TRUST NO SINGLE NOISE MODEL: CERTIFIED ACTIVE DATASET REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

A recorded label is an imperfect observation of the label we want, separated from it by a corruption process that recurs across records and can be modelled, so that a small audit budget repairs some records directly and, through the model, others indirectly. But the noisy labels alone do not identify that process: several models fit the same data and prescribe different audits and repairs, and existing methods commit to one and offer no guarantee when it is wrong. We treat the problem as a game: nature picks one corruption model from a declared set of candidates; the learner chooses audits sequentially and then corrects the unaudited labels. Solving the game yields a repaired dataset with a certificate, the largest error expected to remain under any candidate, computable without gold labels. The minimax value is the Bayes value under a least-favourable weighting of the candidates, which makes the game solvable by dynamic programming, and choosing audits one at a time can be arbitrarily worse than planning the sequence. On the CIFAR-10H test set the minimax policy attains the lowest certified risk at every budget among 23 arms, including eight published cleaners: every cleaner's certified risk stays above 5% while ours falls from 5.4% to 0.9%, at a price of about one point of true error against a Bayesian weighting.

1 INTRODUCTION

Consider a biodiversity archive with hundreds of thousands of volunteer-labelled photographs. Some species are hard to tell apart, some volunteers are more careful than others, and each photograph carries one recorded label. A curator can re-examine only a small fraction. How many labels remain wrong afterwards? The answer depends on which species volunteers confuse, how often, and whether their errors recur across photographs—patterns the recorded labels alone cannot establish.

The same structure appears wherever labels are produced at scale: crowd workers mislabel a few percent to a third of benchmark records (Wei et al., 2022), scraped product images inherit webpage tags, and screening archives retain one reading per case from readers with differing habits. Existing methods offer partial remedies: robust learning trains classifiers despite noise while leaving the dataset unchanged; detectors rank suspicious records using confident learning, training dynamics or neighbourhood votes (Northcutt et al., 2021); and active label cleaning audits the highest-ranked records (Bernhardt et al., 2022). However, no detector consistently dominates across noise regimes (Goswami et al., 2023; George et al., 2024).

The pieces share a hidden premise. Which records to audit, how to correct the unaudited ones, and how many errors remain afterwards all depend on a model of the corruption process—which classes get confused, how often, and by whom—and each existing method commits to one such model before it starts—explicitly, as an estimated class-confusion matrix (Northcutt et al., 2021), or implicitly, through the heuristic that ranks records: a single classifier's posterior (Bernhardt et al., 2022), training dynamics (Pleiss et al., 2020; Toneva et al., 2019), or a neighbourhood vote (Zhu et al., 2022). Yet one recorded label per record does not determine that model (Liu et al., 2023): "the classifier is wrong on these photographs" and "this volunteer is unreliable" fit the same archive and prescribe different audits, different repairs, and different counts of remaining errors. A method that commits to one model can still rank and repair, but whatever it implies about the errors left behind is conditional on assumptions that cannot be verified from the recorded labels alone.

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.

We propose CARD (Certified Active Repair of Datasets) to make this cost measurable. Instead of one model, CARD declares a set of candidate models—each a classifier's belief about the true labels, a pattern of which classes are confused with which, and a per-annotator noise level that every audit updates—and lets the audits weigh them. Under each candidate, the recorded labels and the audit answers determine the probability that each unaudited label is wrong, and hence the expected fraction of records still wrong after repair; the largest of these expectations over the candidates is the *certified risk*. It needs no gold labels, it is defined for any audit-and-repair method, and it is a promise conditional on the candidates: if the true corruption process is among them (*coverage*), it bounds the true residual error (Section 2). The certificate is the value of a two-player game: nature selects one candidate from the declared set and keeps it for the whole campaign; the learner chooses audits adaptively and then repairs the unaudited labels. The value of this game is the smallest certificate any policy can carry on the given candidates, and it is attained by an ordinary Bayesian policy under a least-favourable weighting of the candidates, which makes it computable (Section 3). That weighting is one of three we study; the others are the Bayesian posterior over the candidates and a tempered posterior between the two. All three report the same certificate and differ in what they optimise: the minimax weighting yields the tightest certificate and pays for it in true residual error, the posterior the lowest error with a looser certificate, and the tempered weighting the lowest error among our arms on most datasets with a certificate that covers the true error in 24 of 25 dataset–budget pairs. On CIFAR-10H, for instance, auditing 1% of the records under the minimax weighting certifies that at most 3.2% of the labels remain wrong, the tightest certificate of the 23 methods compared, while the labels actually left wrong fall from 5.4% to 2.0% (Section 4.3).

Contributions.

- **A certificate for any audit-and-repair method**: the largest expected residual error over a declared menu, computed without gold, and the coverage condition under which it bounds the true error, checked wherever gold exists (Sections 2 and 4.3).
- **Theory of the tightest certificate**: the least-favourable-law representation and its dynamic-programming solution, the bracket on the minimax value, and the greedy-versus-planning gap (Theorems 3.3, 3.4 and B.4).
- **Three weightings, priced**: one solver runs every weighting of the menu from minimax to the Bayesian posterior; we compare them on certificate and on true residual error, locate the robustness to misspecification in the menu rather than in the weighting by a preregistered experiment (Section 4.3), and give a recommendation (Section 5).
- **Evidence**: one development set and four test sets with human or web labels; 23 arms, namely the three weightings, four internal controls and eight published cleaners, each run as published and with its ranking driving our audit-and-repair step; all protocols frozen before any test result was seen (Section 4).

2 Certified repair

Let $\mathcal{D}_0$ denote the recorded dataset, an annotation table with one row per record $j \in [n]$: a feature vector x_j, a recorded label $o_j \in \mathcal{Y}$ from a finite label set $\mathcal{Y}$, and a source s_j, the annotator or group of records assumed to share one corruption process. All probabilities below are conditional on $\mathcal{D}_0$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ be the latent vector of true labels and $\boldsymbol{z} \in \mathcal{A}$ the released dataset, a repaired label vector with $\mathcal{A} = \mathcal{Y}^n$ unless stated otherwise. The release is scored by a bounded loss $\ell(\boldsymbol{Y}, \boldsymbol{z}) \in [0, 1]$, throughout the weighted Hamming loss

$$\ell(\boldsymbol{Y}, \boldsymbol{z}) = \sum_{j=1}^{n} c_j \mathbf{1}\{Y_j \neq z_j\}, \qquad c_j \geq 0, \quad \sum_j c_j = 1. \tag{1}$$

With $c_j = 1/n$ it is the fraction of released labels that are wrong, the residual noise rate of the dataset: the quantity the user of a repaired dataset asks about (Section 1) and the quantity label-cleaning studies report (Bernhardt et al., 2022). Unequal weights encode unequal record importance. The loss is separable across records, which the dynamic programme of Section 3 exploits.

Audit protocol. Repair proceeds in B rounds. In round t the learner selects an unaudited record i_t and a trusted oracle reveals its true label Y_{i_t}; the selection may depend on $\mathcal{D}_0$ and on all previous

answers. After round B the learner releases $\boldsymbol{z}$: each audited record carries its revealed label, and each unaudited record carries the label the learner infers for it from the recorded labels and the audit answers. The loss (1) is evaluated on all n records, so it measures the error left in the whole released dataset rather than the accuracy of the audit selection; audited records contribute nothing to it. Evaluated against gold labels withheld from every method, $\ell(\boldsymbol{Y}, \boldsymbol{z})$ is the *hidden-gold loss*, the true residual error of Section 1.

Candidate laws. We are given K candidate conditional laws of the true labels, the *menu*, $P_k(\mathrm{d}\boldsymbol{y}) = P_{\boldsymbol{\theta}_k}(\boldsymbol{Y} \in \mathrm{d}\boldsymbol{y} \mid \mathcal{D}_0)$, $k \in [K]$. Each is specified by $\boldsymbol{\theta} = (q, \boldsymbol{\Omega}, (a, b), r)$ and has the form

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \mathcal{D}_0) = \prod_s \int_0^1 p^r_{\boldsymbol{\theta}}(\rho \mid \mathcal{D}_0) \prod_{j:\, s_j = s} \frac{q_j(y_j)\, T_{\boldsymbol{\Omega}}(o_j \mid y_j, \rho)}{m_j(\rho)}\, \mathrm{d}\rho, \tag{2}$$

where $T_{\boldsymbol{\Omega}}(o \mid y, \rho) = (1-\rho)\mathbf{1}\{o = y\} + \rho\, \boldsymbol{\Omega}_{yo} \mathbf{1}\{o \neq y\}$ and $m_j(\rho) = \sum_{y \in \mathcal{Y}} q_j(y)\, T_{\boldsymbol{\Omega}}(o_j \mid y, \rho)$. A law has four components. The *truth model* q is a classifier: $q_j(y) = q(y \mid x_j)$ is its probability that the true label of record j is y. The *corruption channel* $T_{\boldsymbol{\Omega}}$ describes how a recorded label arises from a true one: with probability $1 - \rho$ the label is recorded correctly, otherwise it is replaced by a class drawn from row y of the *destination matrix* $\boldsymbol{\Omega}$, a row-stochastic matrix with zero diagonal. The *severity* $\rho \in [0, 1]$ is the corruption probability of a source, shared by all its records and given the prior $\mathrm{Beta}(a, b)$, so that an audit of one record informs the severity of every record of the same source. The *root* r specifies which observations update the severity: all recorded labels of the source (joint likelihood) or the audit answers alone (conditional anchor). Under a candidate, the recorded label o_j is an observation of Y_j through the channel, and the factor $q_j(y_j) T_{\boldsymbol{\Omega}}(o_j \mid y_j, \rho)/m_j(\rho)$ in (2) is the resulting posterior of Y_j at severity ρ.

Policies and certified risk. Let B be the audit budget. The history after t audits is $\boldsymbol{h}_t = ((i_1, Y_{i_1}), \ldots, (i_t, Y_{i_t}))$, the audited indices with their revealed labels, with $\boldsymbol{h}_0 = \varnothing$. A deterministic policy is a decision tree π that maps every reachable history $\boldsymbol{h}_t$, $t < B$, to an unaudited index and every terminal history $\boldsymbol{h}_B$ to a release $\boldsymbol{z}_\pi(\boldsymbol{h}_B) \in \mathcal{A}$; let Π_B denote the finite set of such trees and $\Delta(\Pi_B)$ the set of probability distributions on it. A randomized policy $\mu \in \Delta(\Pi_B)$ draws one tree from μ before the first audit. Let $\boldsymbol{H}_B(\boldsymbol{Y})$ denote the terminal history that π generates when the true labels are $\boldsymbol{Y}$. The risk of π under candidate k is $R_k(\pi) = \mathbb{E}_{\boldsymbol{Y} \sim P_k}\big[\ell\{\boldsymbol{Y}, \boldsymbol{z}_\pi(\boldsymbol{H}_B(\boldsymbol{Y}))\}\big]$, and $R_k(\mu) = \mathbb{E}_{\pi \sim \mu} R_k(\pi)$. The *certified risk* of μ relative to the menu is

$$\mathcal{C}(\mu) := \max_{k \in [K]} R_k(\mu), \tag{3}$$

the largest expected residual error of the released dataset over the candidate laws. It is the certificate this paper attaches to a policy: it is computed from $\mathcal{D}_0$ and the audit answers without gold labels, it is defined for every policy, and it is the objective that the policies of Section 3 minimize. Equation (3) is an ex ante quantity; every table reports its ex post counterpart for the released $\boldsymbol{z}$ given the transcript $\boldsymbol{h}_B$, $\mathcal{C}(\boldsymbol{z} \mid \boldsymbol{h}_B) := \max_k \mathbb{E}_{P_k}[\ell(\boldsymbol{Y}, \boldsymbol{z}) \mid \boldsymbol{h}_B]$.

Validity. Let $P^\dagger$ be the true conditional law of $\boldsymbol{Y}$ given $\mathcal{D}_0$ and $R_P(\mu) = \mathbb{E}_{\pi \sim \mu} \mathbb{E}_{\boldsymbol{Y} \sim P}[\ell\{\boldsymbol{Y}, \boldsymbol{z}_\pi(\boldsymbol{H}_B(\boldsymbol{Y}))\}]$ the risk of μ under a law P, so that $R_k = R_{P_k}$. This risk is linear in P, so if $P^\dagger$ lies in the convex hull of $\{P_1, \ldots, P_K\}$, then $R_{P^\dagger}(\mu) \leq \mathcal{C}(\mu)$ for every policy: the certificate is exact relative to the menu and valid whenever the menu *covers* the true law. Coverage is a property of the menu alone; no weighting can supply it, and the recorded labels cannot verify it. Where gold labels exist, a necessary condition is checked after the fact, the certified risk lying at or above the realised hidden-gold loss, and every certificate in this paper is subjected to that check (Table 2). Proposition B.3 gives the general form: if a data-dependent set of laws contains $P^\dagger$ with probability at least $1 - \alpha$, the certificate bounds the true conditional risk on that event, and the unconditional true risk exceeds the expected certificate by at most α.

Section 3 turns to policies. Under a fixed weighting $\boldsymbol{\lambda}$ of the candidates the audit problem is Bayesian sequential design under the mixture law $\sum_k \lambda_k P_k$, which dynamic programming solves exactly; the policies we compare differ only in $\boldsymbol{\lambda}$: the least-favourable weighting, which attains the smallest certified risk any policy can carry (minimax); the posterior over the candidates given the audit answers (posterior); and a tempered posterior between the two (tempered). All three report the certificate (3); they differ only in which laws the audits are planned against and the unaudited labels repaired under.

3 AUDIT POLICIES: MINIMAX, POSTERIOR AND TEMPERED WEIGHTINGS

Every CARD policy runs the same solver: audits are planned and the unaudited labels repaired under a mixture of the menu's laws, and only the weights of that mixture differ between policies. Let Δ_K denote the probability simplex on $[K]$. For $\boldsymbol{\lambda} \in \Delta_K$, let $P_{\boldsymbol{\lambda}} = \sum_{k=1}^{K} \lambda_k P_k$ be the mixture law and $V_B(\boldsymbol{\lambda}) := \min_{\pi \in \Pi_B} \sum_{k=1}^{K} \lambda_k R_k(\pi)$ its Bayes audit value, the smallest expected loss under $P_{\boldsymbol{\lambda}}$ that a policy with budget B can achieve.

