LEARNING FROM PARTIALLY CORRECTED LABELS: A FRAMEWORK FOR HUMAN-IN-THE-LOOP DATA CURATION WITH STATISTICAL GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models rely heavily on high-quality data, making annotation and curation central to their development. As models become more capable, curation is increasingly shifting from labeling from scratch toward reviewing, correcting, or rejecting existing outputs. These workflows naturally produce before/after correction records, a structure largely overlooked in conventional learning from noisy labels. We formalize this setting as *Learning from Partially Corrected Labels* (PCL), where every instance has a potentially noisy label and a subset additionally reveals its correction. We show that this additional observation fundamentally changes identifiability under instance-dependent label noise: while the noise process is generally non-identifiable from noisy labels alone, partial correction pairs make the underlying correction process identifiable under mild audit conditions. Building on this result, we develop a learning framework with finite-sample guarantees, achieving an $\mathcal{O}(1/\sqrt{n_{\mathrm{e}}})$ correction-estimation rate with n_{e} observed corrections. We evaluate PCL across image and text settings, including human-noise benchmarks and historical curation data, and further translate the approach into a real annotation pipeline through a prospective user study. Our results show that correction histories provide reusable supervision for improving downstream learning and data curation, suggesting a promising direction for future human-AI annotation systems.

1 INTRODUCTION

Artificial intelligence (AI) systems, particularly large language models (LLMs), rely heavily on the quality and composition of training and adaptation data, making data annotation and curation central to modern machine learning (Whang et al., 2023; Li et al., 2024; Zha et al., 2025). Traditionally, large-scale supervision has often been collected by asking human annotators to label examples from scratch, as in crowdsourced language annotation and large-scale image classification (Snow et al., 2008; Sheng et al., 2008; Deng et al., 2009). More recently, annotation pipelines increasingly combine machine-generated labels or suggestions with human verification and correction (Kim et al., 2024a; Wang et al., 2024; Schroeder et al., 2025). Human supervision remains central as models become more capable: instruction-tuning and human-feedback pipelines use human demonstrations, preferences, and feedback to shape model behavior (Ouyang et al., 2022; Scheurer et al., 2023).

At the same time, a growing class of annotation workflows no longer asks humans to construct every label or output from scratch. Instead, humans inspect an existing machine-generated or previously annotated candidate and verify, correct, or reject it (Wang et al., 2024; Kim et al., 2024a). Machine translation systems, for example, record system outputs together with translator post edits (Green et al., 2013; Domingo et al., 2019), language-feedback pipelines refine model-generated responses based on human feedback (Scheurer et al., 2023), and data-cleaning systems iteratively identify and repair problematic records (Krishnan et al., 2016; Mahdavi & Abedjan, 2020). Such workflows naturally produce a before/after record: an instance x, its original label or output $\tilde{y}$, and its revised label y when corrected. Crucially, this creates paired label observations for the same instance, with both the original and revised labels being observable. Recent work has begun to exploit related forms of paired supervision: Tjandra & Wiens (2023) use a small *alignment set* containing both observed and ground-truth labels to mitigate instance-dependent label noise, while Li et al. (2026) solicit human corrections for vision-language model (VLM)-generated labels when model predictions

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.

disagree. These studies demonstrate the practical value of paired or corrected supervision, but leave open a more fundamental question: *what does a partially observed correction record identify about an unrestricted instance-dependent annotation process, and under what review mechanisms?*

The central obstacle in learning with noisy labels is the noise transition function $T_{yj}(x) = \Pr(\tilde{Y} = j \mid Y = y, X = x)$. When T does not depend on x, corresponding to class-conditional noise, a mature family of correction methods exists. Noise rates or transition matrices can be identified or estimated under conditions such as irreducibility or anchor points (Scott et al., 2013; Liu & Tao, 2015), and, once estimated, can be incorporated into corrected learning objectives (Natarajan et al., 2013; Patrini et al., 2017). However, these guarantees rely critically on the corruption mechanism being shared across instances within each class. In practice, annotation errors may depend strongly on the particular instance, with ambiguous or difficult examples being more susceptible to mislabeling (Xia et al., 2020; Chen et al., 2021; Cheng et al., 2022). When the transition becomes instance-dependent, $T = T(x)$, the problem is substantially harder: the observed noisy-label distribution generally no longer determines the clean posterior and the transition function separately. With only noisy labels observed, the transition function is not identifiable in general and generic identification at the instance level can require multiple noisy labels per instance (Liu et al., 2023). Existing approaches therefore recover tractability by imposing additional structure, such as part-dependent transition models (Xia et al., 2020), structural causal models (Yao et al., 2021), Bayes-label transitions (Yang et al., 2022), or geometric regularity of the instance-dependent transition function (Cheng et al., 2022).

We address this question by treating the correction record itself as an observation of the annotation process. Every reviewed instance can reveal not only a clean target y, but also the label $\tilde{y}$ that preceded it on the same instance. Unlike a fresh clean label, the pair $(\tilde{y}, y)$ therefore contains information about how the original annotation process behaved. We formalize this setting as *Learning from Partially Corrected Labels* (PCL), where every instance carries a potentially noisy label and only a subset additionally reveals its reviewed label. We show that this seemingly small change in observation structure fundamentally changes what can be learned under instance-dependent label noise. While the underlying noise process is generally non-identifiable from noisy labels alone, partially observed correction pairs identify the correction mechanism under mild conditions on how instances are selected for review. Based on this result, we develop a learning framework with a finite-sample error bound whose correction-estimation term decreases at an $\mathcal{O}(1/\sqrt{n_{\mathrm{e}}})$-type rate in the number of observed correction pairs. We then evaluate PCL on human-noise benchmarks and historical curation data, showing that correction histories contain reusable learning signal for both downstream learning and selective data curation. Finally, we translate this principle into an existing human-in-the-loop annotation workflow and conduct a prospective user study to assess its practical utility.

2 Learning from Partially Corrected Labels

We introduce the partially corrected labels (PCL) setting as an observation model for learning from historical label corrections. We formalize the observed data and learning objective, define the correction posterior together with the audit assumptions used in the subsequent analysis, and characterize how partial corrections change the identifiability of instance-dependent label noise. Building on this structure, we then develop learning procedures and establish finite-sample guarantees.

2.1 Setting and Correction Posterior

Let $\mathcal{X}$ denote the instance space and $\mathcal{Y} = \{1, \ldots, K\}$ the label space, and let $\mathcal{D}$ be the joint distribution over $\mathcal{X} \times \mathcal{Y}$. Nature draws $(X, Y) \sim \mathcal{D}$ with clean posterior $\eta_y(x) := \mathbb{P}(Y = y \mid X = x)$. An annotation process produces a potentially noisy label $\tilde{Y}$ via the instance-dependent transition $T_{yj}(x) := \mathbb{P}(\tilde{Y} = j \mid Y = y, X = x)$, inducing the noisy-label marginal $\mu_j(x) := \mathbb{P}(\tilde{Y} = j \mid X = x) = \sum_{y \in \mathcal{Y}} \eta_y(x) T_{yj}(x)$. Some instances are reviewed. Let $O \in \{0, 1\}$ indicate whether review reveals the clean label, with propensity $\pi(x, \tilde{y}, y) := \mathbb{P}(O = 1 \mid X = x, \tilde{Y} = \tilde{y}, Y = y)$, and marginal review propensity $\pi(x, \tilde{y}) := \mathbb{P}(O = 1 \mid X = x, \tilde{Y} = \tilde{y})$. The learner observes $(X, \tilde{Y}, O, O \cdot Y)$. If $O = 1$, $(x, \tilde{y}, y)$ is observed; otherwise only $(x, \tilde{y})$ is observed. Review may leave the label unchanged ($\tilde{y} = y$). Given N i.i.d. observations, let $n_{\mathrm{e}} := \sum_{i=1}^{N} o_i$ denote the number with revealed clean labels. Our goal is to learn a predictor $f \in \mathcal{F}$ with small clean risk $R(f) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(f(X), Y)]$, where ℓ is bounded by $\bar{\ell}$.

