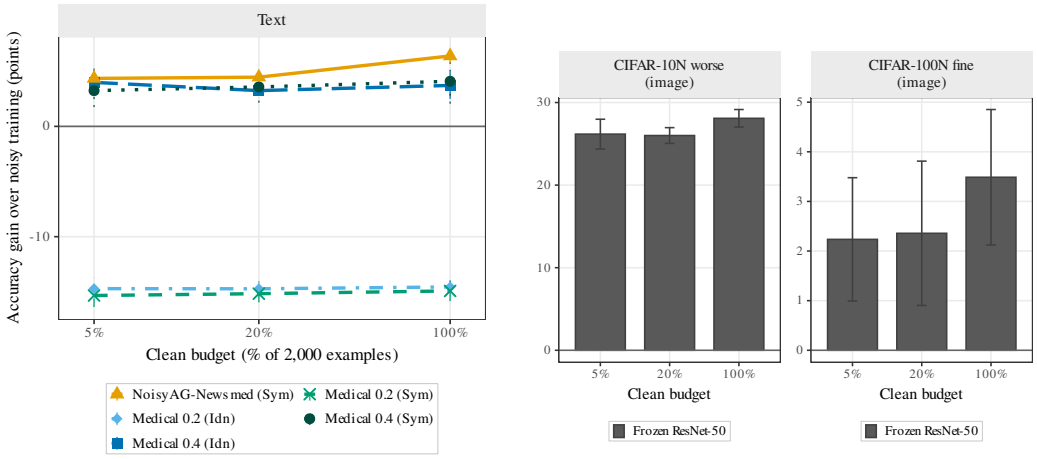


Figure E17: Accuracy gain over the noisy stream against the edit rate γ, one panel per setting of Table E3. Point size and labels encode the recovery rate. Shaded bands are the 95% confidence interval over seeds.

Table E4: Feature-space boundary: frozen encoders at the edit-rate coefficient γ=0.5, η=0.1, one entry per clean budget, five seeds. Recovery rate, accuracy gain (Δacc, points) and transition net (repairs minus damage). The frozen Swin corrector runs pixel mode at the same 32×32 input as the ResNet-50 configuration. Bold marks the frozen ResNet-50 entries, the encoder adopted as the operating point for text.

	CIFAR-10N worse			CIFAR-100N fine		
Corrector	Recov.	Δacc	Net	Recov.	Δacc	Net
Frozen ResNet-50, clean set 2,000	0.82	**+28.1**	+1194	0.46	**+3.4**	+143
Frozen ResNet-50, clean set 3,000	0.83	**+28.5**	+1212	0.47	**+4.4**	+188
Frozen ResNet-50, clean set 5,000	0.84	**+29.3**	+1244	0.49	**+5.3**	+225
Frozen Swin, clean set 2,000	0.87	+31.7	+1346	0.54	+10.0	+426
Frozen Swin, clean set 3,000	0.87	+32.1	+1363	0.56	+10.9	+461
Frozen Swin, clean set 5,000	0.87	+31.9	+1358	0.57	+11.0	+468



(a) Text. (b) Image.

Figure E18: Clean-budget ladder across every setting of the study. In (a) the real-crowd NoisyAG-News med stream and the four injected Medical-abstracts settings, two noise rates crossed with two noise types, are drawn as lines over the budget percentages, which are spaced logarithmically. In (b) the two image datasets are drawn as bars at the same three budgets. Marks are the mean accuracy gain over noisy training in points, and the vertical segments the 95% confidence intervals over runs, five per point. Each panel has its own legend and its own vertical range. Table E5 and Table E6 give the exact values.

Table E5: Clean-budget ladder on text (frozen encoder, the edit-rate coefficient γ=0.5, η=0.1), one column per text setting. Accuracy gain (Δacc, points) and the recovery rate at each nested budget of the 2,000-example clean reference. Five seeds per entry. The NoisyAG-News med stream carries a measured flipped-label fraction of 19.8%. Bold marks the real-crowd response. Underlines mark the entries whose accuracy falls below the noisy stream.

	NoisyAG-News med	Medical 0.2 (Idn)	Medical 0.2 (Sym)	Medical 0.4 (Idn)	Medical 0.4 (Sym)
Δacc (points), by clean budget					
2,000 x 5% (100)	**+4.3**	-14.7	-15.3	+4.0	+3.2
2,000 x 20% (400)	**+4.5**	-14.7	-15.2	+3.2	+3.6
2,000 x 100% (2,000)	**+6.4**	-14.6	-14.9	+3.7	+4.1
Recovery rate, by clean budget					
2,000 x 5% (100)	0.38	0.62	0.63	0.63	0.63
2,000 x 20% (400)	0.39	0.63	0.63	0.62	0.64
2,000 x 100% (2,000)	0.46	0.62	0.64	0.63	0.64

Table E6: Clean-budget ladder on real image noise (the edit-rate coefficient γ=0.5, η=0.1, five seeds). Recovery rate, accuracy gain (Δacc, points) and transition net (repairs minus damage). The image side of the study runs the frozen ResNet-50 corrector at all three budgets. Bold marks the accuracy gain.