Bayes best response. Fix $\boldsymbol{\lambda}$. For a history $\boldsymbol{h}$ of any length, let $S(\boldsymbol{h}) \subset [n]$ be the set of records not audited in $\boldsymbol{h}$, let $P_{\boldsymbol{\lambda}}(\cdot \mid \boldsymbol{h})$ be the posterior of $\boldsymbol{Y}$ under $P_{\boldsymbol{\lambda}}$ given the answers in $\boldsymbol{h}$, and let $\boldsymbol{h} \oplus (i, a)$ be $\boldsymbol{h}$ extended by the audit of record i with answer a. Let $W_b^{\boldsymbol{\lambda}}(\boldsymbol{h})$ denote the smallest expected loss attainable from history $\boldsymbol{h}$ with b audits remaining. It is given by the backward recursion, for $b = 1, \ldots, B$,

$$W_0^{\boldsymbol{\lambda}}(\boldsymbol{h}) := \min_{\boldsymbol{z} \in \mathcal{A}} \mathbb{E}_{P_{\boldsymbol{\lambda}}}[\ell(\boldsymbol{Y}, \boldsymbol{z}) \mid \boldsymbol{h}], \tag{4}$$

$$W_b^{\boldsymbol{\lambda}}(\boldsymbol{h}) := \min_{i \in S(\boldsymbol{h})} \sum_{a \in \mathcal{Y}} P_{\boldsymbol{\lambda}}(Y_i = a \mid \boldsymbol{h})\, W_{b-1}^{\boldsymbol{\lambda}}\{\boldsymbol{h} \oplus (i, a)\}, \tag{5}$$

where ℓ is the loss of Eq. (1).

Proposition 3.1 (Exact full-horizon best response)**.** *$V_B(\boldsymbol{\lambda}) = W_B^{\boldsymbol{\lambda}}(\varnothing)$, and an index attaining the minimum in* (5) *at every history defines a Bayes-optimal adaptive audit tree. Under weighted Hamming loss* (1)*,*

$$W_0^{\boldsymbol{\lambda}}(\boldsymbol{h}) = \sum_{j=1}^{n} c_j \left[1 - \max_{a \in \mathcal{Y}} P_{\boldsymbol{\lambda}}(Y_j = a \mid \boldsymbol{h})\right]. \tag{6}$$

The recursion is standard finite-horizon Bayesian experimental design (Huan & Marzouk, 2016; Shen & Huan, 2023). With one law and $B = 1$ it is the nominal one-step rule (NOM-GREEDY), which a one-law menu reproduces to four decimals (Table 8).

Weightings. Let $\boldsymbol{w}(\boldsymbol{h}) \in \Delta_K$ denote the posterior over the candidates given the audit answers, $w_k(\boldsymbol{h}) \propto P_k(\boldsymbol{h})$ from a uniform prior, and let $\boldsymbol{r}(\boldsymbol{h}) = (r_1, \ldots, r_K)$ be the expected residual error of the current plan under each candidate. The policies compared in this paper set the weights as follows. The *posterior* weighting takes $\boldsymbol{\lambda} = \boldsymbol{w}(\boldsymbol{h})$, Bayesian model averaging over the menu (CARD-posterior). The *tempered* weighting tilts the posterior toward the candidates under which the current plan does worst, $\lambda_k \propto w_k(\boldsymbol{h}) \exp\{r_k(\boldsymbol{h})/\tau\}$ with a fixed temperature $\tau > 0$ (CARD-tempered). The *minimax* weighting is the least-favourable $\boldsymbol{\lambda}$ of Theorem 3.3 below, computed by the iterative scheme of Eq. (9) (CARD-minimax). Fixed equal weights (CARD-uniform) serve as the control. The three weightings lie on one path: the tempered tilt is the worst case over a Kullback–Leibler ball around the posterior, whose radius grows from zero (the posterior weighting) to the whole simplex (the minimax weighting) as τ decreases (Proposition B.2). The weighting determines which laws the audits are planned against and which laws the unaudited labels are repaired under; it does not change what is certified. Every policy's certificate is the maximum (3) over the whole menu, and the minimax weighting attains the smallest value that maximum can take.

The minimax weighting. The least-favourable weighting is defined through the game between the learner and nature.

Definition 3.2. The *minimax value* $V_B^\star$ of the audit game with budget B is the smallest certified risk attainable by a randomized policy on the menu,

$$V_B^\star := \min_{\mu \in \Delta(\Pi_B)} \max_{k \in [K]} R_k(\mu) = \min_{\mu \in \Delta(\Pi_B)} \mathcal{C}(\mu), \tag{7}$$

and a *minimax policy* is a μ attaining it. Nature's candidate k is chosen once, before the first audit, and is not observed by the learner, who observes only the audit answers, generated under P_k.

The adversary is static: a candidate that could change between audits would make this a rectangular robust control problem (Iyengar, 2005; Nilim & El Ghaoui, 2005), certifying against corruptions that no single process generates. $V_B^\star$ is the tightest certificate any policy can carry on the menu. The following representation makes it computable.

Theorem 3.3 (Least-favourable-law representation)**.** *For finite label space, audit pool, budget and action set, and loss in* $[0, 1]$*,*

$$V_B^\star = \max_{\boldsymbol{\lambda} \in \Delta_K} V_B(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda} \in \Delta_K} \min_{\pi \in \Pi_B} \mathbb{E}_{\boldsymbol{Y} \sim P_{\boldsymbol{\lambda}}} \ell\{\boldsymbol{Y}, \boldsymbol{z}_\pi(\boldsymbol{H}_B)\}. \tag{8}$$

Moreover, a saddle pair $(\mu^\star, \boldsymbol{\lambda}^\star)$ *exists;* $\boldsymbol{\lambda}^\star$ *is a least-favourable mixture of the candidate laws.*

The identity is classical (Sion, 1958; Berger, 1985; Lehmann & Casella, 1998); what is specific to our setting is the object mixed, complete conditional laws of the true labels, so that the inner problem is ordinary sequential design under $P_{\boldsymbol{\lambda}}$ and Proposition 3.1 solves it. Theorem 3.3 identifies the value, not the policy: in Example B.5 either pure query has certified risk $1/4$ and their half–half mixture $1/8$, so a minimax policy may mix several Bayes best responses.

Computing the least-favourable weighting. Proposition 3.1 solves the inner problem $V_B(\boldsymbol{\lambda})$ at fixed weights; the outer maximization in (8) is carried out by Hedge (Freund & Schapire, 1997; 1999). At round t, compute an exact best-response tree, evaluate its risk vector $\boldsymbol{R}(\pi) = (R_1(\pi), \ldots, R_K(\pi))$, and tilt the weights toward the candidates on which that tree does worst:

$$\pi_t \in \arg\min_{\pi \in \Pi_B} \boldsymbol{\lambda}_t^\top \boldsymbol{R}(\pi), \qquad \lambda_{t+1,k} = \frac{\lambda_{t,k} \exp\{\eta R_k(\pi_t)\}}{\sum_m \lambda_{t,m} \exp\{\eta R_m(\pi_t)\}}. \tag{9}$$

Let μ_T be the uniform mixture over $\pi_1, \ldots, \pi_T$, let $\bar{\boldsymbol{\lambda}}_T = T^{-1} \sum_t \boldsymbol{\lambda}_t$, and define

$$L_T := V_B(\bar{\boldsymbol{\lambda}}_T), \qquad U_T := \max_k \frac{1}{T} \sum_{t=1}^{T} R_k(\pi_t). \tag{10}$$

Theorem 3.4. *With exact best responses, exact risk vectors, and* $R_k(\pi) \in [0, 1]$*,*

$$L_T \le V_B^\star \le U_T, \qquad U_T - L_T \le \frac{\log K}{\eta T} + \frac{\eta}{8}. \tag{11}$$

For $\eta = \sqrt{8 \log K / T}$*, the gap is at most* $\sqrt{\log K / (2T)}$*.*

$U_T = \mathcal{C}(\mu_T)$ is the certified risk (3) of the returned policy; L_T is nature's value at $\bar{\boldsymbol{\lambda}}_T$, below which no policy can be certified on this menu; the gap between them is optimization slack, not statistical error, and shrinks with the number of rounds. Corollary B.1 extends the bound to an ε-approximate best response and a risk vector known to within δ (excess $\varepsilon + 2\delta$); with estimated risk vectors, as from particle posteriors (Chopin, 2002; Del Moral et al., 2006), only the upper certificate survives.

Dataset scale. The exact tree of Proposition 3.1 is computable for one source's records, not for a dataset. There every CARD policy runs as a receding-horizon planner: the exact one-step calculation of Proposition 3.1 ($B = 1$) under the current weights, one committed audit, a weight update by the rule of its weighting (for minimax the Hedge step (9) on the current plan's risk vector), and re-planning; after the last audit the unaudited labels are repaired under the terminal mixture (for minimax the average of the weight path) and the ex post certified risk is reported (Section 2). A receding-horizon step is not an exact best response, so the bracket does not carry over: at dataset scale the certificate is an upper bound without a lower bound.

4 Experiments

Data. We use five image-classification datasets whose recorded labels are noisy and for which gold labels exist and are withheld from every method. CIFAR-10N (Wei et al., 2022) provides one crowd label per image for the CIFAR-10 training images in three labelling rounds, with worker identifiers; CIFAR-10H (Peterson et al., 2019) provides crowd labels for the CIFAR-10 test images, with annotator identifiers; Dopanim (Herde et al., 2024) provides crowd labels from 20 annotators for images of 15 animal species; LabelMe-AMT (Rodrigues & Pereira, 2018) provides crowd labels from 59 workers for 1,000 scene images of 8 classes; and Clothing1M (Xiao et al., 2015) provides web-scraped labels for clothing images of 14 classes together with a clean-labelled subset. Each dataset is an annotation table of (image, annotator) records with one recorded label each; Clothing1M

carries no annotator identity, so its records are grouped into 20 sources by clustering image features. CIFAR-10N round 1 is the development set, on which every design decision was taken; rounds 2 and 3, which relabel the same images with other workers, replicate it. The other four datasets are test sets. Sizes, noise rates and audit budgets (Table 3), access, provenance and preprocessing are given in Appendix C.1.

Evaluation. We report two quantities per run, both as a percentage of the records in the released dataset. The *realised error* is the share of released labels that disagree with the withheld gold labels (Eq. (1) with $c_j = 1/n$): an entry of 15.4 means that 15.4% of the records are still wrong after repair. The *certified error* is the ex post certified risk of Eq. (3), computed from the recorded labels and the audit answers without gold: an entry of 15.4 means that the method certifies that, under every law in the menu, at most 15.4% of the records are expected to be wrong. A certificate covers when it is at or above the realised error; this check is reported for every arm and budget (Table 2). Differences between arms are paired by block (worker, annotator or source cluster) and given as 95 % block-bootstrap intervals with a sign-flip permutation p-value (Appendix C.5).

4.1 BASELINES

We compare 23 arms in three groups.

- **Published cleaners** (16 arms). Eight methods: ALC (Bernhardt et al., 2022), Confident Learning (Northcutt et al., 2021), ACTIVELAB (Goh & Mueller, 2023), SIMIFEAT (Zhu et al., 2022), CINCER (Teso et al., 2021), AUM (Pleiss et al., 2020), FORGETSCORES (Toneva et al., 2019) and SMALLLOSS (Han et al., 2018), each run in two modes. *As published*: the cleaner selects the audits by its own ranking and disposes of the unaudited rows by its own rule (Confident Learning and SIMIFEAT rewrite them; the other six keep the recorded label and so remove at most B errors). *On our repair*: only the cleaner's ranking is used, as the audit rule, and the audit answers are handed to the mixture repair of Section 3, so the contrast isolates audit placement and the arm carries a certificate. Provenance and adaptation of every method: Appendix E.
- **Internal controls** (4 arms). Random audits (RAND); most-probably-corrupted (MPC), which audits the record with the largest error probability under the nominal law; the nominal one-step rule (NOM-GREEDY), one law and a one-step rule, repairing under that law; and CARD-uniform, the planner of Section 3 with fixed equal weights.
- **CARD** (3 arms). The posterior (CARD-posterior), tempered (CARD-tempered, $\tau = 0.1$) and minimax (CARD-minimax) weightings of Section 3.

4.2 PROTOCOL

Budgets and scale. Every arm audits $\lceil \rho n \rceil$ records, $\rho \in \{0.25, 0.5, 1, 2, 5\}\%$, and is run at dataset scale by the receding-horizon planner of Section 3, with the annotators or sources of Table 3 as blocks. The exact game is evaluated separately at the scale of one annotator's records ($B = 2$ audits over eight records) in Appendix C.6.

Menu. Every menu has the same skeleton: candidate truth models crossed with two destination matrices (soft class-confusion and uniform), two severity priors (Beta(2,8) and flat) and two roots; two truth models give the default sixteen-law menu, with DINOv2 / soft / Beta(2,8) / joint as the nominal law. The truth models are chosen per dataset from what is known about its labels: out-of-fold probes on the recorded labels on CIFAR-10N, Dopanim and LabelMe-AMT; probes trained on the disjoint CIFAR-10 training labels on CIFAR-10H; and on Clothing1M a third truth model, a probe fit on the dataset's clean-labelled images outside our sample. How coverage and tightness respond to menu size is measured with eight- and 24-law variants (Section 4.3(c)). Appendix C.2 lists every menu; Section 5 gives guidelines for choosing one.

Design decisions. The menu, the temperature, the planner constants and the statistical procedures were fixed on the development set. Each test set was run once, under a protocol written before any of its results existed, and analysed as written; the protocol documents and the record of every deviation are in the supplement.

Table 1: Results at 1% of the rows audited, in % of rows: error left after repair and certified error. Bold: lowest error and tightest covering certificate among the columns shown; a certificate below the error above it does not cover. Development row: mean of CIFAR-10N rounds 1–3. All arms and budgets: Appendix D.

dataset	% of rows	nominal one-step	posterior	tempered	minimax	ALC as published	SIMIFEAT as published
CIFAR-10N (dev.)	error left	14.4	4.6	4.3	4.3	16.6	**2.8**
	certified	20.8	15.2	14.8	**14.8**	22.0	18.4
CIFAR-10H	error left	1.1	**1.0**	1.0	2.0	4.4	1.0
	certified	8.1	6.9	6.7	**3.2**	9.2	7.8
Dopanim	error left	27.6	**22.3**	23.4	24.0	32.5	26.2
	certified	36.9	29.4	**23.8**	22.1	40.2	43.3
LabelMe-AMT	error left	23.3	19.0	**17.9**	20.1	24.9	18.1
	certified	24.6	**19.6**	21.3	20.4	25.7	24.7
Clothing1M	error left	36.4	20.9	**20.7**	27.8	36.7	37.4
	certified	27.1	34.0	36.2	**28.8**	27.9	35.5

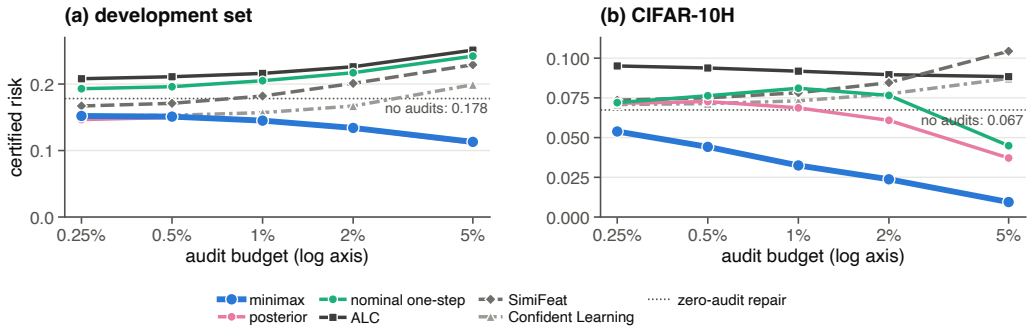

Figure 1: Certified error against audit budget (log axis) on (a) the development set and (b) CIFAR-10H for six arms; dotted: the zero-audit mixture repair. The remaining arms: Appendix D.