The central object in PCL is the *correction posterior*, $g_y(x, \tilde{y}) := \mathbb{P}(Y = y \mid X = x, \tilde{Y} = \tilde{y})$, which describes the distribution of the clean label given the instance and its observed noisy label. Intuitively, g characterizes how an observed label should be corrected in the context of the particular instance. Whether this correction posterior can be learned from reviewed examples depends on how instances are selected for review. We distinguish two regimes. In a *solicited-audit* design, the system selects instances for verification based only on observed information, such as $(x, \tilde{y})$. In a *voluntary-correction* regime, annotators may instead intervene because they notice that the displayed label is incorrect, so the review probability may additionally depend on the unknown clean label y.

Assumption 1 (EAR with Coverage)**.** *The review mechanism satisfies:*

(i) Edits at random (EAR): $\pi(x, \tilde{y}, y) = \pi(x, \tilde{y})$ *for* $P_{X,\tilde{Y},Y}$*-almost every* $(x, \tilde{y}, y)$.
(ii) Coverage: *there exists* $\pi_{\min} > 0$ *such that* $\pi(x, \tilde{y}) \geq \pi_{\min}$ *for* $P_{X,\tilde{Y}}$*-almost every* $(x, \tilde{y})$.

The first condition, *Edits at Random* (EAR), rules out outcome-dependent review conditional on $(X, \tilde{Y})$, while the second ensures that every relevant $(x, \tilde{y})$ region can be reviewed. Randomized audits with a positive audit probability satisfy both conditions by construction. Voluntary corrections, in contrast, may violate EAR; we revisit this setting in Section 2.4.

2.2 From Non-identifiability to Identification

We begin with the negative result that motivates PCL. Consider the binary case, and write the clean posterior as $\eta(x) := \mathbb{P}(Y = 1 \mid X = x)$, the instance-dependent flip probabilities as $e_1(x) := \mathbb{P}(\tilde{Y} = 0 \mid Y = 1, X = x)$, $e_0(x) := \mathbb{P}(\tilde{Y} = 1 \mid Y = 0, X = x)$, and the noisy-label marginal as $\mu(x) := \mathbb{P}(\tilde{Y} = 1 \mid X = x) = (1 - e_1(x))\eta(x) + e_0(x)(1 - \eta(x))$.

Theorem 2 (Non-identifiability from noisy labels alone)**.** *Suppose only* $(X, \tilde{Y})$ *is observed, with noisy-label marginal* $\mu(\cdot)$*. For every measurable pair* (e_0, e_1) *satisfying* $e_0(x) + e_1(x) < 1$*,* $e_0(x) \leq \mu(x) \leq 1 - e_1(x)$ *pointwise, the clean posterior* $\eta(x) = \frac{\mu(x) - e_0(x)}{1 - e_0(x) - e_1(x)}$ *defines a valid instance-dependent noise model that induces exactly the same observed law of* $(X, \tilde{Y})$*. Consequently, the observed distribution is compatible with a continuum of latent clean-label and noise models. Distinct compatible models may disagree on both the Bayes classifier and the clean risk* $R(f)$ *of a fixed predictor. In particular, for any* $\varepsilon \in (0, \frac{1}{2})$*, setting* $A_\varepsilon = \{x : \frac{1}{2} < \mu(x) < \frac{1+\varepsilon}{2}\}$*,* $e_0 = \varepsilon \mathbf{1}_{A_\varepsilon}$*, and* $e_1 = 0$ *yields an observationally equivalent model with the opposite Bayes label on* A_ε*.*

Theorem 2 shows that the difficulty is not merely estimation error: the observed noisy-label distribution itself does not determine the underlying clean problem. See Appendix A.1 for the proof. Intuitively, $\mu(x)$ provides only one observable quantity at each x, while $\eta(x)$, $e_0(x)$, and $e_1(x)$ remain jointly unknown. Anchor-point and irreducibility conditions can resolve this ambiguity under class-conditional noise, where the transition is a single matrix, but not for an unrestricted instance-dependent transition function. Indeed, Liu et al. (2023) show that generic instance-level identification from noisy labels alone requires multiple noisy labels per instance. PCL, however, changes the observation structure: on a reviewed subset, the learner observes the original noisy label together with its clean correction. The next result shows that, under Assumption 1, this additional pairing is sufficient to resolve the ambiguity above.

Theorem 3 (Identification from Partial Corrections)**.** *Under EAR in Assumption 1, for* $P_{X,\tilde{Y}}$ *almost every* $(x, \tilde{y})$ *with* $\pi(x, \tilde{y}) > 0$ *and every* $y \in \mathcal{Y}$*,*

$$\mathbb{P}(Y = y \mid X = x, \tilde{Y} = \tilde{y}, O = 1) = g_y(x, \tilde{y}). \tag{1}$$

If the coverage condition in Assumption 1 also holds, the observed law identifies:

(i) the correction posterior $g_y(x, \tilde{y})$ $P_{X,\tilde{Y}}$*-almost everywhere;*
(ii) the clean posterior $\eta_y(x) = \sum_{j \in \mathcal{Y}} g_y(x, j)\mu_j(x)$*;*
(iii) wherever $\eta_y(x) > 0$*, the instance-dependent transition* $T_{yj}(x) = g_y(x, j)\mu_j(x)/\eta_y(x)$ *;*
(iv) for every measurable f *and bounded loss, the clean risk*

$$R(f) = \mathbb{E}_{X,\tilde{Y}}\left[\sum_{y \in \mathcal{Y}} g_y(X, \tilde{Y})\ell(f(X), y)\right]. \tag{2}$$

In particular, $R(f)$ is identified without recovering or inverting T.

See Appendix A.2 for the proof. Eq. 1 gives the key operational consequence of the theorem: on reviewed instances, predicting the revealed clean label from $(X, \tilde{Y})$ is an ordinary supervised learning problem whose target is exactly the population correction posterior g. EAR allows this conditional relationship to be transported from the reviewed subset to the full population, while coverage ensures that no region of $(X, \tilde{Y})$ is systematically unobserved.

2.3 Learning with an Estimated Correction Posterior

The identification result above immediately suggests a two-stage learning procedure. Let $\mathcal{S} = \{(x_i, \tilde{y}_i, o_i, o_i y_i)\}_{i=1}^N$ denote the full pool, and let $\mathcal{S}_{\mathrm{e}} = \{(x_i, \tilde{y}_i, y_i) : o_i = 1\}$ be the reviewed subset of size n_{e}. We first estimate how the observed label should be corrected conditional on the instance. Specifically, we fit $\hat{g} \in \mathcal{G}$ to predict Y from $(X, \tilde{Y})$ on the reviewed subset: $\hat{g} \in \arg\min_{q \in \mathcal{G}} -\frac{1}{n_{\mathrm{e}}} \sum_{i:o_i=1} w_i \log q_{y_i}(x_i, \tilde{y}_i)$, where $\mathcal{G}$ is a class of candidate correction models, $w_i = 1$ under uniform audits and $w_i \propto 1/\pi(x_i, \tilde{y}_i)$ may be used when review propensities vary and are known. Under EAR, the reviewed conditional distribution has the same target g as the full population; the weighting additionally accounts for nonuniform representation of $(X, \tilde{Y})$ in the reviewed subset.

For every instance, $\hat{g}(x_i, \tilde{y}_i)$ provides an instance-specific soft distribution over the clean label. We therefore replace the observed noisy label by this estimated correction posterior and minimize the resulting empirical clean-risk surrogate over all N instances:

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{R}_{\hat{g}}(f), \qquad \hat{R}_{\hat{g}}(f) := \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \hat{g}_y(x_i, \tilde{y}_i) \ell(f(x_i), y). \tag{3}$$

Thus, the n_{e} reviewed examples are used to learn how labels should be corrected, while the resulting correction is propagated to the entire pool. If the true g were known, Eq. 2 implies that $\hat{R}_g(f)$ is an unbiased estimate of $R(f)$ using all N samples rather than only the reviewed subset. In the finite-sample setting, the remaining error comes from two sources: estimating g from n_{e} corrections and estimating the corrected risk from N pool samples. For the analysis below, $\hat{g}$ is fit on an independent split; in practice, we use cross-fitting (Chernozhukov et al., 2018) to retain data efficiency.