	CIFAR-10N worse			CIFAR-100N fine		
Encoder, clean-set size	Recov.	Δacc	Net	Recov.	Δacc	Net
Frozen ResNet-50, clean set 100	0.79	**+26.2**	+1112	0.44	**+2.2**	+95
Frozen ResNet-50, clean set 400	0.79	**+26.0**	+1105	0.44	**+2.4**	+100
Frozen ResNet-50, clean set 2,000	0.82	**+28.1**	+1194	0.46	**+3.5**	+148

G The Repair Ceiling of a Single-Proposal Rule and an Open Extension

The selective rule of Section B.1 keeps, for each candidate example, either its current label or the single label the corrector proposes. This appendix states the limit that such a rule inherits from

Table E7: Clean-test accuracy gain (points) over the never-correct stream, in the five settings of Section F that carry a thresholded retrain, the three crowd-labelled NoisyAG-News streams and two image datasets. The corrector is held fixed at frozen backbones and five epochs (trusted-set provenance qualified in Section A), five downstream seeds each. The unselective column applies every correction, the two threshold columns apply the proximity-and-robustness rank threshold at 10% and 50% coverage. The last two columns give the mechanism at 50% coverage, measured on the held-out audit slice, namely the share of the corrector's damaging edits reverted and the share of its repairs retained. Bold marks the strongest of unselective correction and the two threshold coverages per entry (ties bolded), and underlines mark the reverted damaging share.

	Never-correct	Gain over never-correct (pts)			At 50% coverage	
Configuration	clean-test acc. (%)	Unselective (100% coverage)	Threshold, 10% coverage	Threshold, 50% coverage	Damaging removed	Repairs retained
Crowd-labelled text						
NoisyAG-News worst	75.9 ± 0.4	$+1.0 \pm 1.4$	$+0.4 \pm 0.6$	$+1.2 \pm 1.1$	$\underline{100.0 \pm 0.0}$	59.8 ± 1.4
NoisyAG-News med	84.3 ± 0.3	$+0.7 \pm 0.3$	$+0.2 \pm 0.3$	$+0.6 \pm 0.3$	$\underline{99.9 \pm 0.3}$	72.9 ± 3.2
NoisyAG-News best	88.6 ± 0.1	$+0.2 \pm 0.2$	$+0.0 \pm 0.1$	$+0.1 \pm 0.2$	$\underline{84.1 \pm 2.1}$	100.0 ± 0.0
Image (frozen ResNet-50)						
CIFAR-10N (worse)	85.2 ± 0.7	$+1.2 \pm 0.9$	$+0.5 \pm 0.7$	$+0.9 \pm 0.7$	$\underline{100.0 \pm 0.0}$	58.6 ± 1.2
CIFAR-100N (fine)	59.3 ± 0.3	$+2.1 \pm 0.4$	$+0.3 \pm 0.5$	$+1.3 \pm 0.2$	$\underline{100.0 \pm 0.0}$	91.2 ± 1.8

Table E8: The same margins as Table 3, each with the interquartile range of the per-setting margins beneath it and one standard error over the settings behind the column. A case is one setting with one corrector seed, five seeds per setting; settings missing any of their five seeds drop. Columns count only cases whose coverage retains at least 30 corrections, and each column's settings appear above. Bold marks every entry that attains the largest margin, ties included.

Threshold score	Text, 10%	Text, 50%	Image, 10%	Image, 50%
Settings per column	5	17	11	17
Confidence	$+0.23 \pm 0.06$ $[+0.16, +0.31]$	$+0.92 \pm 0.17$ $[+0.30, +1.60]$	$+0.84 \pm 0.25$ $[+0.18, +1.79]$	$+2.77 \pm 0.85$ $[+0.28, +1.87]$
Proximity	$+0.17 \pm 0.07$ $[+0.05, +0.20]$	$+1.91 \pm 0.24$ $[+1.20, +2.60]$	$+1.29 \pm 0.40$ $[+0.12, +2.33]$	$+4.98 \pm 1.45$ $[+1.51, +8.56]$
Noise-robustness	$\mathbf{+0.31 \pm 0.07}$ $[+0.20, +0.40]$	$+1.57 \pm 0.34$ $[+0.64, +2.76]$	$\mathbf{+1.50 \pm 0.45}$ $[+0.19, +2.85]$	$+5.25 \pm 1.71$ $[+0.99, +8.02]$
Plausibility	-0.54 ± 0.16 $[-0.60, -0.28]$	-0.52 ± 0.08 $[-0.60, -0.32]$	-0.45 ± 0.10 $[-0.53, -0.22]$	-0.37 ± 0.09 $[-0.64, -0.13]$
Proximity $\wedge$ robustness	$+0.20 \pm 0.08$ $[+0.12, +0.20]$	$\mathbf{+1.96 \pm 0.26}$ $[+1.28, +2.80]$	$\mathbf{+1.50 \pm 0.45}$ $[+0.15, +2.89]$	$\mathbf{+5.45 \pm 1.67}$ $[+1.47, +8.58]$

Table F9: End-to-end clean-test accuracy (%) for the cross-entropy MLP, mean $\pm$ standard error over five corrector seeds (five downstream seeds each, 25 runs per entry). Corrector training provenance is qualified in Section A. The 10% entry is the specified operating coverage; the 50% and 90% entries widen the same rank threshold. Co-teaching is computed for this model only. dopanim's corrector is a pixel-mode Swin, a different family from the frozen-feature correctors used on the CIFAR settings. Table F10 gives the other heads. The dopanim setting carries no clean-label reference; its (d) entry is blank. Bold marks the better of the thresholded stream at 90% coverage and Co-teaching.