Experiments. (i) The 23-arm grid on the five datasets at five budgets (Section 4.3); (ii) its replication on CIFAR-10N rounds 2 and 3 (Appendix D.1); (iii) a menu-size ablation with eight-, sixteen- and 24-law menus on the development set and sixteen- and 24-law menus on CIFAR-10H and Dopanim (Section 4.3(c)); (iv) a semi-synthetic study of planning depth; and (v) a semi-synthetic noise-rate sweep (both in the supplement).

4.3 RESULTS AND DISCUSSION

(a) Tightness of the certificate. Yes, and the minimax weighting makes it the tightest. On CIFAR-10H its certified error is the lowest of the 23 arms at every budget, with all 110 paired intervals below zero; it falls from 5.4% before any audit to 3.2% at 1% and 0.9% at 5%, where no published cleaner certifies below 5.2% in either mode (Appendix D). On every other dataset the tightest covering certificate at 1% also belongs to an arm that repairs with the mixture of Section 3 (Table 1), and on Clothing1M every certificate a published cleaner reports lies below its own error, whereas all three of ours cover. Figure 1 shows why a suspicion score is not a certificate: for every suspicion-ranked arm on the development set, and for eight of eighteen on CIFAR-10H, the certified error rises as more records are audited, and at 5% every one of them certifies at least four times ours. Each discovered error is extrapolated by the most pessimistic candidate, and a method that rewrites under one law does so ever more confidently.

(b) Price of tightness and gain of the repair. Figure 2 places every arm by error and certificate; on every dataset one of our three weightings sits on the lower-left frontier. Against the published cleaners the gains are large. At 1,024 audits on CIFAR-10H the minimax release leaves 2.0% of the records wrong against 4.4% for ALC as published (-0.024 $[-0.028, -0.021]$ over 512 worker blocks, the preregistered primary contrast), replicated in sign on Dopanim; ALC's ranking on our repair recovers most of that margin, so the gain is mostly repair. On Clothing1M the tempered weighting leaves

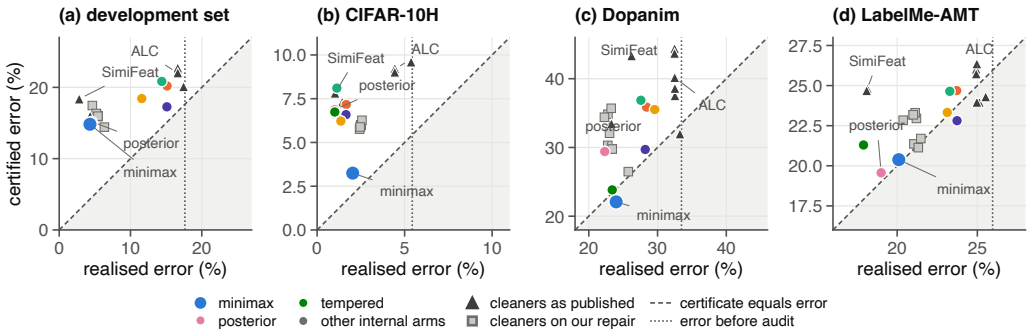

Figure 2: Realised error after repair against certified error, in % of rows, at 1% of the rows audited. Above the dashed diagonal the certificate covers the error; dotted vertical line: error before any audit. Labelled: the three weightings, SIMIFEAT and ALC as published.

20.7% wrong against 36.7% for the best published cleaner, and on LabelMe-AMT it has the lowest error of all 23 arms. Among our weightings the certificate has a price. On CIFAR-10H, where the nominal law is already right, the minimax weighting leaves twice the error of the posterior (2.0% against 1.0%) for a certificate 2.1× tighter at 1% and 4× at 5%; on Dopanim it pays 1.7 points and on Clothing1M 6.9, for certificates 7.3 and 5.2 points tighter. The tempered weighting matches the posterior's error on CIFAR-10H, beats it on LabelMe-AMT and Clothing1M, and has the tighter certificate on three of five datasets. The lowest error on the development set belongs to SIMIFEAT as published (2.8%), which rewrites every row by a majority vote over annotations of the same image, an information source our per-record arms do not use, at a certificate 3.6 points looser than ours; the secondary contrasts are in Appendix D.

(c) Validity of the certificate. Tightness and validity are read together in Table 2. The tempered and posterior certificates cover the realised error in 24 of the 25 dataset–budget pairs each, the minimax certificate in 18: at every budget on the development set and CIFAR-10H, and in seven of the 15 pairs on the other three datasets it falls short, by at most 1.9 points. Validity is a property of the menu, and the menu-size ablation measures it directly. On the development set, removing the conditional-anchor family leaves an eight-law menu under which every weighting collapses to the same error, 14.3% against 3.6–4.5% with sixteen laws (Figure 4); adding a kNN truth-model family gives a 24-law menu under which the minimax certificate covers at every Dopanim budget instead of two, at 1.7–3.4 points of certificate (Appendix D.3). The weighting cannot substitute for the menu: the ablation's preregistered difference-in-differences, -0.004 $[-0.007, -0.001]$ with the same sign on rounds 2 and 3, shows the least-favourable weighting losing slightly more than the posterior when the menu does not cover. A wider menu buys validity with certificate width, which is why the menu is declared per dataset (Section 4.2) and coverage is checked wherever gold exists.

(d) Source of the gain. From the menu and the planner together. On the development set the zero-audit mixture repair alone takes the error from 17.2% to 13.1%, and 109 game-placed audits reach 6.5% where 109 audits placed by the nominal rule with the same repair reach 12.8% (Table 8); rounds 2 and 3 replicate all 45 paired contrasts (Table 7). Under injected corruption of three kinds the minimax certificate is below every suspicion-ranked arm's in 107 of 108 cases (supplement), and the greedy-versus-planning gap of Theorem B.4 is realised on real images when the truth model cannot see the structure the audits must uncover (supplement). One limit: at the two smallest LabelMe-AMT budgets the tempered and posterior certificates are tighter than the minimax one, with intervals excluding zero; with almost no evidence the least-favourable weighting over-hedges.

5 CONCLUSION, LIMITATIONS AND FUTURE WORK

Conclusion. CARD treats budgeted dataset repair as a game against a declared menu of candidate laws and returns a repaired dataset with a certified risk, the largest expected residual error over the menu, computed without gold. On CIFAR-10H the minimax weighting gives the tightest certificate of 23 arms at every budget (5.4% to 3.2% at 1% and 0.9% at 5%) for about one point more true error than the posterior; the tempered weighting has the lowest error of our arms on most datasets,

Table 2: The three weightings at every budget: error left after repair and certified error, in % of rows. Bold, per budget within a dataset: lowest error, and tightest certificate among those that cover the error.

dataset	setting	0.25%		0.5%		1%		2%		5%	
		error	cert.	error	cert.	error	cert.	error	cert.	error	cert.
CIFAR-10N (dev.)	minimax	6.5	15.7	5.3	15.3	4.3	**14.8**	3.4	**13.9**	2.0	11.5
	posterior	5.2	**15.4**	5.1	15.4	4.6	15.2	3.3	14.0	**1.9**	**11.5**
	tempered	**5.0**	**15.4**	**5.0**	**15.3**	**4.3**	14.8	**3.3**	14.0	**1.9**	**11.5**
CIFAR-10H	minimax	2.0	**5.4**	2.1	**4.4**	2.0	**3.2**	1.8	**2.4**	0.7	**0.9**
	posterior	**1.5**	7.2	**1.2**	7.3	**1.0**	6.9	**0.7**	6.1	**0.3**	3.7
	tempered	1.5	7.1	**1.2**	7.1	1.0	6.7	0.7	6.1	**0.3**	3.7
Dopanim	minimax	26.0	27.6	25.1	24.4	24.0	22.1	22.9	22.5	19.8	**21.9**
	posterior	24.5	31.6	**24.0**	**29.0**	**22.3**	29.4	**21.6**	26.9	**19.4**	25.3
	tempered	**24.4**	**27.0**	25.6	24.2	23.4	**23.8**	21.9	**26.0**	20.0	24.4
LabelMe-AMT	minimax	22.4	22.5	21.8	21.6	20.1	20.4	19.0	18.5	16.3	15.9
	posterior	20.4	22.4	21.0	21.7	19.0	**19.6**	**17.9**	**19.2**	16.1	**19.2**
	tempered	**19.4**	**20.2**	**18.0**	**19.9**	**17.9**	21.3	18.1	20.0	**15.9**	19.5
Clothing1M	minimax	29.5	28.3	28.6	**29.4**	27.8	**28.8**	27.3	**30.5**	25.5	**28.0**
	posterior	32.6	24.7	21.4	33.3	20.9	34.0	20.1	34.9	**18.2**	34.2
	tempered	**22.1**	**33.3**	**21.0**	36.9	**20.7**	36.2	**19.9**	35.0	18.2	34.5

covers in 24 of 25 dataset–budget pairs, and on Clothing1M leaves 20.7% wrong against 36.7% for the best published cleaner. Where a bound must be signed, use the minimax weighting; otherwise the tempered weighting with its certificate reported.

Limitations. The limitation is coverage: the certificate bounds the true error only if the menu covers the true law, which the recorded labels cannot verify. Here coverage moved with the menu, not the weighting: eight laws collapsed every weighting to the same error, 24 widened coverage at the price of certificate width, and the minimax certificate under-covered in seven of 25 pairs on three test sets, by at most two points (Section 4.3).

Future work. Possible directions include a data-driven menu with a coverage guarantee of the form of Proposition B.3, exact planning beyond one step at dataset scale, records with several labels per item, and modalities beyond images.

AI USE STATEMENT

We used generative AI tools to assist with software development, primarily the integration of baseline methods, and to identify relevant baseline literature. We also used these tools to improve the grammar, conciseness and flow of the manuscript. No datasets, labels or annotations were generated using these tools. The authors reviewed all AI-assisted code and text and take responsibility for the final content of this work, including its code, claims and artifacts.

ETHICS STATEMENT

All datasets used are public or were obtained under their stated access conditions: CIFAR-10N, CIFAR-10H, Dopanim and LabelMe-AMT are publicly released with their original crowd-worker identifiers, and Clothing1M was obtained from its authors by e-mail and is not redistributed with this submission. No new human annotation was collected, and the worker identifiers we use are the anonymous ones already present in the public releases. Third-party baseline code is used under its own licences, recorded per method in Appendix E.

Reproducibility statement

Proofs are in Appendix B. Appendix C gives every dataset with its source, access conditions and preprocessing, the candidate laws and the menu of every dataset, the dataset-scale protocol, the planner, the statistical procedures and the exact-game evaluation; Appendix D holds every result table behind the main text; Appendix E records the paper, repository, commit, licence and adaptation of each published cleaner; and Appendix F the environments, seeds and runtimes. Every number in the paper is generated from the released result summaries by the scripts shipped with the code. The supplementary zip contains the anonymised code, the protocol documents with SHA-256 manifests of inputs and code, and the supplement with the complete tables, the two semi-synthetic studies and the numerical audit of the theory. Clothing1M was obtained from its authors and is not redistributed; the other datasets are public.

References

Hany Abdulsamad, Sahel Iqbal, Christian A. Naesseth, Takuo Matsubara, and Adrien Corenflos. Maximin robust Bayesian experimental design. *arXiv preprint arXiv:2603.14094*, 2026. URL https://arxiv.org/abs/2603.14094.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, New York, 2 edition, 1985. doi: 10.1007/978-1-4757-4286-2.

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.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Xi Chen, Qihang Lin, and Dengyong Zhou. Statistical decision making for optimal budget allocation in crowd labeling. *Journal of Machine Learning Research*, 16:1–46, 2015.

Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Alexander Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. doi: 10.1214/20-AOS2004.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1–2):79–103, 1999. doi: 10.1006/game.1999.0738.

Charlie Frogner, Sebastian Claici, Edward Chien, and Justin Solomon. Incorporating unlabeled data into distributionally robust learning. *Journal of Machine Learning Research*, 22(56):1–46, 2021. URL https://jmlr.org/papers/v22/19-1023.html.

Thomas George, Pierre Nodet, Alexis Bondu, and Vincent Lemaire. Mislabeled examples detection viewed as probing machine learning models: concepts, survey and extensive benchmark. *Transactions on Machine Learning Research*, 2024.

Jinwoo Go and Tobin Isaac. Robust expected information gain for optimal Bayesian experimental design using ambiguity sets. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 728–737. PMLR, 2022. URL https://proceedings.mlr.press/v180/go22a.html.

Hui Wen Goh and Jonas Mueller. ActiveLab: Active learning with re-labeling by multiple annotators. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. URL https://arxiv.org/abs/2301.11856. arXiv:2301.11856.

Mononito Goswami, Vedant Sanil, Arjun Choudhry, Arvind Srinivasan, Chalisa Udompanyawit, and Artur Dubrawski. AQuA: A benchmarking tool for label quality assessment. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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*, volume 37, 2024. doi: 10.52202/079017-1617.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Xun Huan and Youssef M. Marzouk. Sequential Bayesian optimal experimental design via approximate dynamic programming. *arXiv preprint arXiv:1604.08320*, 2016. URL https://arxiv.org/abs/1604.08320.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129.

Youjin Jeon, Kyusik Cho, Suhan Woo, and Euntai Kim. A^2LC: Active and automated label correction for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5296–5304, 2026.

Bidur Khanal, Tianhong Dai, Binod Bhattarai, and Cristian Linte. Active label refinement for robust training of imbalanced medical image classification tasks in the presence of high label noise. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Lecture Notes in Computer Science, pp. 37–47. Springer, 2024. doi: 10.1007/978-3-031-72120-5_4.

Hoyoung Kim, Sehyun Hwang, Suha Kwak, and Jungseul Ok. Active label correction for semantic segmentation with foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23924–23940. PMLR, 2024. URL https://proceedings.mlr.press/v235/kim24g.html.

Jan Kremer, Fei Sha, and Christian Igel. Robust active label correction. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 308–316, 2018.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, New York, 2 edition, 1998. doi: 10.1007/b98854.

Xiao Lin, Zeyu Rong, Yan Li, Qizhe Yang, Ping Li, and Wei Huang. A two-stage active cleaning strategy for long-tail label noise. *Neural Networks*, 198:108585, 2026. doi: 10.1016/j.neunet.2026.108585.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21475–21496. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23g.html.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Will Cukierski, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Raje, Max Bartolo, Evan Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Addison Howard, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, D. Sculley, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. DataPerf: Benchmarks for data-centric AI development. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/112db88215e25b3ae2750e9eefcded94-Abstract-Datasets_and_Benchmarks.html.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018. doi: 10.1007/s10107-017-1172-1.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9617–9626, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Peterson_Human_Uncertainty_Makes_Classification_More_Robust_ICCV_2019_paper.html.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17044–17056. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c6102b3727b2a7d8b1bb6981147081ef-Abstract.html.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.09344.

Filipe Rodrigues and Francisco C. Pereira. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 1611–1618, 2018. doi: 10.1609/aaai.v32i1.11506.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 441–448, 2001.