Theorem 4 (Learning with an estimated correction posterior)**.** *Let $\ell \in [0, \bar{\ell}]$, and let $\hat{g}$ be fixed independently of the pool used in Eq. 3. With probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} \left| \hat{R}_{\hat{g}}(f) - R(f) \right| \le \bar{\ell}\, \mathbb{E}_{X,\tilde{Y}} \left[\|\hat{g}(X, \tilde{Y}) - g(X, \tilde{Y})\|_1 \right] + 2 \sum_{y \in \mathcal{Y}} \mathfrak{R}_N(\ell_y \circ \mathcal{F}) + \bar{\ell} \sqrt{\frac{\log(2/\delta)}{2N}}. \tag{4}$$

where $\ell_y \circ \mathcal{F} = \{x \mapsto \ell(f(x), y) : f \in \mathcal{F}\}$. Consequently, $R(\hat{f}) \le \inf_{f \in \mathcal{F}} R(f) + 2 \times$ (right-hand side of Eq. 4). If $\hat{g}$ is a conditional MLE over a well-specified class $\mathcal{G}$ of dimension $d_{\mathcal{G}}$ fit on n_{e} corrections under EAR with constant audit propensity, then $\mathbb{E}\|\hat{g} - g\|_1 = O\left(\sqrt{\frac{d_{\mathcal{G}}}{n_{\mathrm{e}}}}\right)$, yielding excess risk $O\left(\sqrt{\frac{d_{\mathcal{G}}}{n_{\mathrm{e}}}} + \sqrt{\frac{\mathrm{Comp}(\mathcal{F})}{N}}\right)$. Under misspecification, an additional approximation term proportional to $\inf_{q \in \mathcal{G}} \mathbb{E}\|q - g\|_1$ appears.

See Appendix A.3 for the proof. Theorem 4 makes the role of the two sample sizes explicit: n_{e} controls how accurately the correction posterior is learned, whereas N controls concentration of the corrected downstream risk. In the regime of primary interest, $n_{\mathrm{e}} \ll N$, the audit fraction may therefore vanish while consistency remains possible, provided that $n_{\mathrm{e}} \to \infty$.

The plug-in objective above is the learning rule used by PCL. When π is known, the same observation model also admits standard IPW (Horvitz & Thompson, 1952) and doubly robust (DR) (Bang & Robins, 2005) estimators for the clean risk of a fixed predictor; their definitions, variance properties, and synthetic validation are given in Appendix A.4 and B.6. Beyond risk estimation, correction pairs retain both noisy and clean labels on the same instance, providing the same Fisher information about clean-label parameters as a fresh clean label while additionally informing the annotation mechanism; see Appendix A.5.

2.4 SELECTIVE REVIEW

The identification result relies on both components of Assumption 1. If coverage fails, e.g., $\pi(x, \tilde{y}) \in \{0, 1\}$ under deterministic review, Eq. 1 identifies g only on the reviewed region $\{\pi = 1\}$; g remains unconstrained on $\{\pi = 0\}$. A small randomized audit probability over all $(x, \tilde{y})$ restores coverage. A different failure occurs under voluntary correction, where review may depend on the unknown clean label y. In this case EAR fails and reviewed examples can over-represent errors. Writing $\tilde{g}_y(x, \tilde{y}) := \mathbb{P}(Y = y \mid X = x, \tilde{Y} = \tilde{y}, O = 1)$, Bayes' rule gives

$$\tilde{g}_y(x, \tilde{y}) = \frac{\pi(x, \tilde{y}, y) g_y(x, \tilde{y})}{\sum_{y' \in \mathcal{Y}} \pi(x, \tilde{y}, y') g_{y'}(x, \tilde{y})}. \tag{5}$$

Thus $\tilde{g}$ is generally a selection-biased version of g. Without restrictions on $\pi(x, \tilde{y}, y)$, g is not identified; with known relative propensities, $g_y(x, \tilde{y}) \propto \tilde{g}_y(x, \tilde{y})/\pi(x, \tilde{y}, y)$, whereas bounded review-odds ratios yield sensitivity bounds when the selection mechanism is unknown. Formal results are given in Appendix A.6. In practice, a small set of randomly audited logs can be integrated with voluntary corrections to construct an unbiased baseline.

3 EXPERIMENTS

We evaluate PCL in three stages of increasing realism. First, a controlled synthetic testbed with known clean and corruption mechanisms allows us to directly examine the theoretical predictions in Section 2 (Section 3.1). Second, we evaluate PCL on real-world noisy-label benchmarks across image and text domains to test whether the estimated correction posterior improves downstream learning under human annotation noise (Section 3.2). Finally, we study a live human-in-the-loop curation pipeline, first testing whether historical correction records contain reusable predictive signal and then evaluating their use prospectively in the deployed workflow (Section 3.3).

3.1 CONTROLLED VALIDATION OF CORRECTION POSTERIOR

Data generation and setting. We use a controlled one-dimensional testbed in which the clean posterior and the instance-dependent corruption process are known analytically. We draw $X \sim \mathrm{Unif}([-3, 3])$, $\eta(x) = \Pr(Y = 1 \mid X = x) = \sigma\big(3(x^2 - 1)\big)$, so that the clean Bayes decision boundaries lie at $x = \pm 1$. To introduce localized asymmetric corruption, we use Gaussian-shaped noise bumps centered near the two decision boundaries, with $1 \to 0$ errors concentrated near the left boundary and $0 \to 1$ errors near the right. The exact corruption functions and additional testbed details are given in Appendix B.1. We estimate the correction posterior g using three model classes of increasing flexibility: a quadratic logistic model ($d_{\mathcal{G}} = 6$), a logistic model with a cubic spline basis ($d_{\mathcal{G}} = 20$), and a logistic model with Gaussian random Fourier features ($d_{\mathcal{G}} = 602$). All model classes and hyperparameters are fixed before evaluation. For the rate experiment, we directly sample n_{e} independent correction triples $(X, Y, \tilde{Y})$ from the population for each edit budget and average results over 20 independent seeds. Under constant-propensity EAR, this is precisely the distribution of the reviewed sample conditional on its realized size n_{e}. Correction-posterior error is measured in the population norm appearing in Theorem 4, $\mathcal{E}(\hat{g}) = \mathbb{E}_{X,\tilde{Y}}\left[\left\|\hat{g}(X, \tilde{Y}) - g(X, \tilde{Y})\right\|_1\right]$, and is evaluated by numerical quadrature rather than by a finite test set.

Results. Figure 1 first evaluates the correction-estimation component of Theorem 4. For a well-specified finite-dimensional model, the theorem gives $\mathbb{E}\|\hat{g} - g\|_1 = O\left(\sqrt{\frac{d_{\mathcal{G}}}{n_{\mathrm{e}}}}\right)$, while misspecification contributes an additional model-class approximation term. We therefore fit the empirical error curves to $\mathrm{err}(n_{\mathrm{e}}) = c n_{\mathrm{e}}^{-\alpha} + \varepsilon_{\mathcal{G}}$. At $n_{\mathrm{e}} = 800$, the quadratic estimator fails to capture the local shape of the true correction posterior, particularly around the decision boundaries, whereas the spline and RBF estimators track the target more closely. Across edit budgets, the fitted convergence rates are close to the predicted $n_{\mathrm{e}}^{-\frac{1}{2}}$ behavior for all three model classes, with bootstrap confidence intervals containing the reference exponent $\frac{1}{2}$. At larger budgets, however, the misspecified quadratic model levels off at a nonzero error floor while the more flexible estimators continue to improve. Overall, the results are consistent with the two components of the theoretical bound: a sampling-error term that decreases with n_{e} and a model-dependent approximation term that becomes dominant under misspecification.