	Text (NoisyAG-News)			Image		
Training configuration	worst	med	best	CIFAR-10N (worse)	CIFAR-100N (fine)	dopanim
(a) original noisy labels	75.9 $\pm$0.2	84.3 $\pm$0.1	88.6 $\pm$0.1	85.2 $\pm$0.3	59.3 $\pm$0.1	72.3 $\pm$0.3
(b) unselective correction	76.8 $\pm$0.5	84.9 $\pm$0.1	88.8 $\pm$0.1	86.3 $\pm$0.3	61.4 $\pm$0.1	73.5 $\pm$0.3
(c) thresholded, 10% coverage (operating)	76.3 $\pm$0.1	84.5 $\pm$0.1	88.6 $\pm$0.1	85.7 $\pm$0.2	59.6 $\pm$0.1	73.0 $\pm$0.4
(c) thresholded, 50% coverage	77.1 $\pm$0.4	84.9 $\pm$0.1	88.7 $\pm$0.1	86.1 $\pm$0.1	60.6 $\pm$0.1	73.8 $\pm$0.4
(c) thresholded, 90% coverage	**76.8 $\pm$0.6**	**84.9 $\pm$0.2**	**88.9 $\pm$0.1**	86.6 $\pm$0.3	**61.5 $\pm$0.0**	**73.8 $\pm$0.4**
(f) random rejection, 50% coverage	76.8 $\pm$0.5	84.7 $\pm$0.0	88.8 $\pm$0.1	86.4 $\pm$0.3	60.5 $\pm$0.1	73.5 $\pm$0.4
(e) Co-teaching (in-training)	73.6 $\pm$0.5	84.7 $\pm$0.1	88.0 $\pm$0.1	**87.8 $\pm$0.0**	61.3 $\pm$0.1	61.9 $\pm$0.4
(d) clean-label reference	90.1 $\pm$0.1	90.1 $\pm$0.1	90.1 $\pm$0.1	91.1 $\pm$0.0	69.2 $\pm$0.1	–

the corrector, and records an open extension that widens the candidate set. End-to-end evaluation covers the single-proposal rule. The extension below is an exploratory, in-sample label-level study

Table F10: End-to-end clean-test accuracy (%; 25 runs per entry). The first block holds the cross-entropy MLP configurations, the lower blocks the linear and asymmetric-loss heads at the same coverage levels. Bold marks the best non-oracle result per dataset across the three heads (ties bolded). The random-rejection control is the MLP's, as in Table F9; dopanim's corrector is a pixel-mode Swin, a different family from the frozen-feature correctors used on the CIFAR settings. Entries carrying a dagger have one corrector seed diverging from the other four: the (b) entry on med under the asymmetric loss (± 2.4).

	Text (NoisyAG-News)			Image		
Training configuration	worst	med	best	CIFAR-10N (worse)	CIFAR-100N (fine)	dopanim
Three-layer MLP, cross-entropy						
(a) original noisy labels	75.9 ±0.2	84.3 ±0.1	88.6 ±0.1	85.2 ±0.3	59.3 ±0.1	72.3 ±0.3
(b) unselective correction	76.8 ±0.5	84.9 ±0.1	88.8 ±0.1	86.3 ±0.3	61.4 ±0.1	73.5 ±0.3
(c) thresholded, 10% coverage (operating)	76.3 ±0.1	84.5 ±0.1	88.6 ±0.1	85.7 ±0.2	59.6 ±0.1	73.0 ±0.4
(c) thresholded, 50% coverage	77.1 ±0.4	84.9 ±0.1	88.7 ±0.1	86.1 ±0.1	60.6 ±0.1	73.8 ±0.4
(c) thresholded, 90% coverage	76.8 ±0.6	84.9 ±0.2	**88.9 ±0.1**	86.6 ±0.3	61.5 ±0.0	73.8 ±0.4
(f) random rejection, 50% coverage	76.8 ±0.5	84.7 ±0.0	88.8 ±0.1	86.4 ±0.3	60.5 ±0.1	73.5 ±0.4
(e) Co-teaching (in-training)	73.6 ±0.5	84.7 ±0.1	88.0 ±0.1	**87.8 ±0.0**	61.3 ±0.1	61.9 ±0.4
(d) clean-label reference	90.1 ±0.1	90.1 ±0.1	90.1 ±0.1	91.1 ±0.0	69.2 ±0.1	–
Single-layer linear head						
(a) original noisy labels	77.9 ±0.0	84.8 ±0.1	88.1 ±0.0	81.4 ±0.9	59.5 ±0.3	72.4 ±0.3
(b) unselective correction	78.0 ±0.3	85.0 ±0.1	88.7 ±0.1	86.3 ±0.3	62.5 ±0.1	73.4 ±0.4
(c) thresholded, 10% coverage (operating)	78.1 ±0.1	84.8 ±0.1	88.2 ±0.0	82.4 ±0.7	59.5 ±0.4	72.3 ±0.2
(c) thresholded, 50% coverage	**78.5 ±0.2**	**85.0 ±0.1**	88.4 ±0.1	83.9 ±0.3	61.3 ±0.3	73.3 ±0.5
(c) thresholded, 90% coverage	78.0 ±0.3	85.0 ±0.1	88.7 ±0.1	86.1 ±0.4	**62.5 ±0.1**	73.7 ±0.4
(d) clean-label reference	90.0 ±0.0	90.0 ±0.0	90.0 ±0.0	91.1 ±0.0	71.1 ±0.1	–
Three-layer MLP with asymmetric loss						
(a) original noisy labels	74.8 ±0.2	84.0 ±0.1	88.4 ±0.0	85.7 ±0.3	59.0 ±0.2	71.9 ±0.3
(b) unselective correction	76.8 ±0.6	82.5 ±2.4†	88.8 ±0.1	86.5 ±0.3	61.5 ±0.1	73.5 ±0.3
(c) thresholded, 10% coverage (operating)	75.4 ±0.0	84.2 ±0.1	88.5 ±0.1	86.0 ±0.3	59.4 ±0.1	72.6 ±0.5
(c) thresholded, 50% coverage	76.3 ±0.5	84.7 ±0.1	88.6 ±0.1	86.5 ±0.1	60.4 ±0.0	73.7 ±0.4
(c) thresholded, 90% coverage	76.3 ±0.6	84.8 ±0.2	88.8 ±0.1	86.7 ±0.3	61.3 ±0.1	**73.9 ±0.4**
(d) clean-label reference	90.4 ±0.1	90.4 ±0.1	90.4 ±0.1	91.3 ±0.1	68.8 ±0.0	–