Wanggang Shen and Xun Huan. Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning. *Computer Methods in Applied Mechanics and Engineering*, 416:116304, 2023. doi: 10.1016/j.cma.2023.116304.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171.

Roubing Tang, Sabina J. Sloman, and Samuel Kaski. Representative, informative, and de-amplifying: Requirements for robust Bayesian active learning under model misspecification. In *International Conference on Artificial Intelligence and Statistics*, 2026. URL https://arxiv.org/abs/2506.07805. arXiv:2506.07805.

Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, and Andrea Passerini. Interactive label cleaning with example-based explanations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12966–12977. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6c349155b122aa8ad5c877007e05f24f-Abstract.html.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.05159.

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 *International Conference on Learning Representations*, 2022.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5286–5295. PMLR, 2018. URL https://proceedings.mlr.press/v80/wong18a.html.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2691–2699, 2015. doi: 10.1109/CVPR.2015.7298885.

Mohamed Yakout, Ahmed K. Elmagarmid, Jennifer Neville, Mourad Ouzzani, and Ihab F. Ilyas. Guided data repair. *Proceedings of the VLDB Endowment*, 4(5):279–289, 2011.

Dixian Zhu, Zhe Li, Xiaoyu Wang, Boqing Gong, and Tianbao Yang. A robust zero-sum game framework for pool-based active learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 517–526. PMLR, 2019. URL https://proceedings.mlr.press/v89/zhu19a.html.

Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27412–27427. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhu22a.html.

A Related work

Learning with noisy labels. Robust training fits a classifier through the noise by small-loss selection (Han et al., 2018), a transition model from true to recorded label (Xiao et al., 2015), or per-annotator confusions learned with the classifier (Rodrigues & Pereira, 2018). It establishes that a classifier can be trained despite noisy labels; it releases no repaired dataset and bounds no residual error, the quantity CARD certifies.

Label-error detection and active label cleaning. Detectors rank records by suspicion: a confusion matrix estimated from one classifier's posterior (Northcutt et al., 2021), training dynamics such as margin and forgetting (Pleiss et al., 2020; Toneva et al., 2019), neighbourhood votes over features (Zhu et al., 2022), or margins with explanations (Teso et al., 2021). Active cleaning audits from the top of such a ranking, as guided data repair (Yakout et al., 2011), active label correction (Kremer et al., 2018; Kim et al., 2024; Jeon et al., 2026), active label cleaning (Bernhardt et al., 2022), multi-annotator re-labelling (Goh & Mueller, 2023), or schemes scored by downstream performance rather than residual error (Khanal et al., 2024; Lin et al., 2026). Benchmarks pose the task (Mazumder et al., 2023) and find no detector dominating across noise regimes (Goswami et al., 2023; George et al., 2024). Each method commits to one noise model and returns a ranking; none bounds the residual error of the release. CARD accepts any such ranking as its audit rule (Section 4) and adds the certificate.

Crowdsourcing and annotator models. With several labels per item, annotator models estimate per-worker error rates and aggregate (Dawid & Skene, 1979; Rodrigues & Pereira, 2018), budgeted collection is Bayesian allocation over items (Chen et al., 2015), and the crowd datasets we use record annotator identity (Peterson et al., 2019; Wei et al., 2022; Herde et al., 2024). Aggregation shows that redundant labels can identify annotator noise and is the zero-audit alternative our per-record protocol does not use (Section 4.2). With one label per record the transition matrix is not identifiable without further structure (Liu et al., 2023), so the noise model is assumed; CARD declares a menu of such assumptions and certifies against all of them.

Sequential design and the game. Under one law, adaptive auditing is Bayesian sequential design (Lindley, 1956; Chaloner & Verdinelli, 1995), with one-step (Roy & McCallum, 2001; Houlsby et al., 2011) and full-horizon dynamic-programming (Huan & Marzouk, 2016; Shen & Huan, 2023) solutions. The menu makes it a game. Γ-minimax theory equates the minimax risk over a set of priors with the Bayes risk under a least-favourable prior (Berger, 1985; Lehmann & Casella, 1998); Hedge solves finite zero-sum games by multiplicative weights (Freund & Schapire, 1997; 1999); robust Bayesian design maximises worst-case information gain over a KL ball or by maximin (Go & Isaac, 2022; Abdulsamad et al., 2026); distributionally robust learning protects a loss over an ambiguity set (Duchi & Namkoong, 2021; Mohajerin Esfahani & Kuhn, 2018; Frogner et al., 2021), including a per-round adversary in active learning (Zhu et al., 2019); robust MDPs let nature re-choose within a rectangular set at every step (Iyengar, 2005; Nilim & El Ghaoui, 2005), which we reject: one process produced the dataset. These lines supply the solution concept; CARD supplies the object: complete conditional laws of one dataset's true labels, with a terminal repair whose worst-case risk is the certificate.

Certificates and misspecification. Certified robustness bounds a classifier's loss over a perturbation set without knowing the attack (Wong & Kolter, 2018; Raghunathan et al., 2018; Cohen et al., 2019); certified risk is the dataset analogue over a set of laws. Robust Bayesian active learning states what an acquisition rule must satisfy under misspecification (Tang et al., 2026). Both confront misspecification without a checkable guarantee on a released dataset; CARD's certificate is exact relative to the menu and valid under coverage, a condition checked wherever gold exists (Section 4.3).

B PROOFS

B.1 PROOF OF THEOREM 3.3

Index the finite pure-policy set as $\Pi_B = \{\pi_1, \dots, \pi_M\}$ and form the $M \times K$ risk matrix A with $A_{mk} = R_k(\pi_m)$. For $\mu \in \Delta_M$,

$$\max_k R_k(\mu) = \max_{\boldsymbol{\lambda} \in \Delta_K} \mu^\top A \lambda,$$

because a linear function on a simplex attains its maximum at a vertex. Finite minimax duality gives

$$\min_{\mu \in \Delta_M} \max_{\boldsymbol{\lambda} \in \Delta_K} \mu^\top A \lambda = \max_{\boldsymbol{\lambda} \in \Delta_K} \min_{\mu \in \Delta_M} \mu^\top A \lambda.$$

For fixed λ, the inner linear minimum is attained at a vertex of Δ_M, so it equals $\min_{\pi \in \Pi_B} \sum_k \lambda_k R_k(\pi) = V_B(\boldsymbol{\lambda})$. Compactness of the two simplexes and bilinearity give attainment and a saddle pair. □

B.2 PROOF OF PROPOSITION 3.1

At a terminal history h, the conditional Bayes action is by definition the minimizer in (4). With $b > 0$ audits left, choosing i first partitions the conditional law by the possible trusted outcomes $Y_i = c$. Conditional expectation therefore gives the continuation value in (5); backward induction proves optimality and $V_B(\boldsymbol{\lambda}) = W_B^{\boldsymbol{\lambda}}(\varnothing)$. For separable Hamming loss, each coordinate action minimizes its own posterior error independently. Choosing a posterior mode produces error $1 - \max_a P_{\boldsymbol{\lambda}}(Y_j = a \mid \boldsymbol{h})$, and summing with weights c_j yields (6). □

B.3 PROOF OF THEOREM 3.4

The returned uniform mixture has risk $R_k(\mu_T) = T^{-1} \sum_t R_k(\pi_t)$, hence its worst risk is U_T and $V_B^\star \le U_T$. The averaged adversary is feasible, so Theorem 3.3 gives $L_T = V_B(\bar{\boldsymbol{\lambda}}_T) \le V_B^\star$.

The exponential-weights regret inequality for rewards in $[0, 1]$ is

$$\max_k \frac{1}{T} \sum_t R_k(\pi_t) - \frac{1}{T} \sum_t \lambda_t^\top R(\pi_t) \le \frac{\log K}{\eta T} + \frac{\eta}{8}.$$

Exact best response gives $\lambda_t^\top R(\pi_t) = V_B(\lambda_t)$. The pointwise minimum of linear functions, $V_B(\boldsymbol{\lambda})$, is concave. Jensen's inequality therefore gives $T^{-1} \sum_t V_B(\lambda_t) \le V_B(\bar{\boldsymbol{\lambda}}_T) = L_T$. Substitution proves (11); minimizing its right-hand side over η gives the stated choice. □

B.4 PROOF SKETCH OF COROLLARY B.1

Corollary B.1 (Approximate-oracle robust regret)**.** *Suppose round t returns an ε-best response in true weighted risk, and the vector used in the exponential update is uniformly within δ of the true risk vector. Then the returned mixture satisfies*

$$\max_k R_k(\mu_T) - V_B^\star \le \frac{\log K}{\eta T} + \frac{\eta}{8} + \varepsilon + 2\delta. \tag{12}$$

A numerical lower *certificate additionally requires a certified lower bound on $V_B(\bar{\boldsymbol{\lambda}}_T)$; approximate minimization alone supplies an upper, not a lower, bound on this Bayes value.*

Apply the same exponential-weights inequality to the estimated rewards. A uniform δ error changes the comparison with the best fixed coordinate by at most 2δ. The ε-best-response condition changes $\lambda_t^\top R(\pi_t) = V_B(\lambda_t)$ to an upper inequality with additive ε. Concavity and $\max_k R_k(\mu_T) \ge V_B^\star$ give (12). □

B.5 PROOF OF PROPOSITION B.2

The statement, moved here from Section 3 for space, is:

Proposition B.2 (KL ambiguity dual and robustness path)**.** *For every $r \in [0,1]^K$,*

$$\mathcal{R}_\rho(r) = \inf_{\gamma>0} \left\{ \gamma \log\left(\sum_{k=1}^{K} w_k e^{r_k/\gamma} \right) + \gamma\rho \right\}. \tag{13}$$

At $\rho = 0$, $\mathcal{R}_0(r) = w^\top r$. As the ball grows to include a point mass on a worst candidate, $\mathcal{R}_\rho(r) = \max_k r_k$. Hence $\min_\mu \mathcal{R}_\rho\{R(\mu)\}$ connects nominal Bayes and hard minimax repair.

Introduce multiplier $\gamma \geq 0$ for the KL constraint and a multiplier for the simplex normalization. Maximizing the Lagrangian over q gives the exponential tilt

$$q_k(\gamma) = \frac{w_k \exp(r_k/\gamma)}{\sum_m w_m \exp(r_m/\gamma)}$$

and the dual objective in (13). Convexity and the full support of w give strong duality (with the limiting interpretation at the boundary). The cases $\rho = 0$ and inclusion of a worst-coordinate point mass follow directly from the feasible set. □

B.6 Proof of Proposition B.3

The statement, moved here from Section 3 for space, is:

Proposition B.3 (From model-set coverage to an honest risk statement)**.** *Let $\widehat{\mathcal{P}}(\mathcal{D}_0)$ be a data-dependent set of conditional laws such that*

$$\mathbb{P}\{P^\dagger(\cdot \mid \mathcal{D}_0) \in \widehat{\mathcal{P}}(\mathcal{D}_0)\} \geq 1 - \alpha, \tag{14}$$

where $P^\dagger$ is the true conditional audit law. For any data-dependent policy $\widehat{\mu}$ and

$$U(\mathcal{D}_0) = \sup_{P \in \widehat{\mathcal{P}}(\mathcal{D}_0)} R_P(\widehat{\mu}; \mathcal{D}_0), \tag{15}$$

the true conditional risk obeys $R_{P^\dagger}(\widehat{\mu}; \mathcal{D}_0) \leq U(\mathcal{D}_0)$ on the coverage event. If $\ell \in [0,1]$, then

$$\mathbb{E} R_{P^\dagger}(\widehat{\mu}; \mathcal{D}_0) \leq \mathbb{E} U(\mathcal{D}_0) + \alpha. \tag{16}$$

On the event in (14), the true conditional law is one of the laws over which $U(\mathcal{D}_0)$ takes a supremum, proving the conditional inequality. Let E denote this event. Since the loss and hence risk are in $[0,1]$,

$$\mathbb{E} R_{P^\dagger} = \mathbb{E}[R_{P^\dagger} \mathbf{1}_E] + \mathbb{E}[R_{P^\dagger} \mathbf{1}_{E^c}] \leq \mathbb{E}[U \mathbf{1}_E] + \mathbb{P}(E^c) \leq \mathbb{E} U + \alpha.$$

□

B.7 Proof of Theorem B.4

The statement and its construction, moved here from the main text for space, are:

Theorem B.4 (Unbounded risk ratio for one-step acquisition)**.** *For every integer $M \geq 1$, there is a singleton-model binary repair problem with budget $B = 2$ such that exact full-horizon planning has risk*

$$R_{\text{full}} = \frac{1}{2(M+1)}, \tag{17}$$

whereas sequential one-step terminal-risk reduction has risk

$$R_{\text{greedy}} = \frac{M}{2(M+1)}. \tag{18}$$

Thus $R_{\text{greedy}}/R_{\text{full}} = M$.

Let A, B, D be independent fair bits, let M target labels equal $A \oplus B$ and let D be a distractor; only A, B, D can be audited, the loss weights each target and D by $1/(M+1)$ and the calibration bits by zero. Auditing A or B alone leaves the XOR fair and has zero immediate value, so the one-step rule audits D first and its second audit cannot resolve the pair, whereas full-horizon planning audits A then B; the identity is verified to 1.42×10^{-14} for $M \in \{1, 2, 4, \ldots, 64\}$ (the numerical audit and the ratio against M are in the supplement).

Before any audit, every target $A \oplus B$ and the distractor D has Bayes error $1/2$. Revealing A alone (or B alone) leaves $A \oplus B$ fair and has no immediate effect on any positive-weight coordinate. Revealing D removes error $1/[2(M+1)]$, so greedy selects it strictly. With one audit left, revealing only one of A, B still leaves all targets fair, giving risk $M/[2(M+1)]$. The full-horizon tree reveals both A and B, determines all M targets, and leaves only D fair, giving $1/[2(M+1)]$. □

B.8 THE MIXING EXAMPLE

Example B.5 (Why a single Bayes tie-break is not enough)**.** There are two equally sized sources and one audit. Under P_1, all labels from source 1 share a fair bit while source 2 is known zero; under P_2 the roles reverse. Auditing source 1 has risk vector $(0, 1/4)$ and auditing source 2 has $(1/4, 0)$. The least-favourable prior is $(1/2, 1/2)$, under which both pure queries are Bayes optimal. Either arbitrary tie-break has worst risk $1/4$. Sampling the two query trees with probability $1/2$ gives risk $(1/8, 1/8)$ and is minimax. Thus Theorem 3.3 reduces the value to a Bayes problem, but the saddle policy may still need to mix several Bayes best responses.

C EXPERIMENTAL DETAILS

This appendix gives the details needed to read the experiments of Section 4: the datasets, their sources and preprocessing (C.1); the candidate laws and the two menus (C.2); the dataset-scale protocol (C.3); the receding-horizon planner and the two cleaner modes (C.4); the statistical procedures (C.5); and the evaluation of the exact game at episode scale (C.6). Every study ran under a written protocol fixed on the development set before its first test result existed; the protocol documents and the full record of every run are in the supplement.