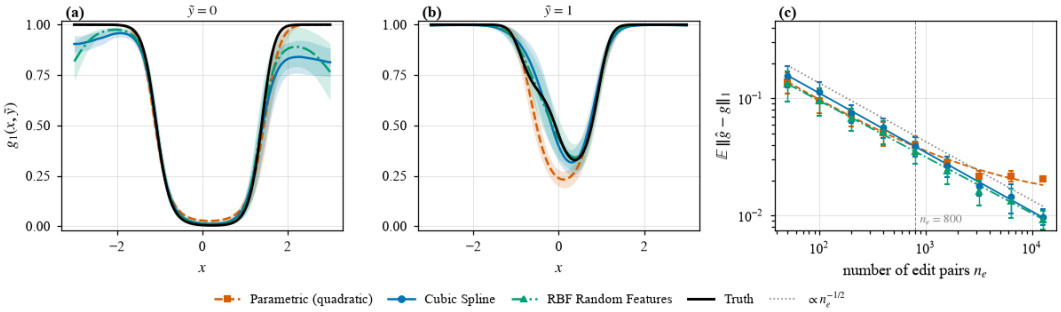


Figure 1: **Correction-posterior recovery under edits at random. (a,b)** Estimated correction posteriors $g(x, \tilde{y})$ at $n_e = 800$ for the two observed-label branches (Panel(a): $\tilde{y} = 0$, (b): $\tilde{y} = 1$), averaged over 20 seeds. The quadratic model shows systematic shape error, whereas the spline and RBF estimators more closely track the population target. **(c)** Population vector-L^1 error versus edit budget. Solid curves show fits of $cn_e^{-\hat{\alpha}} + \hat{\varepsilon}_{\mathcal{G}}$; the dotted line has slope $-1/2$. Error bars denote one standard deviation over 20 seeds.

Complete model specifications, rate estimates and confidence intervals are reported in Appendix B.2 and B.3. We further examine departures from the review assumptions in Appendices B.4 and B.5. Under selective review, randomized exploration restores coverage at the cost of reallocating a fixed finite audit budget; under error-triggered MNAR review, unweighted correction-posterior fitting becomes selection-biased, whereas inverse-propensity weighting with known review propensities substantially reduces this bias. Appendix B.6 separately validates the plug-in, IPW, and doubly robust clean-risk estimators under known audit propensities.

3.2 Learning Under Real Human Label Noise

The controlled experiments in Section 3.1 isolate correction-posterior estimation when the data-generating process is known. We next ask whether paired corrections remain useful under real human label noise. Our experiments address two questions: whether correcting the *same instance* provides information beyond collecting the same number of clean labels on fresh instances, and how PCL compares with established trusted-data and noisy-label methods, including whether its correction signal complements robust learning. We further study the difficulty of correction-posterior estimation in extreme label spaces in Appendix C.2. For a noisy training pool of size N and audit fraction $\rho = \frac{n_e}{N}$, PCL observes $(x, \tilde{y})$ for all examples and additionally receives y on the n_e audited instances. Trusted-data methods receive the same audited subset, while noisy-label methods use only the original noisy pool.

Benchmarks and models. We evaluate on CIFAR-10N and CIFAR-100N (Wei et al., 2022), AlleNoise (Rączkowska et al., 2025), and a paired-training adaptation of Clothing1M (Xiao et al., 2015). For CIFAR-10N, we consider both the *Aggregate* annotations and the more challenging *Worst* annotations. The CIFAR-N experiments use PreAct ResNet-18 (He et al., 2016b) models trained from scratch. Clothing1M uses an ImageNet-pretrained ResNet-50 (He et al., 2016a). AlleNoise uses a four-layer Transformer (Vaswani et al., 2017) trained from scratch. Complete data construction, model architectures, optimization settings, and baseline implementations are given in Appendix C.1.

Does same-instance pairing provide additional information? We isolate the value of the observation structure by comparing PCL with four information-matched alternatives in Figure 2: ordinary noisy-label training (Noisy CE), training only on the audited clean subset (Audit-only), replacing audited noisy labels by their corrections (Replace-corrected), and collecting the same number of clean labels on disjoint fresh instances (Fresh-clean). The all-clean model is shown only as a skyline.

The three CIFAR-N settings reveal different regimes. On CIFAR-10N Aggregate, noisy-label training is already strong, leaving limited headroom, although PCL still improves with the audit fraction. The benefit of pairing is clearer on CIFAR-10N Worst, where PCL increasingly separates from Replace-corrected and Fresh-clean as more edit pairs become available. This suggests that paired corrections provide reusable information about the relation between noisy and clean labels, beyond simply repairing audited examples or adding the same number of clean labels. CIFAR-100N highlights the finite-sample cost of estimating a richer correction posterior. PCL can underperform simpler uses of

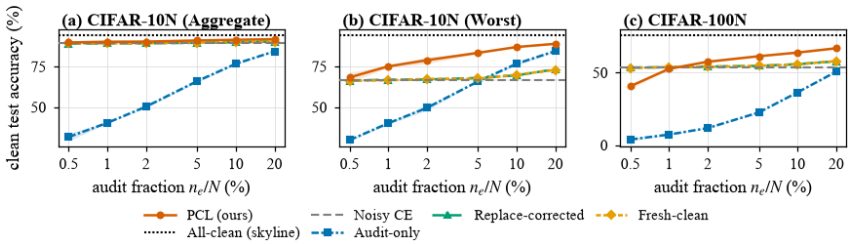


Figure 2: **Value of same-instance correction pairs.** Clean test accuracy as the audit fraction ρ increases on CIFAR-10N Aggregate, CIFAR-10N Worst, and CIFAR-100N. PCL observes paired $(\tilde{y}, y)$ corrections on audited examples, whereas the comparison policies either discard the original noisy label or spend the same clean-label budget on different instances. The dotted line denotes the all-clean skyline. Results show means over three random seeds; error bars denote one standard deviation.

Dataset	ρ	Noisy CE	Baselines & SOTA				Ours		All-clean
			GLC	L2RW	DivideMix	BLTM	PCL	PCL + DivMix	
CIFAR-10N (Agg)	1%	$89.33_{0.19}$	$89.63_{0.18}$	$\mathbf{90.72_{0.30}}$	$89.39_{0.81}$	$89.59_{0.21}$	$90.20_{0.27}$	$89.56_{0.73}$	$94.41_{0.21}$
	5%		$90.02_{0.10}$	$90.90_{0.11}$	$90.02_{0.19}$	$89.64_{0.18}$	$\mathbf{91.14}_{0.12}$	$89.71_{0.20}$	
	10%		$90.00_{0.31}$	$90.94_{0.11}$	$90.71_{0.33}$	$89.12_{0.39}$	$\mathbf{91.75}_{0.09}$	$90.69_{0.43}$	
CIFAR-10N (Worst)	1%	$66.72_{0.27}$	$72.11_{0.51}$	$80.86_{0.12}$	$\mathbf{88.52}_{0.84}$	$68.84_{1.92}$	$74.78_{1.58}$	$86.27_{0.23}$	$94.38_{0.06}$
	5%		$74.12_{0.31}$	$82.28_{0.34}$	$90.06_{0.38}$	$65.98_{1.19}$	$83.33_{0.90}$	$\mathbf{90.31}_{0.24}$	
	10%		$75.60_{0.44}$	$82.51_{0.19}$	$90.84_{0.53}$	$54.76_{3.42}$	$86.63_{0.03}$	$\mathbf{91.21}_{0.49}$	
CIFAR-100N	1%	$53.69_{0.25}$	$53.59_{0.56}$	$56.97_{0.29}$	$\mathbf{65.04}_{0.90}$	$52.77_{0.86}$	$52.58_{0.75}$	$63.05_{0.60}$	$76.04_{0.38}$
	5%		$56.34_{0.02}$	$57.84_{0.14}$	$66.41_{0.29}$	$53.53_{0.56}$	$61.46_{0.27}$	$\mathbf{67.27}_{0.16}$	
	10%		$57.89_{0.26}$	$58.19_{0.43}$	$67.56_{0.11}$	$53.76_{0.47}$	$63.72_{0.39}$	$\mathbf{68.76}_{0.26}$	
Clothing1M	1%	$57.13_{0.41}$	$64.32_{1.16}$	$66.31_{0.38}$	$62.94_{1.02}$	$61.59_{0.11}$	$63.16_{1.23}$	$\mathbf{67.24}_{1.23}$	$76.30_{0.38}$
	5%		$65.45_{0.10}$	$66.51_{0.46}$	$65.63_{0.29}$	$61.60_{0.68}$	$69.75_{0.88}$	$\mathbf{72.35}_{0.56}$	
	10%		$66.57_{0.64}$	$66.67_{0.72}$	$68.05_{0.35}$	$61.89_{0.24}$	$71.93_{0.42}$	$\mathbf{73.04}_{0.26}$	
AlleNoise	1%	$61.46_{0.12}$	$47.20_{2.78}$	$60.34_{0.18}$	$60.96_{0.21}$	$\mathbf{62.47}_{0.23}$	$27.59_{1.59}$	$41.64_{0.71}$	$71.99_{0.03}$
	5%		$56.65_{0.23}$	$60.47_{0.07}$	$62.37_{0.02}$	$\mathbf{63.21}_{0.49}$	$56.24_{0.26}$	$56.63_{0.41}$	
	10%		$64.25_{0.54}$	$60.92_{0.17}$	$62.96_{0.02}$	$63.68_{0.28}$	$\mathbf{66.54}_{0.17}$	$65.51_{0.19}$	