Table F11: Clean-label-referenced and label-free score variants in a separate archived diagnostic cohort. *Clean-label* uses the true label as score reference; *label-free* uses the emitted label. Both columns measure WR versus WW AUROC on the same held-out examples within each setting-seed pair. Means use six corrector seeds per text/CIFAR setting and three for dopanim (33 pairs), rather than the five-by-five downstream cohort. The dopanim archive tag differs from the Swin retraining tag; no direct model-matched comparison is claimed. Bold marks label-free scores.

			Proximity		Noise-robustness	
Setting	Type	Seeds	clean-label	label-free	clean-label	label-free
NoisyAG-News worst	text	6	0.861	**0.633**	0.921	**0.625**
NoisyAG-News med	text	6	0.865	**0.626**	0.947	**0.432**
NoisyAG-News best	text	6	0.869	**0.634**	0.958	**0.321**
CIFAR-10N worse	image	6	0.782	**0.761**	0.949	**0.720**
CIFAR-100N fine	image	6	0.864	**0.744**	0.972	**0.571**
dopanim	image	3	0.523	**0.506**	0.989	**0.676**
Mean over every setting-seed pair			0.819	0.664	0.953	0.547

without downstream retraining. Let y_i be the clean label of $\mathbf{x}_i$, used only for evaluation, and let the current stream label be $\ell_i = \tilde{y}_i$, the noisy annotation. The corrector proposes $\hat{c}_i = \mathcal{C}(\mathbf{x}_i, \ell_i)$. Over the examples whose current label is wrong, $\mathcal{S} = \{i : \ell_i \neq y_i\}$, the corrector recovers the share $\bar{v} = |\mathcal{S}_1|/|\mathcal{S}|$, where $\mathcal{S}_1 = \{i \in \mathcal{S} : \hat{c}_i = y_i\}$ are the recovered examples, the recovery ceiling of Section 4. A selective rule whose only actions are to keep ℓ_i or adopt $\hat{c}_i$ can return the clean label only on $\mathcal{S}_1$, where the proposal itself is clean. On a wrong example the corrector does not recover, $\mathcal{S} \setminus \mathcal{S}_1$, both ℓ_i and $\hat{c}_i$ are wrong, and no threshold over the audit scores can return y_i. The rule therefore realises a repair rate of at most $\bar{v}$ times the fraction of $\mathcal{S}_1$ it keeps. The scores govern only that multiplier, deciding which recovered examples to keep while rejecting damaging relabels, while the ceiling $\bar{v}$ itself is set by the corrector's recovery rate. The ceiling is a property of offering one label per example. A wrong example the argmax misses may still place its clean label among the runner-up classes of the corrected-label posterior $p_c(y \mid \mathbf{x}, \ell)$. That posterior is the distribution over classes induced by decoding the corrected latent through the shared decoder g_ϕ, whose mode

Table G12: Top-w coverage of the clean label in the corrected-label posterior, per setting and width (share of the wrong examples whose clean label falls in the top w; mean over five seeds). Bold marks w=1, the recovery ceiling of the single-proposal rule; underlines mark the operating width w=2. Candidate-extension study over five corrector seeds, one encoder family per data type.