C.1 DATASETS: SOURCES, ACCESS AND PREPROCESSING

Table 3: Datasets. Rows are (image, annotator) records with one recorded label; annotators or sources are the blocks of the paired bootstrap; "wrong before audit" is the share of recorded labels that disagree with the withheld gold labels; budgets are $\lceil \rho n \rceil$ audits, shown at $\rho = 1\%$. [a]Rounds 2 and 3 (43,360 and 43,480 rows) replicate it. [b]42 / 369 / 589 images carry 1 / 2 / 3 labels. [c]38.27 % of the 37,497 images with both a web and a clean label disagree; the 8,000 rows are a seeded sample.

Dataset	Role	Rows / images	Labels per image	Annotators / sources	Wrong before audit	Audits at 1%
CIFAR-10N R1	development[a]	43,370 / 43,370	one	280 workers	17.2 %	434
CIFAR-10H	test	102,400 / 10,000	1–23 (mean 10.24)	512 workers	5.4 %	1,024
Dopanim	test	10,484 / 10,484	one	20 annotators	33.5 %	105
LabelMe-AMT	test	2,547 / 1,000	about 2.5[b]	59 workers	26.0 %	26
Clothing1M	test (web labels)	8,000 / 8,000	one	20 feature clusters	37.7 %[c]	80

CIFAR-10N (Wei et al., 2022). Crowd labels for the 50,000 CIFAR-10 training images, collected on Amazon Mechanical Turk in three rounds ("Random 1–3"), each giving one label per image with a worker identifier; the CIFAR-10 training labels serve as gold. Public release. We use the workers of round 1 with their complete annotation sets as the development set (43,370 rows, 280 worker blocks); rounds 2 and 3 (43,360 and 43,480 rows) relabel the same images with other worker batches and serve as replications. Truth models are out-of-fold logistic probes on the recorded labels (3 folds $\times$ 2 repeats) over frozen DINOv2 ViT-S/14 (Oquab et al., 2024) and ImageNet ResNet-50 (He et al., 2016) features.

CIFAR-10H (Peterson et al., 2019). Crowd labels for the 10,000 CIFAR-10 test images with anonymised annotator identifiers (2,571 workers, 514,200 annotations); the CIFAR-10 test labels serve as gold. Public raw annotation archive. We use the complete annotation sets of 512 workers with 200 annotations each, selected by a salted hash (102,400 rows; 1–23 labels per image, mean 10.24). The truth models are probes on the same DINOv2 and ResNet-50 features, trained on the 50,000 CIFAR-10 training labels; the soft destination matrix is estimated from all public CIFAR-10H annotations without gold.

Dopanim (Herde et al., 2024). Crowd labels for 10,484 images of 15 animal species from 20 annotators (159–1,052 rows each), with expert gold labels. Zenodo record 14016659. The whole table is used with the annotators as blocks; truth models are out-of-fold probes on the recorded labels over frozen DINOv2 embeddings.

LabelMe-AMT (Rodrigues & Pereira, 2018). Crowd labels for 1,000 scene images of 8 classes from 59 Mechanical Turk workers (2,547 labels; 42 / 369 / 589 images with 1 / 2 / 3 labels), with

gold labels. The authors' archive is served truncated (16,244,736 bytes); the answer matrix and gold were taken from the peerannot reproduction repository (`computorg/published-202402-lefort-peerannot` at commit `d904396f`), 998 of whose 1,000 images are byte-identical to the original release. Rows are (image, worker, label) records in 59 worker blocks; truth models are out-of-fold probes on the recorded labels with image-grouped folds; an image-cluster bootstrap is the secondary interval.

Clothing1M (Xiao et al., 2015). One million clothing images of 14 classes with labels scraped from the surrounding web text, of which 37,497 also carry a clean label assigned by the dataset's authors (38.27 % disagreement). Obtained from the authors on request and not redistributed. We use a seeded sample of 8,000 of the images with both labels, the web label as the recorded label and the clean label as gold. The dataset carries no annotator identity, so sources are k-means clusters of frozen DINOv2 features ($k = 20$ primary; $k = 10$ and 50 secondary). Two truth models are out-of-fold probes on the web labels; the third is a probe fit on the clean labels of the 29,497 both-label images outside the sample.

Four further datasets were examined and not run, all before any result: NoisyCXR (access-gated, no annotator identifiers) and Controlled Noisy Web Labels (distribution file unavailable) for access reasons, Food-101N and mini-WebVision because they lack the corrected gold labels a repair needs.

C.2 Candidate laws and the two menus

Every law of Eq. (2) is built from the same channel. For a record i of source s, let o_{si} be the recorded class, Y_{si} the trusted class, $q_{si}(c)$ a source-blind content probability and ρ_s the source's severity; D_s is the passive block of source s, $\mathcal{T}_s$ its target records and $\mathbf{\Omega}$ a zero-diagonal row-stochastic destination matrix. Then

$$T_{\mathbf{\Omega}}(o \mid c, \rho) = \begin{cases} 1-\rho, & o = c, \\ \rho \mathbf{\Omega}_{co}, & o \neq c, \end{cases} \qquad m_{si}(\rho) = \sum_c q_{si}(c) T_{\mathbf{\Omega}}(o_{si} \mid c, \rho), \tag{19}$$

and the trusted audit distribution at fixed severity is

$$p(Y_{si} = c \mid \rho, D_s) = \frac{q_{si}(c) T_{\mathbf{\Omega}}(o_{si} \mid c, \rho)}{m_{si}(\rho)}. \tag{20}$$

A *joint-likelihood* candidate lets all passive labels identify severity,

$$p_{\text{joint}}(\rho \mid D_s) \propto \text{Beta}(\rho; a, b) \prod_{i \in D_s} m_{si}(\rho), \tag{21}$$

whereas a *conditional-anchor* candidate sets

$$p_{\text{anchor}}(\rho \mid D_s) = \text{Beta}(\rho; a, b), \quad p_{\text{anchor}}(Y_{s\mathcal{T}} \mid \rho, D_s) = \prod_{i \in \mathcal{T}_s} p(Y_{si} \mid \rho, D_s), \tag{22}$$

a coherent joint conditional law once the passive annotation table is fixed: it expresses doubt that unverified history can estimate its own corruption severity, and trusted audits still update ρ by ordinary conditioning, so that every candidate is a probability law. Table 4 lists the four axes; menu-16 is their 2^4 cross, the nominal law the first level of every axis (`q0_soft_beta28_joint`), severity integrated on 64 Gauss–Legendre nodes. Content models per dataset: on CIFAR-10H, q_0 is an averaged-SGD multinomial logistic probe on DINOv2 ViT-S/14 features (Oquab et al., 2024) and q_1 the same probe on ImageNet ResNet-50 features (He et al., 2016), both trained on the 50,000 canonical CIFAR-10 training labels only, with the soft destination $\mathbf{\Omega}$ estimated from all 514,200 public CIFAR-10H annotations; on the other annotation tables the two content models are out-of-fold logistic probes on the recorded labels (3 folds $\times$ 2 repeats, seeds 0 and 1; LabelMe folds image-grouped). *Menu-24*, the declared secondary menu, appends a third content family q_2, a cosine k-nearest-neighbour estimator over the frozen DINOv2 features ($k = 10$, leave-self-out, Laplace smoothing 0.1, votes from the annotation table's recorded labels), crossed with the other three axes; it reports its certified risk on the sixteen shared laws. On the annotation-row CIFAR-10H annotation table the row-unit q_2 sees the other annotations of the same image as zero-distance neighbours, the information set of SimiFeat's native repair. Menu-24 is the 24-law arm of the menu-size ablation, run on the development set, CIFAR-10H and Dopanim.

Table 4: The four axes of the sixteen-law menu; the nominal law takes the bold level of every axis. The 24-law ablation menu adds the eight laws with $q = q_2$.

Axis	Symbol in Eq. (2)	Levels	What it hedges
Content model	q	q_0 **(DINOv2 probe)**, q_1 (ResNet-50 probe)	what the features say the label is
Destination	$\mathbf{\Omega}$	**soft** (gold-free confusion), uniform	where an error lands
Severity prior	(a, b)	**Beta(2,8)**, Beta(1,1)	how often a source errs
Root	r	**joint** (21), conditional anchor (22)	whether passive labels may estimate severity

Menu per dataset. CIFAR-10N (development, rounds 1–3): the sixteen-law menu with q_0 and q_1 out-of-fold logistic probes on the recorded labels over DINOv2 ViT-S/14 and ResNet-50 features. CIFAR-10H: the sixteen-law menu with q_0 and q_1 probes trained on the CIFAR-10 training labels and $\mathbf{\Omega}$ estimated from all public CIFAR-10H annotations. Dopanim: the sixteen-law menu with out-of-fold probes over DINOv2 embeddings. LabelMe-AMT: the sixteen-law menu with out-of-fold probes on image-grouped folds. Clothing1M: the sixteen-law menu plus the eight laws of a third truth model q_2, a probe fit on the clean labels of the 29,497 both-label images outside the sample, crossed with the other three axes; its certificates are the maximum over all 24 laws. Menu-size ablation: the eight-law menu drops the conditional-anchor root; the 24-law ablation menu adds the kNN family above, with certificates over the sixteen shared laws.

C.3 Dataset-scale protocol

Each annotation table is the union of source blocks; each block is one game context with all its rows as repairable targets; the terminal repair rewrites the whole table; the budget is $\lceil \rho n \rceil$ audits; every run records the whole-dataset hidden-gold loss, the unaudited-only loss, the per-law expected terminal risk of the released labels and its maximum over the menu, the weight trajectory, the audit log and the runtime. The oracle firewall is a replay contract: loaders take no gold argument, the driver commits one row at a time to an oracle that reveals only that row, and the evaluation reads gold after the budget is exhausted and checks that the reveal count equals the budget. On CIFAR-10H the preregistered primary contrast is the minimax arm minus ALC as published on hidden-gold loss at 1 % over the 512 worker blocks, with same-sign replication on Dopanim; the co-primary claim is the lowest certified risk against all 22 comparators at every budget. The misspecification experiment on the development set runs the four weight rules on menu-16 and on the eight-law menu without the conditional-anchor family, plus leave-one-axis-out menus, at 434 audits; its primary reading is the paired difference in differences $[\mathrm{gold}(\mathrm{CARD} - posterior, 8) - \mathrm{gold}(\mathrm{CARD} - posterior, 16)] - [\mathrm{gold}(\mathrm{CARD} - minimax, 8) - \mathrm{gold}(\mathrm{CARD} - minimax, 16)]$, read as directional (CI > 0, $p < 0.05$), bounded equivalence (CI within ± 0.005) or inconclusive, with the same sign required on rounds 2 and 3.

C.4 The receding-horizon planner, the weight rules and the two cleaner modes

Figure 3 states the dataset-scale planner: the one-step calculation is exact (Proposition 3.1 with $B = 1$), the receding-horizon commitment and the lazy refresh are the approximation, and no exact-game optimality bracket attaches to any receding-horizon run (Corollary B.1 regime). Only the weight rule distinguishes the four λ-arms; NOM-GREEDY is the same planner with λ fixed at the nominal one-hot and $\eta = 0$, which is why the one-law menu reproduces it to four decimals on the development set (0.1396; Table 8). Sensitivities on the development set at 434 audits (CARD-minimax, 16 laws): refresh tolerance 0.02, 0.10 and 0.5 give hidden-gold loss 0.0414, 0.0415 and 0.0420, disabling re-planning gives 0.0437, and 32, 64 and 128 quadrature nodes give 0.0416, 0.0415 and 0.0415. Each published cleaner runs on identical annotation tables, budgets and oracle in two modes. *As published*: its own ranking selects the audits and its own repair rule, where one exists, rewrites the unaudited rows—Confident Learning takes the argmax of its predicted probabilities on the rows it flags, SIMIFEAT its kNN majority vote ($k = 10$, cosine), which on an annotation-row annotation table sees the other annotations of the same image as zero-distance neighbours (mean 9.2 of 10 on CIFAR-10H)—otherwise every unaudited label is kept, so the release contains at most B fewer errors than the input. *On our repair* ("chassis"): the cleaner's ranking replaces step 2(a) of Figure 3 and the engine's mixture repair, Hedge weights, firewall and evaluation are unchanged, so every

CARD-minimax and its λ-arms.

Input: annotation table $\mathcal{D}_0$ as blocks w (one source each, every row a target), menu $P_1, \ldots, P_K$, budget B, weight rule. *Constants:* 64 severity quadrature nodes; refresh tolerance 0.10; Hedge rate $\eta = \sqrt{8 \ln K/B}$ (Freund & Schapire, 1997).

1. Initialise λ_0 (uniform). For every block w and every law k, compute the per-law current terminal risk and, for every unaudited row i, the exact one-step expected post-audit terminal risk $M_w[i,k] = \sum_o P_k(o \mid i)\, \mathrm{risk}_k(\mathrm{child}(i,o))$, where the child's terminal action is the mixture-MAP repair under the current weights conditioned on the hypothetical outcome o (Proposition 3.1, $B = 1$).
2. Repeat B times: (a) score every unaudited row by the λ-weighted exact risk reduction $\lambda^\top \mathrm{cur}_w - \lambda^\top M_w[i]$, re-weighted exactly under the current λ; (b) commit the global arg-max row, reveal its trusted label (one oracle reveal), condition its block; (c) update λ by the weight rule; (d) refresh the physics (step 1) of the block just audited, and lazily of the arg-max block whenever $\|\lambda_{\mathrm{now}} - \lambda_{\mathrm{phys}}\|_1 > 0.10$, iterating until the winner is fresh.
3. Terminal repair: audited rows take the revealed label; unaudited rows take the mixture-MAP action under the repair weights $\bar{\lambda}$. Report the hidden-gold loss (evaluation only) and the certified risk $\max_k R_k$ of the released labels under $\bar{\lambda}$'s menu (Eq. (3), on the 16-law subset for menu-24).

Weight rules ($\widehat{R}_t$ = current annotation table-average per-law expected terminal risk vector):

- CARD-minimax (Hedge, adversarial): $\log w_k \mathrel{+}= \eta\, \widehat{R}_{t,k}$, $\lambda_t = \mathrm{softmax}(\log w)$; repair weights $\bar{\lambda}$ = the average of the played λ_t (the Hedge-average convention of Section 3).
- CARD-posterior (posterior): λ_t = the running posterior over the menu given the revealed audit outcomes; repair under λ_B.
- CARD-tempered (risk-tilted posterior, Proposition B.2 path): $\lambda_t \propto \mathrm{posterior}_t \cdot \exp(\widehat{R}_t/\tau)$, $\tau = 0.1$ ($\tau \to \infty$ recovers CARD-posterior); repair under λ_B.
- CARD-uniform (control): $\lambda_t \equiv 1/K$, frozen.
- NOM-GREEDY: $\lambda \equiv$ nominal one-hot, $\eta = 0$; nominal-law repair. MPC: audit the largest current nominal posterior error probability; nominal repair. RAND: uniform over unaudited rows, eight frozen seeds, mixture-averaged.