Table 1: **Comparison with noisy-label and trusted-data baselines.** Clean test accuracy (%), mean $\pm$ one standard deviation over seeds, at audit rates $\rho = \frac{n_e}{N}$. All methods use the same audit budget; PCL and transition-based methods additionally use paired corrections $(\tilde{y}, y)$. Best non-skyline result per row in bold.

trusted labels at very small audit fractions, but improves substantially with additional edit pairs. This is qualitatively consistent with the $\sqrt{\frac{d_{\mathcal{G}}}{n_e}}$ term in Theorem 4: positive coverage ensures identification, while accurate estimation of an instance-dependent correction posterior still requires sufficient paired supervision.

Comparison with established noisy-label and trusted-data methods. We compare against GLC (Hendrycks et al., 2018), L2RW (Ren et al., 2018), DivideMix (Li et al., 2020), and BLTM (Yang et al., 2022), covering trusted-data correction, meta-reweighting, robust noisy-label learning, and instance-dependent transition modeling; full implementations are given in Appendix C.1.

Table 1 shows a clear sample-size tradeoff. At small audit rates, simpler trusted-data or robust-learning methods can outperform vanilla PCL, whereas PCL becomes increasingly competitive as more correction pairs are observed, particularly on CIFAR-10N Worst and CIFAR-100N. Correction-pair supervision is also complementary to robust learning: combining PCL with DivideMix improves over vanilla PCL on the harder benchmarks and can outperform DivideMix alone, with particularly clear gains on Clothing1M. Thus, paired corrections can provide an additional source of supervision rather than serving as a replacement for robust noisy-label learning. AlleNoise represents a more challenging regime because of its much larger label space. Its low-audit degradation is analyzed in Appendix C.2, where we show that performance tracks the difficulty of estimating the correction posterior as the number of classes grows. Overall, these results show that paired corrections provide reusable supervision beyond simply repairing audited examples, although sufficiently many corrections are needed to estimate the correction posterior reliably. The resulting information can also complement existing robust noisy-label learning methods.

Table 2: **Final dataset quality in the prospective study.** Model-guided curation uses learned candidate ranking and action recommendations, while the baseline uses random presentation without recommendations.

Metric	Model-guided	Baseline	Fisher p
Accepted-set accuracy	**193/193 (100%)**	197/210 (93.8%)	0.0002
Missed-error rate	**0/80 (0%)**	9/82 (11.0%)	0.0031
Failed-edit rate	**0/113 (0%)**	4/128 (3.1%)	–

3.3 Real-World Human-in-the-Loop Data Curation

The experiments above establish that paired corrections provide useful supervision for downstream learning, but evaluate corrections only offline. We next ask whether information learned from historical corrections can improve the curation process that generates them. We study an existing Japanese-language annotation pipeline in which reviewers accept, edit, or reject machine-generated multiple-choice questions (MCQs), and evaluate a history-informed workflow prospectively. System protocols and implementation details are provided in Appendix D.1.

Our formal analysis treats the pre-review output as $\tilde{Y}$ and its reviewed outcome as Y, with correction posterior $g(Y \mid X, \tilde{Y})$. In the MCQ workflow, however, the reviewed outcome is a structured object rather than a finite-class label. We therefore do not attempt to estimate a structured g directly. Instead, we learn the downstream curation action $A \in \{\textsc{Accept}, \textsc{Edit}, \textsc{Reject}\}$ from the pre-review candidate $Z = (\text{question}, \text{answer}, \text{distractors})$, i.e., $q_\phi(A \mid Z)$. This operationalizes the broader PCL hypothesis that historical correction records contain reusable information about future curation needs; extending the formal PCL model to structured corrections is left to future work.

Learning from historical curation. Historical records from previous runs of the same annotation pipeline contain 684 completed judgments spanning 156 question, answer families and 167 distinct MCQ versions. Before deployment, we use these records to test whether past curation outcomes provide predictive signal that transfers to unseen questions. To avoid outcome leakage, the model observes only the pre-review candidate, excluding final edits, annotator identity, interaction history, and the eventual curation decision; full details of the family-level cross-validation protocol and input construction are provided in Appendix D.2. Under this protocol, Japanese BERT (Devlin et al., 2019; Tohoku NLP Group, 2023) achieves a three-way macro-F1 of 0.3974, outperforming a constant predictor. Exact action prediction remains inherently limited because identical inputs can receive different human judgments, yielding an empirical deterministic content-only ceiling of 0.6734. Nevertheless, the learned rejection scores are useful for prioritization: ranking held-out MCQ versions by out-of-fold $p(\textsc{Reject})$ captures 27.39% of historical reject mass within a 10% review budget, corresponding to $2.69\times$ enrichment over uniform selection. These results suggest that historical curation records contain transferable signal for routing future candidates even when the exact human action is only partially predictable.

Prospective deployment. Having established that the historical records contain reusable routing signal, we fit BERT as the predictor $q_\phi(A \mid Z)$ and integrate its predictions into the annotation pipeline. The *model-guided* condition combines model-based curation with action recommendations, whereas the *baseline* samples candidates uniformly and provides no recommendation. All model probabilities are computed before the user study and fixed in the interface, so model-inference latency does not affect the comparison. Eight participants take part in the experiment. Each participant completes two annotation blocks, with the order counterbalanced across two groups: four participants use the model-guided workflow first and the rest use the baseline first. Across the study, participants review 510 candidate presentations: 235 under the model-guided workflow and 275 under the baseline. For every presented question, annotators may accept it without modification, edit and accept it, reject it, or leave it unresolved. Further details of the study protocol are provided in Appendix D.3.

Table 2 shows that model-guided curation yields a cleaner accepted set: all 193 accepted questions are correct, whereas the baseline admits 13 incorrect questions while retaining a similar number of correct items (197 versus 193). Baseline errors arise from both missed errors and failed edits, neither of which occurs under model guidance. Item-level Fisher tests give $p = 0.0002$ for accepted-set accuracy and $p = 0.0031$ for missed errors. Taken together, these results close the loop from historical corrections to future curation. Historical review records first provide a learnable signal for routing

new candidates. This signal is then integrated as decision support in the annotation system, and prospective deployment produces a cleaner accepted dataset while retaining a similar number of correctly accepted questions. Additional protocol details, participant-level results, order-stratified analyses, and complete behavioral outcomes are reported in Appendix D.

4 Related work

Identification under label noise. For class-conditional noise, transition matrices are identifiable under irreducibility, anchor-point, or related geometric conditions (Scott et al., 2013; Liu & Tao, 2015; Patrini et al., 2017). Instance-dependent noise is harder: noisy labels alone do not generally identify the channel, while generic instance-level identification requires multiple noisy labels per instance (Liu et al., 2023), as exploited by crowdsourcing estimators (Zhang et al., 2016; Guo et al., 2023). Alternative approaches impose structure through part-dependent transitions (Xia et al., 2020), causal models (Yao et al., 2021), Bayes-label transitions (Yang et al., 2022), or confidence scores (Berthon et al., 2021). PCL instead changes the observation structure: reviewed instances reveal noisy-clean pairs $(\tilde{y}, y)$, allowing us to characterize identification of an *unrestricted* instance-dependent channel with only one noisy label per instance.