Setting	w=1	w=2	w=3	w=5	w=10
NoisyAG-News worst	**0.56**	<u>0.85</u>	0.97	1.00	1.00
NoisyAG-News med	**0.38**	<u>0.85</u>	0.98	1.00	1.00
NoisyAG-News best	**0.24**	<u>0.79</u>	0.96	1.00	1.00
CIFAR-10N	**0.59**	<u>0.83</u>	0.91	0.97	1.00
CIFAR-100N	**0.23**	<u>0.47</u>	0.60	0.73	0.84

is $\hat{c}$. Decoding averages over the corrected latent, which makes the posterior the expectation of the decoder's class distribution over latent samples, estimated by an importance-weighted average of the per-sample distributions,

$$p_c(y \mid \mathbf{x}, \ell) \;\approx\; \sum_{k=1}^{M} w_k \left[g_\phi(\mathbf{x}, \mathbf{z}_k)\right]_y, \tag{8}$$

where w_k are the sample weights of the correction. Evaluating the expectation at the mean corrected latent instead is a plug-in estimate and need not equal this average, because g_ϕ is nonlinear in the latent. Equation (8) keeps the averaging at the label level. For a candidate width w, let $\mathcal{W}_i$ be the w most probable classes of $p_c(\cdot \mid \mathbf{x}_i, \ell_i)$. An unrecovered wrong example whose clean label lies in $\mathcal{W}_i$ then carries a candidate equal to the clean label and can in principle be recovered. A width $w = 1$ rule, by contrast, stops at the single proposal. The classes of $\mathcal{W}_i$ are considered in order of posterior probability and the first one that differs from ℓ_i and clears a label-free trust floor is adopted. If none clears the floor, the current label is kept. The population coverage rule then ranks the candidate examples and retains the top fraction. Two constraints bound the extra freedom that wider candidates introduce. An agreement check withholds a relabel when an independent classifier is already confident in the current label, which suppresses damage to already-correct examples and the substitution of one wrong label for another. Relabels that move a wrong label to a different wrong label are counted separately as wasted. The width w and the trust floor are chosen experimentally. The quantities of interest per setting are the top-w coverage and the share of wrong examples whose clean label falls in $\mathcal{W}_i$. Two more are the flip rate realised under the rule and the damage rate relative to width $w = 1$. Whether wider candidates raise the repair rate above $\bar{v}$ is open. Examples the model labels with high confidence yet misclassifies are the ones expected to resist re-conditioning, which may leave the top-w coverage close to $\bar{v}$. The evaluation emits five quantities for every pool example of the five primary settings over five corrector seeds. The first is the corrected-label posterior of (8), together with the plug-in estimate at the mean corrected latent. The remaining three are the w most probable classes and their probabilities, and the posterior mass of the current label. The last is an independent classifier's probability of the current label and of each candidate class. The clean label scores the outcome and never drives the rule. Three trust floors are applied over widths $w \in \{1, 2, 3, 5, 10\}$, namely a floor relative to the posterior mode, the status-quo floor, and their conjunction. An agreement check is applied separately, withholding a relabel when an independent classifier prefers the current label to the candidate by a factor λ. Throughout, a pool example carries the transition group of the corrector's action, as introduced above. The rule is scored by four counts, in examples, over the pool of size n. Repairs gained are examples the corrector left alone that the rule moves onto the clean label. Damage added counts the correctly labelled examples the rule moves off it. Repairs reverted are the corrector's own WR recoveries the rule declines to adopt. Wasted moves are wrong examples moved to a different wrong label. Their combination is reported as a net delta in accuracy points, $(\text{gained} - \text{added} - \text{reverted})/n \times 100$, the change against the single-proposal rule on the same examples. Quoted quantities are means over the five seeds. Table G12 reports the top-w coverage. The clean label is the runner-up for 29 to 55% of the wrong examples on the crowd text, which is why coverage nearly doubles between w=1 and w=2 there. On CIFAR-100N it is spread over a longer tail, where 16% of wrong examples place it beyond rank 9. On the examples the corrector leaves alone, coverage at w=1 is exactly zero. The single proposal is the current label on such an example by construction, which is the ceiling of Section 4 restated at the level of the candidate set. Those examples are a minority of the pool, 7 to 22%, against 57 to 88% of examples the corrector also leaves alone while labelling correctly, a ratio of 2.6 to 12.7. The two groups are what a per-candidate score has to separate, since both carry a confident posterior. The mean mass

Table G13: The rule's actions by transition group at floor ρ=0.5 and width w=3, without the agreement check; one entry per transition group, averaged over five seeds (pool sizes 44,500, 45,500, 45,000, 37,000, 37,000). Bold marks the realised gains (WW onto the clean label) and losses (RR off it). Candidate-extension study over five corrector seeds, one encoder family per data type.