Figure 3: The receding-horizon planner at dataset scale and the weight rules of its arms. A cleaner on our repair replaces step 2(a) by the cleaner's ranking.

such run carries a menu-relative certified risk and isolates acquisition from repair. The probability lane (ALC, Confident Learning, ACTIVELAB, CINCER) scores q_0; the feature lane (SIMIFEAT, AUM, FORGETSCORES, SMALLLOSS) scores the frozen DINOv2 features; ACTIVELAB runs in the single-label regime, where its ranking coincides with Confident Learning's (identical ranking hash `6afad8c583`; Appendix E).

C.5 STATISTICS

All arms audit the same rows at the same budgets, so every contrast is paired by block (worker, annotator or source cluster): 95 % cluster percentile bootstraps of the block-paired, row-weighted mean difference with 10,000 draws and one fixed seed per study; a sign-flip permutation p-value on the per-block contributions; an image-cluster bootstrap where images recur across blocks; Holm correction within each dataset and family for secondary contrasts; no pooling across datasets or across k. Certified-risk contrasts resample blocks, recompute the per-law risk vector of the table and take its maximum, so that the two columns of every table are compared like with like. Seeds are in the supplement.

C.6 EXACT-GAME EVALUATION AT EPISODE SCALE

The main text runs the receding-horizon planner at dataset scale. This section evaluates the exact game of Section 3 at the scale at which the full-horizon tree is computable: one annotator's records, with $B = 2$ audits over eight target records (an *episode*). Every Bayes best response is computed by exact multiclass dynamic programming (Proposition 3.1) with 64-node severity quadrature, and the minimax weighting by Hedge with $T = 64$ rounds (an independent $T = 128$ run serves as the certificate-stability check), so that the certificate is the upper end U_T of the bracket (10). Arms share the menu (menu-16) and the terminal repair and differ in acquisition: RAND (64 frozen draws), ALC on the frozen q_0, NOM-DP (exact full-horizon planning under the nominal law), ROB-GREEDY (its

one-step counterpart under the menu) and the exact minimax policy (CARD-minimax in Table 5); hidden gold is loaded only after every acquisition tree is committed, and a randomized policy is scored as the expectation over its components. Episodes: CIFAR-10H, 512 workers selected by a salted hash from the 2,571 of Peterson et al. (2019), each worker's 200 annotations the passive context, 4,096 targets over 3,363 images (worker and image-cluster intervals, 10,000 and 20,000 draws); Dopanim, 16 disjoint episodes per annotator (320), exact 15-class dynamic programming, annotator bootstrap (20,000 draws); LabelMe-AMT, 293 episodes from the 46 workers with at least eight rows, exact 8-class dynamic programming.

On the 512 CIFAR-10H episodes the exact minimax policy lowers the certified risk from 0.027460 (NOM-DP) to 0.010005, a paired difference of -0.017455 $[-0.019774, -0.015187]$ (63.6 %). The nominal law is right here (NOM-DP leaves no residual error), so the minimax policy's gold loss of 0.000755 is a premium, $+0.000755$ $[+0.000031, +0.001701]$, the price of insurance at episode scale; against ROB-GREEDY the two-step lookahead lowers the certified risk by -0.000587 $[-0.000808, -0.000397]$ with the gold difference unresolved. Gold loss minus the upper certificate U_T averages -0.009249 (one-sided 95 % bound -0.008111), so the certificate covers the hidden-gold loss; the $T = 128$ Hedge gaps shrink relative to $T = 64$ and stay inside the envelope of Theorem 3.4 in all 512 episodes, and the budget-one exact/greedy identity holds to discrepancy 0. On Dopanim (320 episodes) and LabelMe-AMT (293 episodes) the unchanged design gains on gold and on certified risk against NOM-DP (Table 5b): domains where the nominal law is wrong enough that both improve together. The per-policy tables with MPC and NOM-GREEDY, the paired-difference figure and the full calibration record are in the supplement.

Table 5: Episode scale ($B = 2$, sixteen-law menu, exact game). (a) Hidden-gold loss / certified risk averaged over episodes (CIFAR-10H 512, Dopanim 320, LabelMe-AMT 293 episodes; eight targets each). Bold: lowest certified risk; underline: lowest gold. (b) Paired contrasts, exact minimax minus comparator, with 95% cluster-bootstrap intervals; LabelMe adds the fraction of episodes in which minimax is strictly better / tied.

(a) Hidden-gold loss / certified risk

	CIFAR-10H		Dopanim		LabelMe-AMT	
Policy	gold	certified	gold	certified	gold	certified
RAND	0.008923	0.060383	0.204449	0.264926	0.1757	0.1861
ALC	0.001221	0.029519	0.151172	0.214849	0.1011	0.1269
NOM-DP	<u>0.000000</u>	0.027460	0.140234	0.212496	0.0956	0.1233
ROB-GREEDY	0.000580	0.010592	0.131097	0.180733	0.0811	0.1023
CARD-minimax (exact)	0.000755	**0.010005**	<u>0.129706</u>	**0.178152**	<u>0.0800</u>	**0.0994**

(b) Paired contrasts, exact minimax minus comparator

Dataset	Comparator	Metric	Difference	95% CI	Note
CIFAR-10H	NOM-DP	gold	0.000755	[0.000031, 0.001701]	the premium
CIFAR-10H	NOM-DP	certified risk	−0.017455	[−0.019774, −0.015187]	relative reduction 63.6%
CIFAR-10H	ROB-GREEDY	certified risk	−0.000587	[−0.000808, −0.000397]	gold difference unresolved
Dopanim	NOM-DP	gold	−0.010529	[−0.020410, −0.001355]	
Dopanim	NOM-DP	certified risk	−0.0343	[−0.0425, −0.0270]	
Dopanim	ALC	gold	−0.021466	[−0.037696, −0.006409]	
Dopanim	ROB-GREEDY	gold	−0.001392	[−0.004291, 0.001379]	unresolved
LabelMe	NOM-DP	gold	−0.0156	[−0.0231, −0.0082]	better 0.235, tied 0.638
LabelMe	NOM-DP	certified risk	−0.0239	[−0.0265, −0.0215]	better 0.894, tied 0.014
LabelMe	ROB-GREEDY	gold	−0.0011	[−0.0032, 0.0008]	better 0.113, tied 0.809
LabelMe	ROB-GREEDY	certified risk	−0.0029	[−0.0034, −0.0023]	better 0.474, tied 0.392
LabelMe	ALC	gold	−0.0211	[−0.0298, −0.0127]	better 0.266, tied 0.611
LabelMe	ALC	certified risk	−0.0275	[−0.0305, −0.0247]	better 0.928, tied 0.003
LabelMe	RAND	gold	−0.0957	[−0.1076, −0.0840]	better 0.782, tied 0.126
LabelMe	RAND	certified risk	−0.0867	[−0.0907, −0.0825]	better 1.000, tied 0.000

D FULL RESULT TABLES

This appendix carries the result tables the main text draws on. Conventions are those of Section 4: *gold* is the hidden-gold loss (fraction of records still wrong after repair, against withheld gold labels); *certified risk* ($\mathcal{C}$, Eq. (3)) is the maximum over the 16-law menu of the expected fraction of records still wrong in the released dataset, computed from the annotation table and the audit transcript alone

(24-law arms report it on the sixteen shared laws); lower is better on both. Intervals are 95 % cluster percentile bootstraps (10,000 draws unless stated) over the block named in each caption; p is a sign-flip permutation p-value. Arm names follow Section 4.1, with the engine names of Sections 3–3 beside them. Every per-dataset table lists every arm run on that dataset at every budget; the exact game at episode scale is evaluated in Appendix C.6.

D.1 DEVELOPMENT SET: CIFAR-10N R1, WITH ROUNDS 2 AND 3

The development set is CIFAR-10N round 1 (43,370 rows, 280 worker blocks, initial hidden error 0.1716) (Wei et al., 2022); rounds 2 and 3 (43,360 and 43,480 rows) relabel the same images with different workers and replicate the four arms of the original grid. Every design decision of the paper was made here; nothing in this subsection is test-set evidence. Table 6 is the full grid (six internal arms, the two menu-24 arms and the eight cleaners in both modes at five budgets); Table 7 the replication; Table 8 separates repair prior from audit placement and varies the menu size. Each cleaner is run *as published* (its own ranking; its own repair where one exists—native repair for Confident Learning and SIMIFEAT, keep-observed otherwise) and *on our repair* (its ranking as our audit rule, our mixture repair, our protocol); a keep-observed cleaner can fix at most B records, which is why ALC, ACTIVELAB, CINCER, AUM and SMALLLOSS share the gold loss 0.122 at 5 %. All 80 paired development contrasts (cleaner minus CARD-minimax, worker-cluster bootstrap, 10,000 draws, seed 20260818) exclude zero: the native-repair cleaners win on gold at small budgets (-0.0212 $[-0.0242, -0.0179]$ at 0.25 % for Confident Learning as published, flipping to $+0.0024$ $[+0.0004, +0.0043]$ at 1 %; -0.0383 $[-0.0417, -0.0347]$ at 0.25 % and -0.0056 $[-0.0076, -0.0037]$ at 2 % for SIMIFEAT as published, flipping to $+0.0063$ $[+0.0045, +0.0082]$ at 5 %; negative = the cleaner ahead), and every mount on our repair loses to CARD-minimax at every budget (smallest margin $+0.0038$ $[+0.0017, +0.0059]$, SIMIFEAT on our repair at 1 %); the 80 intervals are in the supplement. Downstream heads trained on the released labels (supplement) move little because frozen-feature classifiers are noise-robust: a ridge classifier on ResNet-50 features rises from 0.9152 on the raw labels to 0.9230 at 5 % with CARD-minimax's release against 0.9235 with gold, and the DINOv2 heads move by at most 0.3 points.

Table 6: Development set (CIFAR-10N round 1): hidden-gold loss / certified risk of every arm at every budget. Bold: lowest certified risk in the column; underline: lowest gold, over the sixteen-law-menu arms. Zero-audit floors: nominal repair $.1494/.1893$, mixture repair $.1305/.1782$.

Arm	Mode	109 (0.25%)	217 (0.5%)	434 (1%)	868 (2%)	2,169 (5%)
RAND (random)	internal	.1487/.1816	.1480/.1773	.1466/.1709	.1438/.1628	.1365/.1478
MPC (most probably corrupted)	internal	.1491/.1915	.1483/.1938	.1469/.1986	.1433/.2094	.1217/.2348
NOM-GREEDY (nominal one-step)	internal	.1466/.1926	.1440/.1965	.1396/.2048	.1309/.2167	.1030/.2422
CARD-uniform (uniform)	internal	.1247/.1799	.1207/.1812	.1120/.1800	.0956/.1815	.0579/.1754
CARD-posterior (posterior)	internal	.0426/**.1465**	.0507/.1497	.0451/.1455	.0297/**.1332**	.0184/**.1109**
CARD-minimax (minimax)	internal	.0653/.1521	.0519/.1515	.0415/.1451	.0326/.1341	.0202/.1135
menu-24 CARD-minimax	internal	.0546/.159	.0436/.155	.0337/.150	.0247/.144	.0131/.121
menu-24 CARD-tempered	internal	.0389/.161	.0377/.150	.0402/.156	.0298/.146	.0130/.130
ALC	as published	.169/.208	.167/.211	.162/.216	.152/.226	.122/.251
ALC	on our repair	.076/.156	.065/.155	.053/.157	.043/.165	.041/.190
Confident Learning	as published	.044/.152	.044/.153	.044/.157	.044/.167	.042/.199
Confident Learning	on our repair	.076/.156	.065/.155	.052/.157	.043/.164	.041/.190
ACTIVELAB	as published	.169/.208	.167/.211	.162/.217	.152/.227	.122/.253
ACTIVELAB	on our repair	.076/.156	.065/.155	.052/.157	.043/.164	.041/.190
SIMIFEAT	as published	.027/.167	.027/.171	.027/.182	.027/.201	.026/.229
SIMIFEAT	on our repair	.071/.162	.057/.163	.045/.174	.042/.190	.051/.214
CINCER	as published	.169/.210	.167/.215	.162/.223	.152/.234	.122/.258
CINCER	on our repair	.075/.157	.063/.157	.051/.161	.042/.170	.043/.199
AUM	as published	.169/.211	.167/.216	.162/.225	.152/.235	.122/.260
AUM	on our repair	.075/.158	.062/.158	.050/.162	.041/.170	.043/.197
FORGETSCORES	as published	.171/.202	.171/.199	.170/.196	.168/.192	.162/.185
FORGETSCORES	on our repair	.079/.151	.070/**.146**	.062/**.140**	.056/.135	.050/.128
SMALLLOSS	as published	.169/.208	.167/.212	.162/.217	.152/.227	.122/.253
SMALLLOSS	on our repair	.077/.156	.064/.156	.053/.157	.042/.165	.042/.191

D.2 MECHANISM: MENU COVERAGE VERSUS WEIGHTING RULE (A4)

Figure 4 is the misspecification ablation that Section 4.3(c) summarises: development set round 1, 434 audits, hidden-gold loss, paired per worker block (280 blocks).

Table 7: Replication on CIFAR-10N rounds 2 and 3: hidden-gold loss / certified risk of the four replicated arms. Bold: lowest certified risk in the column; underline: lowest gold. All 45 paired gold contrasts of CARD-minimax against RAND, MPC and NOM-GREEDY exclude zero.

Round	Arm	0.25%	0.5%	1%	2%	5%
R2: audits 109, 217, 434, 868, 2,168						
2	RAND (random)	.1573/.1881	.1565/.1835	.1551/.1760	.1523/.1665	.1442/.1530
2	MPC (most probably corrupted)	.1575/.1977	.1568/.2000	.1552/.2055	.1512/.2148	.1305/.2416
2	NOM-GREEDY (nominal one-step)	.1553/.1998	.1525/.2036	.1477/.2112	.1392/.2230	.1119/.2490
2	CARD-minimax (minimax)	.0674/**.1593**	.0565/**.1535**	.0443/**.1513**	.0346/**.1434**	.0212/**.1179**
R3: audits 109, 218, 435, 870, 2,174						
3	RAND (random)	.1538/.1832	.1531/.1782	.1515/.1712	.1487/.1636	.1413/.1503
3	MPC (most probably corrupted)	.1542/.1942	.1535/.1963	.1519/.2018	.1484/.2125	.1269/.2400
3	NOM-GREEDY (nominal one-step)	.1517/.1967	.1492/.2006	.1444/.2091	.1355/.2213	.1082/.2451
3	CARD-minimax (minimax)	.0631/**.1595**	.0520/**.1544**	.0435/**.1482**	.0343/**.1406**	.0195/**.1150**

Table 8: Development set: (a) decomposition at 109 audits, separating the mixture repair from audit placement; (b) menu size for CARD-minimax at 434 audits: menus of 1, 2, 4 and 8 laws reach a gold loss of about 0.14 against 0.0415 with sixteen laws (a smaller menu's certificate is the maximum over that menu). Menu-1 reproduces NOM-GREEDY.