Trusted subsets and correction records. Noisy data are commonly combined with limited trusted supervision, including trusted-batch correction (Hendrycks et al., 2018; Kye et al., 2022), meta-reweighting (Ren et al., 2018), relabeled subsets (Xiao et al., 2015), semi-verified learning (Charikar et al., 2017; Zeng & Shen, 2023), and active label correction (Kremer et al., 2018; Kim et al., 2024b). Closest to PCL, Tjandra & Wiens (2023) use alignment sets containing paired observed and ground-truth labels for instance-dependent label correction, while Li et al. (2026) solicit human corrections for VLM-generated labels under selective review. Correction records also arise in translator post-editing (Turchi et al., 2017; Peris et al., 2017), human edits and language feedback for model training (Ouyang et al., 2022; Scheurer et al., 2023), interactive imitation learning (Kelly et al., 2019), and human-LLM annotation systems (Wang et al., 2024). These works primarily use corrected examples as training signal, reweighting information, or review-scheduling input. PCL instead characterizes when partially observed noisy-clean pairs identify the correction process, including failures under incomplete coverage and outcome-dependent review. Robust preference optimization under estimated flip rates (Chowdhury et al., 2024) illustrates one downstream use of learned noise mechanisms.

Missing-data and validation-sample lineage. PCL is related to internal-validation or double-sampling designs (Tenenbein, 1970), in which an inexpensive noisy measurement is collected broadly and a gold-standard measurement on a subsample. The review indicator is a missingness indicator on Y: EAR corresponds to MAR (Rubin, 1976; Little & Rubin, 2019), while voluntary error-triggered review induces a nonignorable selection problem. Inverse-propensity and doubly robust estimators (Robins et al., 1994) and odds-ratio sensitivity analyses (Rosenbaum, 2005; Manski, 2003) are the classical tools we build on. Prediction-powered inference and design-based supervised learning (Angelopoulos et al., 2023; Egami et al., 2023) similarly combine large collections of imperfect annotations with small gold-standard samples to estimate population quantities. Our target is predictive learning under an instance-dependent annotation process, where a correction posterior learned from n_{e} paired corrections is transported to a pool of size N, with separate statistical dependence on the two sample sizes.

5 Conclusion

We introduced Learning from Partially Corrected Labels (PCL), a framework for learning from historical correction records. Under edits at random with coverage, correction pairs identify the correction posterior; this, in turn, enables identification of the clean posterior, the instance-dependent noise transition, and the clean risk. We derived finite-sample guarantees for learning from partial corrections and empirically characterized when same-instance corrections are beneficial. A prospective study further shows that historical curation records can improve subsequent human-in-the-loop data curation. Overall, correction histories provide reusable supervision for both learning from noisy data and improving future curation.

ETHICS STATEMENT

The study followed a process approved by our institution's ethics review board. This work is motivated by improving the quality and efficiency of human-in-the-loop data curation. Our prospective study involves human annotators interacting with model-generated recommendations. The model is used only as decision support, and final decisions to accept, edit, reject, or defer an item remain with the human annotator. Because the historical annotation records and candidate questions contain private curation data, the deployed model is run locally rather than through an external API.

More broadly, model-guided curation may influence annotator behavior or introduce systematic biases if model recommendations are inaccurate or over-trusted. We therefore view such systems as tools for supporting, rather than replacing, human judgment. As with other data-curation methods, improvements in dataset quality may also benefit downstream systems whose eventual applications can have both beneficial and harmful consequences; these downstream uses are outside the scope of this work.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results, including the assumptions they rely on, are provided in Appendix A. For the controlled testbed (Section 3.1) and the real-noise benchmark experiments (Section 3.2), we submit the source code of PCL and a few baselines as supplementary material, together with the scripts for data generation, training, and evaluation; the corresponding settings, model classes, and hyperparameters are described in Appendices B and C. For the real-world curation study (Section 3.3), we do not release the historical curation records, the user-study logs, or the deployed system: the records and logs contain personal information of the participating annotators, and the system is tied to internal deployment and infrastructure. Instead, to support reproducibility, we describe the pre-processing pipeline, the review interface, and the annotation guidelines (Appendix D.1), the construction of the historical data and the family-disjoint evaluation protocol, the model selection and all hyperparameters of the deployed predictor, which uses publicly available checkpoints (Appendix D.2), and the full study protocol, outcome definitions, and statistical analyses (Appendices D.3–D.6). The source corpus is a publicly available Japanese Wikipedia article.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology and experiments, implement methods, and support qualitative and thematic data analysis. We did not use generative AI tools to generate synthetic datasets, propose or refine hypotheses, assist with translation, clean or reformat datasets, or interpret experimental results.

Additionally, we used generative AI tools to assist with the preparation or modification of scientific figures, suggest experimental parameters, create or edit software code, draft and edit parts of the paper, summarize or analyze existing literature, discover research topics or identify gaps, brainstorm research directions, source or search for information, identify relevant literature, format references, suggest the structure of the paper, and propose titles or keywords.

We reviewed all AI-assisted work. Mathematical claims and proofs were independently checked by the authors, and LLM-generated or LLM-assisted code was verified and tested for correctness. Literature references suggested by generative AI tools were manually verified against the original sources. We take responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with the aid of generative AI.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *International conference on machine learning*, 2021.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th annual ACM SIGACT symposium on theory of computing*, 2017.

Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI conference on artificial intelligence*, 2021.

De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *International Conference on Machine Learning*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Miguel Domingo, Mercedes García-Martínez, Amando Estela Pastor, Laurent Bié, Alexander Helle, Alvaro Peris, Francisco Casacuberta, and Manuel Herranz Pérez. Demonstration of a neural machine translation system with online learning for translators. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, 2023.

Spence Green, Jeffrey Heer, and Christopher D Manning. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013.

Hui Guo, Boyu Wang, and Grace Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. In *Advances in Neural Information Processing Systems*, 2023.

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*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, 2016b.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*, 2021.

Michael Heilman and Noah A. Smith. Good question! Statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in neural information processing systems*, 2018.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *International Conference on Robotics and Automation*, 2019.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. Meganno+: A human-llm collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024a.

Hoyoung Kim, Sehyun Hwang, Suha Kwak, and Jungseul Ok. Active label correction for semantic segmentation with foundation models. In *International Conference on Machine Learning*, 2024b.

Jan Kremer, Fei Sha, and Christian Igel. Robust active label correction. In *International conference on artificial intelligence and statistics*, 2018.

Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, Jiannan Wang, and Eugene Wu. Activeclean: An interactive data cleaning framework for modern machine learning. In *Proceedings of the 2016 International Conference on Management of Data*, 2016.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204, 2020.

Seong Min Kye, Kwanghee Choi, Joonyoung Yi, and Buru Chang. Learning with noisy labels by efficient transition matrix estimation to combat label miscorrection. In *European Conference on Computer Vision*, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Kamal Mohamed Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Joshua P Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander T Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, 2024.

Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Zhongnian Li, Lan Chen, Yixin Xu, Shi Xu, and Xinzheng Xu. Human-corrected labels learning: Enhancing labels quality via human correction of vlms discrepancies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. Distractor generation for multiple choice questions using learning to rank. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2018.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Mohammad Mahdavi and Ziawasch Abedjan. Baran: Effective error correction via a unified context representation and transfer learning. *Proceedings of the VLDB Endowment*, 13(12):1948–1961, 2020.

Charles F Manski. *Partial identification of probability distributions*. Springer, 2003.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*, volume 2. MIT press Cambridge, 2012.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, 2022.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Alvaro Peris, Luis Cebrián, and Francisco Casacuberta. Online learning for neural machine translation post-editing. *arXiv preprint arXiv:1706.03196*, 2017.

Qwen Team. Qwen3.5-9b. `https://huggingface.co/Qwen/Qwen3.5-9B`, 2026. Open-weight model release.