	NoisyAG-News worst	NoisyAG-News med	NoisyAG-News best	CIFAR-10N	CIFAR-100N
RR examples (correct, untouched)	26,246	35,451	39,481	21,423	21,213
moved off the clean label	**383**	**248**	**232**	**151**	**735**
RW examples (damaged by the corrector)	1,437	1,045	1,112	726	940
left damaged	1,437	1,045	1,112	726	940
WR examples (repaired by the corrector)	9,459	3,418	1,070	8,813	3,379
adopted	9,459	3,418	1,070	8,813	3,379
WW examples (wrong, unrecovered)	7,358	5,586	3,337	6,038	11,468
moved	2,056	784	362	2,245	4,819
onto the clean label	**512**	**256**	**118**	**311**	**639**
to another wrong label	1,544	528	244	1,934	4,181

of the current label is 0.65 to 0.95 on the missed examples against 0.87 to 0.99 on the correctly labelled ones. A floor that requires the candidate to exceed the current label returns the corrector's own labels at every width, verified on all 25 setting-seed combinations at $w \in \{2, 3, 5, 10\}$. On an example the corrector edits, the posterior mode already exceeds the current label and is adopted. On an example it leaves alone, the current label is the mode, and the floor holds it in place. The posterior mode is therefore the reference a floor needs to act at all. The net of the rule decomposes into two effects that live in different transition groups. The rule can only operate on examples the corrector left alone, and those are exactly the RR and WW examples. Every repair it gains therefore comes out of WW, and every label it breaks comes out of RR. Table G13 writes the accounting out. Two structural facts fall out of it. Without the agreement check the WR group is adopted wholesale at every width and floor, because the posterior mode on such an example is the clean label, which clears both floors. The extension's gains therefore come from the examples the corrector leaves alone, and the ceiling of Section 4 is untouched on the repair channel the corrector already covers. The agreement check changes that, and it limits damage by giving up adopted repairs (the paragraph below). The RW group is likewise untouched, in the opposite sense, since the rule never reverts a damaging edit the corrector proposed. The extension is therefore orthogonal to the damage control the rank threshold performs in the main text (Section 5.2). Damage control stays with the threshold, while the extension operates on the repair channel. Table G14 varies the width at the operating floor, and widths beyond w=2 return the same examples, though the coverage of Table G12 is still rising at w=10. The two quantities are consistent, and the rule's scan gives the reason. The rule walks the candidate classes in posterior order, skips the current label, and stops at the first class that clears the floor. On every example it is allowed to touch, the current label is by definition the posterior mode, and the scan begins at the runner-up. If the runner-up fails the floor, every later class fails it too, because the probabilities are sorted in descending order. No class beyond the runner-up is ever adopted, at any width and at any positive floor. The two-class decision between the current label and its nearest competitor is the most this family of floors can express. Coverage counts something else, namely the share of examples whose clean label sits anywhere in the top-w, aggregated over all pool examples including the ones the corrector edits. A posterior whose tail holds the clean label at rank four therefore raises the coverage while leaving the rule's decisions unchanged. The difference between the two quantities is exactly what a signal able to rank deeper candidates could add. Width matters only where the agreement check runs with the floor set to zero. There the check can decline the runner-up on grounds other than its probability, the scan continues into the tail, and the wider settings move monotonically downward, from -3.13 to -3.84 accuracy points on CIFAR-10N between w=2 and w=10 and from -2.41 to -2.76 on NoisyAG-News best, the tail offering weaker candidates than the runner-up it replaces. Widening helps when the new candidates outrank the one they displace, and supplying that ranking is the open target the audit's scores set for future per-candidate work. Table G15 varies the floor at w=3. A floor at zero, which admits the runner-up whenever it exists, relabels a majority of the examples the corrector left alone and destroys 49 to 83 accuracy points. The posterior's second choice exists on almost every untouched example, and most of those examples are labelled correctly. Tightening the floor trades the gains and the losses down together and leaves the sign pattern intact. The pattern is positive on the two settings where the corrector is damage-heavy, CIFAR-10N and the severe crowd text, and on NoisyAG-News med. It is negative where the corrector is already reliable (NoisyAG-News best) and marginal on CIFAR-100N. The extension is therefore a smaller version of the same trade-off the threshold makes. It helps where damage risk is real, and where it is not the thresholded stream already suffices. The

Table G14: Net delta in accuracy points against the single-proposal rule, by width w at floor ρ=0.5 and no agreement check (mean over five seeds). Bold marks the operating width w=2. Candidate-extension study over five corrector seeds, one encoder family per data type.

Width	NoisyAG-News worst	NoisyAG-News med	NoisyAG-News best	CIFAR-10N	CIFAR-100N
w=1	0.00	0.00	0.00	0.00	0.00
w=2	**+0.29**	**+0.02**	**-0.25**	**+0.43**	**-0.26**
w=3	+0.29	+0.02	−0.25	+0.43	−0.26
w=5	+0.29	+0.02	−0.25	+0.43	−0.26
w=10	+0.29	+0.02	−0.25	+0.43	−0.26

Table G15: Net delta in accuracy points against the single-proposal rule, by trust floor at width w=3, no agreement check. The floor is a fraction of the posterior mode. Bold marks the operating floor; underlines mark the zero floor. Candidate-extension study over five corrector seeds, one encoder family per data type.