(a) Decomposition ladder, R1, hidden-gold loss

Initial error of the annotation table	0.1716
Zero-audit nominal repair (NOM-GREEDY, $B = 0$)	0.1494
Zero-audit mixture repair (CARD-minimax, $B = 0$)	0.1305
nominal one-step placement, 109 audits, mixture repair	0.1280
nominal one-step placement, 868 audits, mixture repair	0.1116
nominal one-step placement, 2,169 audits, mixture repair	0.0909
CARD-minimax placement, 109 audits, mixture repair	0.0653

(b) Menu cardinality, R1, CARD-minimax at 434 audits

Menu	Gold	Unaudited	Certified risk (over that menu)
1	0.1396	0.1410	0.0396
2	0.1411	0.1425	0.0396
4	0.1432	0.1447	0.0400
8	0.1438	0.1452	0.0404
16	0.0415	0.0419	0.1451

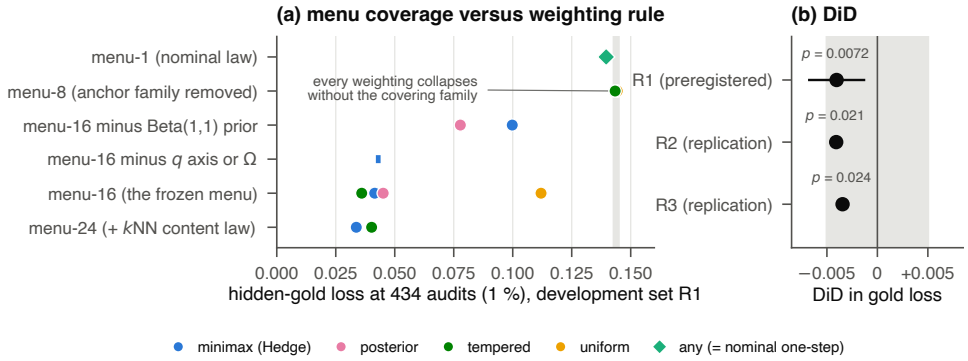

Figure 4: Misspecification ablation on the development set at 434 audits. (a) Hidden-gold loss per (menu, weighting): without the conditional-anchor family (eight laws) all four weightings collapse to 0.143–0.144; without the flat severity prior the loss rises to 0.0998 (minimax) and 0.0778 (posterior); with sixteen laws posterior, tempered and minimax lie within one point. (b) The difference-in-differences $[\text{gold}(\text{posterior}, 8) - \text{gold}(\text{posterior}, 16)] - [\text{gold}(\text{minimax}, 8) - \text{gold}(\text{minimax}, 16)]$ with its 95% interval on round 1 and the replications on rounds 2 and 3; the band is the ± 0.005 equivalence margin.

D.3 MENU-SIZE ABLATION: THE 24-LAW MENU, AND THE IMAGE-UNIT DIAGNOSTIC

Menu-24 adds the kNN truth-model family to the 16-law menu (Appendix C.2) and is run on the development set (the two arms are rows of Table 6; grid in the supplement), CIFAR-10H and Dopanim. Its certified risk is always reported on the 16-law subset. On CIFAR-10H the arm removes most of the premium: menu-24 CARD-minimax has gold $0.0171 \to 0.0030$, below the sixteen-law arm by -0.0029 $[-0.0036, -0.0023]$ at 0.25% and -0.0041 $[-0.0047, -0.0036]$ at 5%, and its first interval-supported gold win over SIMIFEAT arrives at 2% (-0.0017 $[-0.0025, -0.0010]$), inside the preregistered window; the larger hull is paid for in certificate ($+0.0082$ $[+0.0068, +0.0095]$ at 0.25%, $+0.0374$ $[+0.0356, +0.0391]$ at 5% on the 16-law subset). The tempered weighting on menu-24 crosses at 1% (-0.0017 $[-0.0023, -0.0011]$; P-M24b holds) and has the lowest gold of every internal arm at every budget ($0.0112 \to 0.0018$) at a 16-subset certificate of 0.0700–0.0474 against 0.0539–0.0094 for the sixteen-law arm (supplement). The image-unit diagnostic that Section 4.3(b) uses to decompose the gap to SIMIFEAT at 1%: of the 0.0101 between the sixteen-law minimax arm (0.0203) and SIMIFEAT as published (0.0102), 0.0052 is closed by the kNN law without access to the image's other rows, a further 0.0028 by letting it see them (the row-unit arm), and the remaining 0.0021 by SIMIFEAT's uncertified rewrite of every row (derived differences).

D.4 TEST SETS CIFAR-10H AND DOPANIM: ALL ARMS

CIFAR-10H: 102,400 rows (512 workers $\times$ 200 annotations, all 10,000 images; budgets 256, 512, 1,024, 2,048 and 5,120 audits); Dopanim: 10,484 rows, 20 annotator blocks, 15 classes (budgets 27, 53, 105, 210 and 525). Bootstrap: block-cluster percentile with sign-flip permutation, 10,000 draws, seed 20260915. Which law attains the maximum is logged per run but not tabulated. Tables 9 and 10 carry the 23-arm claims of Table 1 and the points of Figure 2.

Table 9: CIFAR-10H: hidden-gold loss / certified risk of every arm at every budget. Bold: lowest certified risk in the column; underline: lowest gold, over the sixteen-law-menu arms (the 24-law rows report certified risk on the sixteen shared laws). Zero-audit floors: mixture repair $.0168/.0674$, nominal repair $.0167/.0709$.

Arm	Mode	256 (0.25%)	512 (0.5%)	1,024 (1%)	2,048 (2%)	5,120 (5%)
RAND (random)	internal	.0166/.0695	.0166/.0682	.0165/.0659	.0163/.0619	.0158/.0534
MPC (most probably corrupted)	internal	.0167/.0708	.0167/.0708	.0166/.0717	.0165/.0760	.0084/.0865
NOM-GREEDY (nominal one-step)	internal	.0151/.0720	.0140/.0763	.0113/.0811	.0079/.0766	.0031/.0449
CARD-uniform (uniform)	internal	.0161/.0696	.0153/.0692	.0136/.0622	.0103/.0469	.0046/.0146
CARD-posterior (posterior)	internal	.0153/.0716	.0122/.0727	.0100/.0687	.0069/.0608	.0030/.0372
CARD-tempered (tempered)	internal	.0154/.0711	.0122/.0714	.0101/.0674	.0070/.0607	.0030/.0371
CARD-minimax (minimax)	internal	.0200/**.0539**	.0209/**.0442**	.0203/**.0325**	.0176/**.0238**	.0071/**.0094**
menu-24 CARD-minimax	internal	.0171/.0620	.0153/.0582	.0123/.0581	.0084/.0570	.0030/.0468
menu-24 CARD-tempered	internal	.0112/.0699	.0099/.0700	.0085/.0673	.0068/.0638	.0018/.0474
ALC	as published	.0519/.0951	.0494/.0938	.0444/.0918	.0344/.0896	.0083/.0884
ALC	on our repair	.0203/.0623	.0220/.0608	.0249/.0599	.0300/.0607	.0416/.0636
Confident Learning	as published	.0144/.0708	.0144/.0714	.0144/.0731	.0143/.0775	.0083/.0872
Confident Learning	on our repair	.0203/.0623	.0220/.0608	.0249/.0599	.0300/.0607	.0410/.0631
ACTIVELAB	as published	.0519/.0951	.0494/.0938	.0444/.0918	.0344/.0896	.0084/.0876
ACTIVELAB	on our repair	.0203/.0623	.0220/.0608	.0249/.0599	.0300/.0607	.0410/.0631
SIMIFEAT	as published	.0103/.0736	.0103/.0751	.0102/.0782	.0102/.0845	.0070/.1044
SIMIFEAT	on our repair	.0201/.0636	.0220/.0627	.0256/.0627	.0322/.0646	.0506/.0752
CINCER	as published	.0519/.0952	.0494/.0936	.0444/.0923	.0345/.0904	.0090/.0868
CINCER	on our repair	.0202/.0623	.0220/.0607	.0251/.0601	.0306/.0612	.0410/.0623
AUM	as published	.0519/.0950	.0494/.0931	.0444/.0900	.0344/.0878	.0093/.0829
AUM	on our repair	.0203/.0622	.0220/.0604	.0247/.0589	.0294/.0597	.0386/.0595
FORGETSCORES	as published	.0542/.0968	.0541/.0966	.0537/.0960	.0533/.0951	.0521/.0924
FORGETSCORES	on our repair	.0201/.0620	.0219/.0599	.0243/.0576	.0271/.0553	.0303/.0519
SMALLLOSS	as published	.0519/.0952	.0494/.0935	.0444/.0902	.0344/.0891	.0082/.0901
SMALLLOSS	on our repair	.0202/.0622	.0220/.0605	.0246/.0588	.0294/.0599	.0423/.0643

Table 10: Dopanim: hidden-gold loss / certified risk of every arm at every budget. Bold: lowest certified risk in the column; underline: lowest gold, over the sixteen-law-menu arms. Zero-audit floors: mixture repair $.3204/.3537$, nominal repair $.2856/.3404$.

Arm	Mode	27 (0.25%)	53 (0.5%)	105 (1%)	210 (2%)	525 (5%)
RAND (random)	internal	.2847/.3226	.2837/.3206	.2821/.2970	.2779/.2903	.2661/.2747
MPC (most probably corrupted)	internal	.2853/.3432	.2848/.3482	.2840/.3587	.2800/.3300	.2680/.3450
NOM-GREEDY (nominal one-step)	internal	.2838/.3669	.2809/.3680	.2758/.3686	.2655/.3463	.2425/.3332
CARD-uniform (uniform)	internal	.3092/.3569	.3068/.3542	.2959/.3552	.2889/.3266	.2577/.3264
CARD-posterior (posterior)	internal	.2451/.3156	.2399/.2901	.2232/.2941	.2157/.2688	.1945/.2535
CARD-tempered (tempered)	internal	.2440/**.2700**	.2555/**.2416**	.2341/.2383	.2185/.2601	.1999/.2436
CARD-minimax (minimax)	internal	.2595/.2760	.2513/.2439	.2398/**.2208**	.2286/**.2249**	.1975/**.2190**
menu-24 CARD-minimax	internal	.2490/.2990	.2506/.2743	.2430/.2552	.2262/.2429	.1985/.2363
menu-24 CARD-tempered	internal	.2715/.2730	.2606/.2481	.2396/.2297	.2195/.2294	.1970/.2394
ALC	as published	.3322/.3870	.3297/.4034	.3250/.4018	.3157/.3711	.2885/.3453
ALC	on our repair	.2626/.3362	.2440/.3368	.2302/.3207	.2239/.2927	.2161/.2573
Confident Learning	as published	.2350/.3471	.2344/.3453	.2330/.3345	.2307/.3266	.2210/.2999
Confident Learning	on our repair	.2632/.3371	.2445/.3236	.2341/.2977	.2302/.2788	.2139/.2507
ACTIVELAB	as published	.3322/.3875	.3298/.3905	.3254/.3754	.3161/.3664	.2893/.3362
ACTIVELAB	on our repair	.2632/.3371	.2445/.3236	.2341/.2977	.2302/.2788	.2139/.2507
SIMIFEAT	as published	.2627/.4087	.2626/.4148	.2622/.4334	.2608/.4565	.2515/.4249
SIMIFEAT	on our repair	.2497/.3477	.2345/.3443	.2325/.3570	.2575/.3717	.2513/.3328
CINCER	as published	.3323/.3735	.3302/.3477	.3253/.3860	.3164/.3345	.2901/.3399
CINCER	on our repair	.2628/.3239	.2462/.2891	.2276/.3030	.2324/.2595	.2275/.2623
AUM	as published	.3322/.3849	.3297/.4049	.3248/.4438	.3149/.4941	.2865/.4887
AUM	on our repair	.2621/.3357	.2422/.3386	.2279/.3483	.2376/.3523	.2621/.3587
FORGETSCORES	as published	.3338/.3864	.3333/.3563	.3326/.3198	.3299/.2969	.3233/.2816
FORGETSCORES	on our repair	.2615/.3225	.2617/.2961	.2574/.2648	.2462/.2377	.2395/.2290
SMALLLOSS	as published	.3322/.3891	.3297/.4041	.3248/.4370	.3148/.4844	.2869/.4172
SMALLLOSS	on our repair	.2614/.3370	.2420/.3361	.2227/.3439	.2304/.3551	.2410/.2875

D.5 LABELME-AMT TEST SET

LabelMe-AMT (Rodrigues & Pereira, 2018): 2,547 crowd labels on 1,000 images from 59 workers (the worker is the source block), 8 classes, initial hidden-gold loss 0.2595; budgets 7, 13, 26, 51 and 128 audits ($\rho \in \{0.25, 0.5, 1, 2, 5\}\%$); worker-block percentile bootstrap (primary) and image-cluster (secondary), 10,000 draws; every contrast in the supplement.

Table 11: LabelMe-AMT: hidden-gold loss / certified risk of every arm at every budget. Bold: lowest certified risk in the column; underline: lowest gold.

Arm	Mode	7 (0.25%)	13 (0.5%)	26 (1%)	51 (2%)	128 (5%)
RAND (random)	internal	.2399/.2310	.2390/.2297	.2373/.2281	.2340/.2237	.2257/.2176
MPC (most probably corrupted)	internal	.2403/.2353	.2399/.2453	.2371/.2469	.2336/.2501	.2132/.2537
NOM-GREEDY (nominal one-step)	internal	.2379/.2285	.2367/.2287	.2328/.2464	.2254/.2542	.2022/.2385
CARD-uniform (uniform)	internal	.2395/.2370	.2367/.2343	.2313/.2333	.2238/.2253	.1998/.2046
CARD-posterior (posterior)	internal	.2038/.2240	.2104/.2174	.1900/**.1956**	.1790/.1916	.1606/.1918
CARD-tempered (tempered)	internal	.1943/**.2024**	.1798/**.1992**	.1790/.2130	.1814/.1996	.1590/.1954
CARD-minimax (minimax)	internal	.2242/.2255	.2175/.2163	.2010/.2038	.1904/**.1851**	.1629/**.1589**
ALC	as published	.2568/.2533	.2544/.2534	.2493/.2573	.2411/.2754	.2152/.2540
ALC	on our repair	.2297/.2331	.2226/.2304	.2120/.2295	.1932/.2362	.1633/.2070
Confident Learning	as published	.1845/.2462	.1837/.2460	.1814/.2482	.1782/.2580	.1696/.2400
Confident Learning	on our repair	.2297/.2331	.2226/.2304	.2101/.2318	.1924/.2376	.1665/.2062
ACTIVELAB	as published	.2568/.2533	.2544/.2534	.2493/.2603	.2403/.2795	.2152/.2531
ACTIVELAB	on our repair	.2297/.2331	.2226/.2304	.2101/.2318	.1924/.2376	.1665/.2062
SIMIFEAT	as published	.1810/.2398	.1810/.2424	.1810/.2469	.1790/.2485	.1704/.2460
SIMIFEAT	on our repair	.2293/.2268	.2207/.2278	.2038/.2284	.1885/.2234	.1606/.2109
CINCER	as published	.2568/.2532	.2544/.2569	.2497/.2636	.2419/.2487	.2163/.2690
CINCER	on our repair	.2301/.2329	.2238/.2328	.2112/.2331	.2002/.2154	.1708/.2169
AUM	as published	.2580/.2309	.2560/.2380	.2513/.2396	.2426/.2508	.2171/.2615
AUM	on our repair	.2332/.2148	.2254/.2187	.2104/.2136	.1943/.2172	.1618/.2099
FORGETSCORES	as published	.2583/.2445	.2568/.2581	.2552/.2430	.2532/.2345	.2470/.2321
FORGETSCORES	on our repair	.2301/.2250	.2226/.2310	.2132/.2113	.2030/.1990	.1885/.1859
SMALLLOSS	as published	.2568/.2462	.2544/.2465	.2497/.2396	.2411/.2479	.2136/.2592
SMALLLOSS	on our repair	.2305/.2276	.2226/.2247	.2148/.2170	.1987/.2181	.1684/.2120

D.6 CLOTHING1M

Table 12: Clothing1M ($k = 20$): hidden-gold loss / certified risk (4 dp) of every arm at every budget. Bold: lowest certified risk in the column; underline: lowest gold. Random audits, uniform weights and the cleaners on our repair are not run under this menu and are not shown.