Alicja Rączkowska, Aleksandra Osowska-Kurczab, Jacek Szczerbiński, Kalina Jasinska-Kobus, and Klaudia Nazarko. Allenoise - large-scale text classification benchmark dataset with real-world label noise. In *International conference on artificial intelligence and statistics*, 2025.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, 2018.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

Paul R Rosenbaum. Observational study. *Encyclopedia of statistics in behavioral science*, 3: 1451–1462, 2005.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

SB Intuitions. Sarashina: Large language models for Japanese. https://huggingface.co/sbintuitions, 2026.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

Hope Schroeder, Deb Roy, and Jad Kabbara. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory*, 2013.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, October 2008.

Masahiro Suzuki, Hiroki Sakaji, Masanori Hirano, and Kiyoshi Izumi. Constructing and analyzing domain-specific language model for financial text mining. *Information Processing & Management*, 60(2):103194, 2023. doi: 10.1016/j.ipm.2022.103194.

Aaron Tenenbein. A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, 65(331):1350–1361, 1970.

Donna Tjandra and Jenna Wiens. Leveraging an alignment set in tackling instance-dependent label noise. In *Conference on Health, Inference, and Learning*, 2023.

Tohoku NLP Group. Pretrained japanese bert models. https://github.com/cl-tohoku/bert-japanese, 2023. Model: tohoku-nlp/bert-base-japanese-v3.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Marco Turchi, Matteo Negri, M Farajian, and Marcello Federico. Continuous learning from human post-edits for neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108 (1):233–244, 2017.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 2024.

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.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: a data-centric ai perspective. *The VLDB Journal*, 32(4):791–813, 2023.

Wikipedia. Sengoku jidai (nihon) [Sengoku period (Japan)]. Japanese Wikipedia, https://ja.wikipedia.org/wiki/%E6%88%A6%E5%9B%BD%E6%99%82%E4%BB%A3_(%E6%97%A5%E6%9C%AC), 2026. Accessed: 2026-09-25.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in neural information processing systems*, 2019.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in neural information processing systems*, 2020.

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*, 2015.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *International Conference on Machine Learning*, 2022.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. In *Advances in neural information processing systems*, 2020.

Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. In *Advances in Neural Information Processing Systems*, 2021.

Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. MCQG-SRefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025.

Shiwei Zeng and Jie Shen. Semi-verified PAC learning from the crowd. In *International conference on artificial intelligence and statistics*, 2023.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42, 2025.

Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In *International conference on machine learning*, 2021.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102): 1–44, 2016.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

APPENDIX CONTENTS

A Proofs and Additional Theoretical Results **17**

A.1 Proof of Theorem 2 17

A.2 Proof of Theorem 3 18

A.3 Proof of Theorem 4 18

A.4 Risk Estimation with Known Audit Propensities 20

A.5 Information Value of Paired Corrections 22

A.6 Selective Review under MNAR 22

B Controlled Testbed and Additional Analyses **24**

B.1 Testbed construction and design choices 24

B.2 Model classes and population error 25

B.3 Edit-budget sweep and rate fitting 26

B.4 Coverage under a fixed expected audit budget 27

B.5 MNAR review and propensity weighting 28

B.6 Clean-risk estimation under known propensities 29

C Real-Noise Benchmarks and Large-Label-Space Analysis **30**

C.1 Real-noise benchmark setup 30

C.2 AlleNoise cardinality stress test 33

D Additional Details for Real-World Human-in-the-Loop Curation **36**

D.1 System and Pipeline Implementation 36

D.2 Historical Curation Data and Evaluation Protocol 37

D.3 Prospective Study Protocol 39

D.4 Editing depth 40

D.5 Participant-Level and Order-Stratified Results 40

D.6 Statistical Scope 41

E Additional Related work **42**

A Proofs and Additional Theoretical Results

A.1 Proof of Theorem 2

Proof. Fix any measurable pair (e_0, e_1) satisfying the conditions of the theorem, and define

$$\eta(x) := \frac{\mu(x) - e_0(x)}{1 - e_0(x) - e_1(x)}.$$

Observational equivalence. Since $e_0(x) + e_1(x) < 1$, the denominator is strictly positive. Moreover, $e_0(x) \le \mu(x) \le 1 - e_1(x)$ implies

$$0 \le \mu(x) - e_0(x) \le 1 - e_0(x) - e_1(x),$$

and hence $\eta(x) \in [0, 1]$ holds pointwise. Thus η defines a valid clean posterior.

Under this model, the probability of observing noisy label 1 is

$$\begin{aligned}\mathbb{P}(\tilde{Y} = 1 \mid X = x) &= (1 - e_1(x))\eta(x) + e_0(x)(1 - \eta(x)) \\ &= e_0(x) + (1 - e_0(x) - e_1(x))\eta(x) \\ &= e_0(x) + \mu(x) - e_0(x) = \mu(x).\end{aligned}$$

Thus every admissible pair (e_0, e_1) induces exactly the same conditional law of $\tilde{Y}$ given X. Since the marginal law of X is unchanged, all such models induce the same observed joint distribution of $(X, \tilde{Y})$. The admissible class contains a continuum of such pairs, so the observed law is compatible with a continuum of latent models.

Disagreement of Bayes classifiers. Consider first the noise-free model

$$\mathcal{M}_{\mathrm{A}}: \qquad e_0^{\mathrm{A}}(x) = e_1^{\mathrm{A}}(x) = 0, \qquad \eta_{\mathrm{A}}(x) = \mu(x).$$

Fix $\varepsilon \in (0, \frac{1}{2})$ and define

$$A_\varepsilon := \left\{ x : \frac{1}{2} < \mu(x) < \frac{1+\varepsilon}{2} \right\}.$$

Now consider the alternative model

$$\mathcal{M}_{\mathrm{B}}: \qquad e_0^{\mathrm{B}}(x) = \varepsilon \mathbf{1}\{x \in A_\varepsilon\}, \qquad e_1^{\mathrm{B}}(x) = 0,$$

with clean posterior

$$\eta_{\mathrm{B}}(x) = \frac{\mu(x) - \varepsilon \mathbf{1}\{x \in A_\varepsilon\}}{1 - \varepsilon \mathbf{1}\{x \in A_\varepsilon\}}.$$

This is an admissible model. Indeed, if $x \in A_\varepsilon$, then $\mu(x) > \frac{1}{2} > \varepsilon$, so $e_0^{\mathrm{B}}(x) = \varepsilon < \mu(x)$; if $x \notin A_\varepsilon$, then $e_0^{\mathrm{B}}(x) = 0$. Also $e_0^{\mathrm{B}}(x) + e_1^{\mathrm{B}}(x) < 1$ everywhere. Hence by the observational-equivalence argument above, $\mathcal{M}_{\mathrm{A}}$ and $\mathcal{M}_{\mathrm{B}}$ induce the same observed law of $(X, \tilde{Y})$.

For every $x \in A_\varepsilon$,

$$\eta_{\mathrm{A}}(x) = \mu(x) > \frac{1}{2},$$

whereas

$$\eta_{\mathrm{B}}(x) = \frac{\mu(x) - \varepsilon}{1 - \varepsilon} < \frac{1}{2} \iff \mu(x) < \frac{1+\varepsilon}{2}.$$

Thus the Bayes classifier predicts class 1 under $\mathcal{M}_{\mathrm{A}}$ and class 0 under $\mathcal{M}_{\mathrm{B}}$ on A_ε. In particular, if $\mathbb{P}_X(A_\varepsilon) > 0$, the two observationally equivalent models disagree on the Bayes classifier on a set of positive probability.

Disagreement of clean risks. Under 0-1 loss, consider the fixed classifier $f \equiv 1$. Its clean risk under a model with posterior η is

$$R(f) = \mathbb{E}[1 - \eta(X)].$$

Since $\eta_A(x) = \eta_B(x)$ outside A_ε,

$$\begin{aligned}R_A(f) - R_B(f) &= \mathbb{E}[\eta_B(X) - \eta_A(X)] \\ &= -\frac{\varepsilon}{1-\varepsilon} \mathbb{E}\left[(1 - \mu(X))\mathbf{1}\{X \in A_\varepsilon\}\right].\end{aligned}$$

Therefore, whenever $\mathbb{P}_X(A_\varepsilon) > 0$, this quantity is strictly negative, so the two observationally equivalent models also assign different clean risks to the same fixed predictor. □

A.2 Proof of Theorem 3

Proof. We proceed in four steps.