Floor ρ	NoisyAG-News worst	NoisyAG-News med	NoisyAG-News best	CIFAR-10N	CIFAR-100N
0	−49.67	−69.11	−82.52	−50.41	−49.15
0.25	+0.31	−0.10	−0.57	+0.74	−1.73
0.50	**+0.29**	**+0.02**	**-0.25**	**+0.43**	**-0.26**
0.75	+0.14	+0.02	−0.11	+0.20	+0.01

agreement check, instantiated as the appendix describes it, with a classifier trained on the same noisy labels, is confident exactly where the labels are wrong. Those are the labels it was fitted to. Under the comparative form at λ=0.5 it therefore declines the corrector's own WR recoveries along with the extension's moves. The adopted repairs fall from the full WR group to 0 on CIFAR-100N and to 3,628 of 9,459 on NoisyAG-News worst. They fall to 1,438 of 3,418 on NoisyAG-News med, 581 of 1,070 on best, and 5,875 of 8,813 on CIFAR-10N. It does limit the damage, reducing the damaging edits at the operating width from 1,820 to 132 on the worst crowd text and from 1,676 to zero on CIFAR-100N. It also removes almost all of the wasted moves. The two effects do not balance. The net delta against the single-proposal rule falls from +0.29 to −10.17 accuracy points on NoisyAG-News worst, from +0.43 to −6.25 on CIFAR-10N, from +0.02 to −2.50 on med, and from −0.26 to −6.59 on CIFAR-100N, while on NoisyAG-News best it rises from −0.25 to +0.45. The check limits damage by giving up the same repairs the corrector makes, so its value depends on the setting. On the cleanest stream there is little to repair, and the damage limiting turns the extension's small loss into a small gain. That is the opposite of what such a check is for. The rank threshold needs no such calibration. The classifier is uncalibrated, and the share of wrong examples whose probability of the current label exceeds 0.9 runs from 0.3% on CIFAR-10N to 96.8% on CIFAR-100N, a spread an absolute threshold would render incomparable across class counts. That loss still carries information. Within the examples the corrector leaves alone, the posterior mass of the current label separates the examples whose label is wrong from the examples whose label is right at an AUROC of 0.786 on NoisyAG-News worst, 0.768 on med, 0.762 on best, 0.883 on CIFAR-10N and 0.827 on CIFAR-100N, and the margin between that mass and the best alternative class separates them equally well. Both are example-level scores that a rank threshold can use directly. Ranking the untouched examples by the margin and relabelling the lowest fraction to their best alternative has an interior optimum. Table G16 reports that the rule improves on CIFAR-10N and on the severe crowd text, is marginal on NoisyAG-News med and CIFAR-100N, and stays flat on NoisyAG-News best. The optimum fraction lies between 1 and 10% of the untouched examples. The precision of the resulting moves stays near one half, so the gain realises a small net over a large symmetric exchange, a conservative estimate of the extension's value. A trustworthy per-candidate signal would add the difference between the coverage of Table G12 and the realised repair rate. The margin, the strongest example-level score the audit already has, realises the first share of it. Two side results bound the interpretation. The label-level average of (8) and the plug-in estimate at the mean corrected latent disagree on the top class for 0.09 to 0.67% of examples (0.67 on NoisyAG-News worst, 0.33 on med, 0.23 on best, 0.26 on CIFAR-10N, 0.09 on CIFAR-100N), showing that the estimators are not identical. We report label-level marginalisation to match Equation (8). The net deltas also carry a seed variance that the means alone hide. At the operating point the delta is consistent across the five corrector seeds only on CIFAR-10N ($+0.43 \pm 0.05$, positive in five of five setting-seed combinations) and on NoisyAG-News best, where the corrector's own stream leaves the least for the extension to add (-0.25 ± 0.07, zero of five). On NoisyAG-News worst ($+0.29 \pm 0.23$) and med ($+0.02 \pm 0.27$)

Table G16: Margin rank threshold on the examples the corrector leaves alone, at the best operating point per setting (mean over five seeds); examples are ranked by the posterior margin and the lowest fraction is relabelled. Net delta is against the single-proposal rule. Bold marks the realised net. Candidate-extension study over five corrector seeds, one encoder family per data type.

Setting	best fraction	moved	gained	added	net
NoisyAG-News worst	0.05	1,610	775	621	+0.35
NoisyAG-News med	0.01	406	201	166	+0.08
NoisyAG-News best	n/a	n/a	n/a	n/a	−0.30
CIFAR-10N	0.08	2,052	980	649	+0.90
CIFAR-100N	0.01	295	77	70	+0.02

the sign varies from seed to seed, and on CIFAR-100N it is negative in four of five (-0.26 ± 0.28). Across the five settings, CIFAR-10N is the one whose gain holds its sign throughout the seed variance. The scope above is bounded in three ways. The rank threshold is calibrated on the pool rather than on the held-out validation examples, and the pool contains the examples the corrector was fitted on. The gains are therefore in-sample and may be optimistic; they do not establish out-of-sample performance. The evaluation stops at the label stream, without retraining a downstream model on the extended stream, and it therefore bounds the repair rate of the rule rather than its end-to-end value. The width and the floor were explored on the same five settings that the rest of the paper reports, with one encoder family per data type. The gap between the coverage of Table G12 at w=3, which reaches 0.91 to 0.98 on four of the five settings, and the realised gains measures what a per-candidate trust score would be worth. Example-level signals recover the share of the excluded examples that such scores can certify, and the remaining gap is what a per-candidate trust score could close. The reported estimator follows (8); the small observed disagreement does not establish the superiority of either estimator for downstream learning.