Arm	Mode	20 (0.25%)	40 (0.5%)	80 (1%)	160 (2%)	400 (5%)
MPC (most probably corrupted)	internal	.3711/.2684	.3698/.2632	.3666/.2805	.3583/.2433	.3334/**.2182**
NOM-GREEDY (nominal one-step)	internal	.3691/.2745	.3675/.2737	.3635/**.2706**	.3559/**.2411**	.3311/.2215
CARD-posterior (posterior)	internal	.3258/**.2470**	.2144/.3331	.2090/.3397	.2010/.3495	.1815/.3423
CARD-tempered (tempered)	internal	.2205/.3330	.2103/.3690	.2065/.3616	.1993/.3503	.1818/.3447
CARD-minimax (minimax)	internal	.2951/.2828	.2863/.2943	.2783/.2882	.2733/.3048	.2545/.2798
ALC	as published	.3740/.2729	.3719/**.2620**	.3670/.2789	.3583/.2761	.3334/.2479
Confident Learning	as published	.3388/.2899	.3379/.2911	.3361/.2863	.3329/.2903	.3200/.2512
ACTIVELAB	as published	.3740/.2714	.3716/.2789	.3671/.2762	.3580/.2888	.3344/.2234
SIMIFEAT	as published	.3765/.3597	.3754/.3560	.3736/.3555	.3703/.3696	.3538/.3334
CINCER	as published	.3741/.2801	.3718/.2890	.3673/.2881	.3593/.2733	.3358/.2510
AUM	as published	.3741/.2651	.3716/.2813	.3671/.3021	.3578/.3406	.3326/.3379
FORGETSCORES	as published	.3754/.3395	.3750/.2954	.3731/.3066	.3696/.3193	.3584/.3465
SMALLLOSS	as published	.3740/.2727	.3716/.2790	.3668/.2981	.3575/.3197	.3328/.2812

E BASELINE METHODS: PROVENANCE, ADAPTATION AND ADEQUACY

This appendix records, for every external method that appears in a reported result, the facts a reader needs to reproduce or to dispute the comparison: paper and venue, lane, whether it is a budgeted method with a repair rule of its own, the repository and commit executed, its licence, the check that our invocation reproduces its scoring, and how it was mapped onto the audit-and-repair protocol of Section 4.2. Four lanes were distinguished when the baseline set was fixed: (a) budgeted cleaning, the exact task, with ALC (Bernhardt et al., 2022) as the direct prior art; (b) label-error detectors, which return a ranking usable as an acquisition rule; (c) LLM-as-oracle relabelling of fresh unlabelled

Table 13: The eight published cleaners: lane (budgeted cleaning, or label-error detector used as a ranking), whether the published method is budgeted and defines its own repair, and the code executed (vendored upstream files or an installed package).

Method	Paper, venue	Lane	Budg.	Own repair	Repository @ commit	Licence	Mode
ALC (Bernhardt et al., 2022)	Bernhardt et al., *Nature Communications* 13:1161, 2022	(a)	yes	audited only	`microsoft/InnerEye-DeepLearning` (`InnerEye-DataQuality`) @ `1606729c`; removed from HEAD by PR #692	MIT	vendored (3 files, 4 shims)
Confident Learning (Northcutt et al., 2021)	Northcutt, Jiang, Chuang, *JAIR* 70, 2021	(b)	no	de facto (see text)	`cleanlab/cleanlab` release 2.9.0 (PyPI); clone `75062574` for reference	Apache-2.0	imported
ACTIVELAB (Goh & Mueller, 2023)	Goh, Mueller, ICLR 2023 Workshop TrustML (workshop)	(a)	yes	consensus (trivial here)	`cleanlab.multiannotator` (ActiveLab since 2.3.0; run under 2.9.0)	Apache-2.0	imported
SIMIFEAT (Zhu et al., 2022)	Zhu, Dong, Liu, ICML 2022	(b)	no	yes, kNN vote	`UCSC-REAL/SimiFeat` @ `534af484`	CC BY-NC 4.0 (non-commercial)	vendored (3 files)
CINCER (Teso et al., 2021)	Teso et al., NeurIPS 2021 (spotlight)	(b)	no	no	`autonlab/aqua` PyTorch port @ `84d3abd1` (MIT); reference `abonte/cincer` @ `58264ffa` (MIT, TF 2.4.1, not installable)	MIT / MIT	vendored (2 files)
AUM (Pleiss et al., 2020)	Pleiss et al., NeurIPS 2020; via the TMLR-2024 benchmark library (George et al., 2024)	(b)	no	no	`Orange-OpenSource/mislabeled` @ `9e5391f1`	MIT	imported
FORGETSCORES (Toneva et al., 2019)	Toneva et al., ICLR 2019; via `mislabeled`	(b)	no	no	as above	MIT	imported
SMALLLOSS (Han et al., 2018; George et al., 2024)	small-loss selection (no single canonical paper); via `mislabeled`	(b)	no	no	as above	MIT	imported

data; and (d) learning with noisy labels, which trains through the noise and releases no dataset. Lanes (a) and (b) are in scope; (c) and (d) do not produce the estimand we score. A method in scope was run if it had an executable implementation whose scoring path could be driven from our frozen inputs (a probability matrix or frozen features) and if its adaptation could be stated in one line without weakening it: the eight methods that met both conditions are in Table 13, their adaptation and validation in Table 14. Every method emits a full descending-priority ranking; the top-B records go to the oracle; whatever the upstream method would natively do with its flagged set (delete, filter, negotiate, request annotations) is replaced by this audit step. A method without a native repair rule can change at most B labels, so its as-published gold loss is bounded below by the initial error minus B/n (the *keep-observed cap*); on the development set at 5 % this is why ALC, ACTIVELAB, CINCER, AUM and SMALLLOSS reach the identical gold loss 0.122 (Table 6).

F REPRODUCIBILITY DETAILS

Platforms, environments and seeds. Two platforms were used: cluster A, a Slurm cluster whose CPU partition ran the exact-game arrays of Appendix C.6 inside a frozen container (`nvcr.io/nvidia/pytorch:24.12-py3`) and whose GPU partition ran the feature-extraction jobs, and cluster B, a single shared node without a scheduler on which every dataset-scale grid, the remaining exact-game grids and both semi-synthetic studies ran as pools of single-threaded processes, one run per process (BLAS threads set to 1). The exact-game engine and the dataset-scale engine are pure NumPy; the latter's one-step scores and posterior updates reproduce the frozen exact-game engine to 10^{-12} on the development set and 10^{-14} on the test sets, and its smoke runs agree to all printed digits between NumPy 1.26.4 and 2.1.2. The eight external cleaners run in one pinned virtual environment (Python 3.12, NumPy 2.5.2, scikit-learn 1.9.0, torch 2.13.0 CPU, cleanlab 2.9.0 (Northcutt et al., 2021), faiss-cpu, the `mislabeled` library (George et al., 2024) at commit

Table 14: Adaptation and validation of the eight cleaners: the score that ranks records, the native repair of the as-published mode, the check that our invocation reproduces the upstream scoring on synthetic inputs (AUROC on planted flips), the Spearman correlation ρ_{MPC} of the ranking with the MPC ranking on the development set, and the identifier of the ranking file (Confident Learning and ACTIVELAB rank identically in the single-label regime).

Method	Ranking (as executed)	Native repair	Validation	ρ_{MPC}	hash
ALC	`PosteriorBasedSelector`, decision rule `JOINT`: normalised cross-entropy of the predictor against the current label minus an ambiguity penalty, $\gamma = 0.30$ hard-coded upstream; descending sort by the official batch-selection function. Static variant: frozen posteriors, no belief update (the update loop needs the retired training stack).	Upstream draws fresh annotations until a strict majority, for selected records only; one oracle reveal replaces that draw. Unaudited records keep the observed label (upstream rule).	JOINT and CE scoring and official batch selection reproduced on Python 3.12 / NumPy 2.5; deterministic; flip AUROC 1.000.	0.515	`7bf1c3bfe1`
Confident Learning	Ascending `get_label_quality_scores(method="self_confidence")` on the frozen probability matrix; repair flag from `find_label_issues` (`n_jobs=1`).	The library exposes no repair API; the de facto rule of its tutorials is applied: unaudited records flagged by `find_label_issues` are relabelled to $\arg\max$ of the predicted probabilities, unflagged records keep the observed label (our instantiation, not an upstream function).	Native `pred_probs` API called unchanged (zero adaptation of the scoring path; evidence script); deterministic; flip AUROC 1.000; repair lowers unaudited error 0.040 $\rightarrow$ 0.012 on the synthetic check.	0.923	`6afad8c583`
ACTIVELAB	Ascending `get_active_learning_scores` with one annotator column (single-label regime; no multi-annotator mount in this campaign).	Upstream consensus label degenerates to "keep observed" with one label per record; none applied.	Single-annotator-column call verified end-to-end (evidence script case 5); deterministic; flip AUROC 1.000.	0.923	`6afad8c583`
SIMIFEAT	Cosine kNN label distribution over frozen DINOv2 features ($k = 10$, `min_similarity=0`, `Tii_offset=1`); the upstream majority-vote (MV) flagged set first, ordered by the upstream `ce` score, then the remainder by the same score (our totalisation of a set-valued rule; the unseeded per-class `rank1` variant excluded).	kNN majority-vote label on unaudited records (the method's own correction).	MV recall 0.96 / precision 1.00; deterministic; flip AUROC 1.000; repair lowers unaudited error 0.040 $\rightarrow$ 0.000 on the synthetic check.	0.658	`b010c66ec4`
CINCER	Margin inspector, one-shot: descending $p_i(\hat{y}_i) - p_i(y_i)$ from the frozen probabilities; the upstream negotiation loop, retraining, influence/Fisher inspectors and counterexamples are excluded (AQuA's own default, `no_ce=true`).	None: the upstream flag-and-relabel step is the audit itself.	Margin reproduces $p(\hat{y}) - p(y)$ to machine precision; deterministic; flip AUROC 1.000.	0.443	`72075e610b`
AUM	Ascending `AreaUnderMargin(...).trust_score` with a logistic probe (`SGDClassifier`, log loss, seeded, `steps=1`) trained on the frozen features.	None.	Four detectors run end-to-end; flip AUROC 0.840 (adapter), 0.936 (evidence script).	0.820	`6d67de5415`
FORGET-SCORES	Ascending `ForgetScores(...).trust_score`, same logistic probe.	None.	Flip AUROC 0.867 (adapter), 0.877 (evidence script).	-0.016	`4ea21442ed`
SMALLLOSS	Ascending `SmallLoss(...).trust_score` with `LogisticRegression` on the frozen features.	None.	`LogisticRegression` probe over frozen features run end-to-end; flip AUROC 1.000 (adapter and evidence script).	0.901	`bbe125f7d6`

9e5391f, vendored ALC and SIMIFEAT sources); the numerics suite (supplement) uses Python 3.11.10, NumPy 1.26.4 and SciPy 1.14.1 (HiGHS). Acquisition is deterministic for every arm except RAND, so the only randomness in the reported numbers is the RAND draws (eight frozen seeds at dataset scale), the salted-hash selection of workers and targets, the out-of-fold probes behind the content models, the semi-synthetic generators, and the bootstrap (seed table in the supplement); the Hedge rate $\eta = \sqrt{8 \ln K/B}$, the refresh tolerance 0.10 and the 64 quadrature nodes are constants. The grid launchers, the per-environment version pins, the seed table and the SHA-256 manifests of code and inputs are reproduced in the supplement.

Oracle firewall. The firewall is enforced in code: loaders take no gold argument and are tested by intercepting every `numpy.load` during loading; the driver commits one row at a time to an oracle object that reveals only committed rows; and the evaluation step, run after the budget is exhausted, raises unless the reveal count equals the budget. Each test-set baseline run stores the SHA-256 of the ranking it consumed, so the identity of the Confident Learning and ACTIVELAB rankings (Appendix E) can be verified on every table.

From runs to printed numbers. An aggregator per study reads the run files and writes a machine-readable summary from which every table is generated. At dataset scale the certified risk of a released dataset (Eq. 3) is computed from the per-block per-law terminal expected loss that every run stores, as the table-wide per-law risk vector followed by its maximum over the 16-law index set, and the paired contrasts recompute the same quantity inside the bootstrap, resampling blocks before taking the maximum over laws; this is the whole-dataset scale of the hidden-gold loss, so the two columns of every table are comparable (at episode scale, Appendix C.6, the certified risk is the upper end U_T of the Hedge bracket). The numbers file of the supplement keeps the full precision of every source; Table 1, Figures 1 and 2, the per-dataset tables and the coverage table of Appendix D are generated from it by the scripts shipped with the paper source.

Cost. Table 15 gives the wall time of the internal arms on CIFAR-10H and Dopanim. Cleaners as published take seconds per run (SIMIFEAT about 150 s for its kNN repair on CIFAR-10H) and cleaners on our repair under half a minute. The CIFAR-10H grid took 172 CPU-hours and the Dopanim grid 134; on the development set the 1 % run took 7,196 s for CARD-minimax against 2,661 s for NOM-GREEDY; a LabelMe-AMT run of CARD-minimax takes 18–70 s.

Table 15: Wall time per run, in hours, for the seven internal arms on CIFAR-10H and Dopanim (RAND: sum over its eight seeds).

	CIFAR-10H, budget (audits)					Dopanim, budget (audits)				
Arm	256	512	1024	2048	5120	27	53	105	210	525
RAND	0.01	0.02	0.03	0.06	0.17	0.00	0.00	0.00	0.00	0.01
MPC	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
NOM-GREEDY	0.39	0.52	0.78	1.33	2.76	0.51	0.78	1.34	2.50	5.99
CARD-uniform	0.39	0.52	0.78	1.32	2.74	0.33	0.43	0.63	1.06	7.12
CARD-posterior	1.04	1.49	2.64	3.42	4.89	1.13	1.83	2.66	6.02	10.30
CARD-tempered	1.07	1.62	2.83	3.63	5.13	1.18	2.04	3.28	5.06	8.05
CARD-minimax	0.66	0.95	1.42	2.11	3.65	0.87	1.34	2.13	3.22	6.65