H Training-Time Pool Enlargement and Seed Robustness

H.1 Training-Time Pool Enlargement

The pool-enlargement study tests whether the rank threshold transfers from the correction audit to training-time supervision. Score-thresholded corrected examples are admitted into the semi-supervised pool of a round-1 retrain. Each seed compares four size-matched training configurations, namely no pool (V0), thresholded admission, unselective admission (a random sample of the candidate pool at the thresholded count), and uniform random admission at the same count, on AG-News with injected noise at rate 0.2, with five seeds per noise type. The proximity threshold is calibrated on the held-out validation split of Section A at $\geq 95\%$ precision. The specified acceptance rule requires thresholded admission to exceed unselective by at least 0.05 median on at least 70% of the settings. The margin is narrow, and where it is the threshold study of Section 5.2 remains the paper's central claim.

On the image side the re-measured study is positive and its margin is small. Thresholded admission exceeds unselective admission by +2.7 accuracy points averaged over the 12 image configurations (0.7173 against 0.6900). It also exceeds size-matched random admission at every coverage the study measures, by +0.8 accuracy points at 10% coverage, +1.4 at 25% and +2.0 at 50%. The margin stays below the specified bar of a median gain of five accuracy points, and the value of that gap is diagnostic.

The calibrated threshold admits 90–92% of the candidate pool at precision 0.98–1.00. These image results support a modest benefit of supervised pool selection. They are distinct from the clean-label diagnostic and label-free end-to-end comparisons.

On the text side the re-measured study covers both injected noise types over five seeds each. It returns the same ordering of the four size-matched configurations at a narrower margin, with thresholded admission first, uniform random admission next, unselective admission after it and no pool last. Thresholded admission exceeds unselective admission by +0.72 accuracy points under symmetric noise (0.9150 against 0.9078), and by +1.46 accuracy points under instance-dependent noise (0.9200 against 0.9054). Against uniform random admission at the same count the margin is +0.58 and

$+1.34$ accuracy points, and against no pool it is $+2.26$ and $+2.42$ accuracy points. Uniform random admission is within 0.14 and 0.12 accuracy points of unselective admission.

The size-matched comparisons show a small advantage for the selected pool on these text settings. The margin between thresholded and unselective admission stays below the specified bar of a median gain of five accuracy points on both noise types. The small margins delimit the observed benefit on both modalities. With the corrector repairing far more than it damages on these settings, the calibrated threshold admits almost the entire candidate pool. These results motivate testing selection at lower coverage while retaining a matched random control.

An earlier version of this appendix reported figures from a study that assigned admitted examples their ground-truth labels rather than the corrector's emitted labels. Those figures are withdrawn.

H.2 Statistical Robustness Check

The statistical robustness check concludes the experiments with a ten-seed replication of the principal results. It re-runs the two primary settings of the audit output (Section 5.1) over ten seeds on the full test slice with $K = 32$ perturbation draws. It re-examines the K reduction with three seeds at $K = 4$. The ten-seed replications preserve the reported mean ordering on these two settings. On both primary settings the recovery rate and accuracy-gain means lie within one standard deviation of the three-seed estimates in Sections E and 5.1. The scores are equally stable, with proximity and robustness holding their ordering relative to chance and plausibility staying flat near it. The robustness damage channel remains a strong damage detector on both settings.

Two calibration points follow. First, the K reduction underestimates robustness, scoring below $K = 32$ on both settings, so the measured $K = 4$ estimates are lower in these two comparisons. Second, the association between the recovery rate and confidence is a measured association rather than a guarantee. The per-seed correlation is $r = 0.61$ ($p = 0.06$) on NoisyAG-News worst and $r = 0.52$ ($p = 0.12$) on CIFAR-10N worse, consistent with the pooled correlation of Section 5.1, though not independently significant at ten seeds, and it is reported as an observed correlation rather than a tested law.

I Scope and Limitations

The association between confidence informativeness and the recovery rate is measured rather than guaranteed. Over the whole inventory the pooled correlation is $+0.94$, the per-seed values are lower, and at ten seeds they do not reach independent significance. The threshold is a conservative policy by design. Its end-to-end rule ranks all candidate edits without clean labels, whereas the diagnostic comparison uses clean-label-referenced scores on WR/RW examples (Section B.1). Its advantage over Co-teaching is observed on the crowd-labelled text streams and not on CIFAR-10N. Where the score saturates on the candidate corrections, the threshold decides how many corrections are admitted rather than which ones. The audit covers the LSNPC corrector family with the tested encoder configurations, and extending these properties further across correction families remains open. Plausibility supplies a lower-density floor; its density alone does not certify a repair or reliably identify damage. The end-to-end claim covers all six settings. The dopanim corrector is a pixel-level model at 224 px rather than the frozen-feature correctors used on the CIFAR settings. Run in the frozen-feature mode instead, the same corrector loses about seven accuracy points on that stream, showing sensitivity to the encoder and training configuration (Huang et al., 2025). The ten-seed figures quoted above come from the seed-robustness study of Section H, which lies outside the five-seed main configuration that the main evaluation uses.