Step 1: the reviewed conditional distribution equals the correction posterior. Fix $(x, \tilde{y})$ with $\pi(x, \tilde{y}) > 0$. By Bayes' rule,

$$\begin{aligned}&\mathbb{P}(Y = y \mid X = x, \tilde{Y} = \tilde{y}, O = 1)\\ &\quad = \frac{\mathbb{P}(O = 1 \mid X = x, \tilde{Y} = \tilde{y}, Y = y)\mathbb{P}(Y = y \mid X = x, \tilde{Y} = \tilde{y})}{\sum_{y' \in \mathcal{Y}} \mathbb{P}(O = 1 \mid X = x, \tilde{Y} = \tilde{y}, Y = y')\mathbb{P}(Y = y' \mid X = x, \tilde{Y} = \tilde{y})}.\end{aligned} \tag{6}$$

By EAR,

$$\mathbb{P}(O = 1 \mid X = x, \tilde{Y} = \tilde{y}, Y = y) = \pi(x, \tilde{y})$$

for every y. Hence the selection propensity cancels:

$$\begin{aligned}\mathbb{P}(Y = y \mid X = x, \tilde{Y} = \tilde{y}, O = 1) &= \frac{\pi(x, \tilde{y}) g_y(x, \tilde{y})}{\pi(x, \tilde{y}) \sum_{y'} g_{y'}(x, \tilde{y})}\\ &= g_y(x, \tilde{y}).\end{aligned}$$

This proves Eq. 1.

Step 2: identification of the correction posterior. The conditional distribution on the left-hand side of Eq. 1 is determined by the observed law wherever $\pi(x, \tilde{y}) > 0$. Coverage guarantees this positivity $P_{X,\tilde{Y}}$-almost everywhere. Therefore g is identified almost everywhere.

Step 3: identification of the clean posterior and transition. Using the law of total probability,

$$\begin{aligned}\eta_y(x) &= \mathbb{P}(Y = y \mid X = x)\\ &= \sum_{j \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x, \tilde{Y} = j)\mathbb{P}(\tilde{Y} = j \mid X = x)\\ &= \sum_{j \in \mathcal{Y}} g_y(x, j)\mu_j(x).\end{aligned}$$

Since both g and the observable noisy-label marginal μ are identified, so is η.

Whenever $\eta_y(x) > 0$, Bayes' rule then yields

$$T_{yj}(x) = \mathbb{P}(\tilde{Y} = j \mid Y = y, X = x) = \frac{g_y(x, j)\mu_j(x)}{\eta_y(x)},$$

so the forward transition is also identified on its natural support.

Step 4: identification of clean risk. For any measurable f and bounded loss ℓ,

$$\begin{aligned}R(f) &= \mathbb{E}[\ell(f(X), Y)]\\ &= \mathbb{E}_{X,\tilde{Y}}\left[\mathbb{E}[\ell(f(X), Y) \mid X, \tilde{Y}]\right]\\ &= \mathbb{E}_{X,\tilde{Y}}\left[\sum_{y \in \mathcal{Y}} g_y(X, \tilde{Y})\ell(f(X), y)\right].\end{aligned}$$

The right-hand side depends only on identified quantities. Therefore the clean risk is identified directly, without first recovering or inverting T. □

A.3 Proof of Theorem 4

Proof. For a fixed estimate $\hat{g}$, define the population corrected risk

$$R_{\hat{g}}(f) := \mathbb{E}_{X,\tilde{Y}}\left[\sum_{y \in \mathcal{Y}} \hat{g}_y(X, \tilde{Y})\ell(f(X), y)\right].$$

For every $f \in \mathcal{F}$, decompose

$$\hat{R}_{\hat{g}}(f) - R(f) = \left[\hat{R}_{\hat{g}}(f) - R_{\hat{g}}(f)\right] + \left[R_{\hat{g}}(f) - R(f)\right]. \tag{7}$$

Error from estimating the correction posterior. By Eq. 2,

$$R(f) = \mathbb{E}_{X,\tilde{Y}}\left[\sum_{y\in\mathcal{Y}} g_y(X,\tilde{Y})\ell(f(X),y)\right].$$

Hence

$$|R_{\hat{g}}(f) - R(f)| = \left|\mathbb{E}\left[\sum_{y\in\mathcal{Y}}(\hat{g}_y - g_y)\ell(f(X),y)\right]\right| \le \bar{\ell}\,\mathbb{E}\Big[\|\hat{g}(X,\tilde{Y}) - g(X,\tilde{Y})\|_1\Big]. \tag{8}$$

The bound is uniform in f, so

$$\sup_{f\in\mathcal{F}} |R_{\hat{g}}(f) - R(f)| \le \bar{\ell}\,\mathbb{E}\|\hat{g} - g\|_1.$$

Concentration of the corrected empirical risk. For fixed $\hat{g}$, define

$$\mathcal{H}_{\hat{g}} := \left\{(x,\tilde{y}) \mapsto \sum_{y\in\mathcal{Y}} \hat{g}_y(x,\tilde{y})\ell(f(x),y) : f\in\mathcal{F}\right\}.$$

Since $\hat{g}(\cdot,\cdot)$ is a probability vector and $\ell \in [0,\bar{\ell}]$, every function in $\mathcal{H}_{\hat{g}}$ takes values in $[0,\bar{\ell}]$. By standard symmetrization and concentration arguments for bounded function classes (Bartlett & Mendelson, 2002; Mohri et al., 2012), with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\left|\hat{R}_{\hat{g}}(f) - R_{\hat{g}}(f)\right| \le 2\mathfrak{R}_N(\mathcal{H}_{\hat{g}}) + \bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}}. \tag{9}$$

By subadditivity of Rademacher complexity,

$$\mathfrak{R}_N(\mathcal{H}_{\hat{g}}) \le \sum_{y\in\mathcal{Y}} \mathfrak{R}_N\left(\hat{g}_y\cdot(\ell_y\circ\mathcal{F})\right).$$

Because $\hat{g}_y(x,\tilde{y}) \in [0,1]$ is fixed independently of the sample used to evaluate the empirical risk, the contraction principle for Rademacher complexity (Mohri et al., 2012) gives

$$\mathfrak{R}_N\left(\hat{g}_y\cdot(\ell_y\circ\mathcal{F})\right) \le \mathfrak{R}_N(\ell_y\circ\mathcal{F}).$$

Substituting this into Eq. 9 and combining with Eq. 8 proves Eq. 4.

Excess risk. Let

$$f^\star \in \arg\min_{f\in\mathcal{F}} R(f)$$

and let

$$\Delta := \sup_{f\in\mathcal{F}}\left|\hat{R}_{\hat{g}}(f) - R(f)\right|.$$

Since $\hat{f}$ minimizes $\hat{R}_{\hat{g}}$,

$$R(\hat{f}) \le \hat{R}_{\hat{g}}(\hat{f}) + \Delta \le \hat{R}_{\hat{g}}(f^\star) + \Delta \le R(f^\star) + 2\Delta,$$

which gives the stated excess-risk bound.

Rate for estimating g. Under EAR with constant audit propensity, the reviewed triples $(X,\tilde{Y},Y)$ have the same population distribution as the corresponding triples before review selection. Thus, conditional on the reviewed sample size, $\hat{g}$ is an ordinary conditional MLE for Y given $(X,\tilde{Y})$.

For a well-specified regular parametric class $\mathcal{G}$ of dimension $d_{\mathcal{G}}$, standard parametric MLE theory implies, under standard regularity conditions,

$$\mathbb{E}\left[D_{\mathrm{KL}}\left(g(\cdot\mid X,\tilde{Y})\;\|\;\hat{g}(\cdot\mid X,\tilde{Y})\right)\right] = O\left(\frac{d_{\mathcal{G}}}{n_{\mathrm{e}}}\right)$$

(Van der Vaart, 2000). By Pinsker's inequality (Tsybakov & Zaiats, 2009) followed by Jensen's inequality,

$$\mathbb{E}\|\hat{g} - g\|_1 \le \mathbb{E}\sqrt{2D_{\mathrm{KL}}(g\|\hat{g})} \le \sqrt{2\mathbb{E}D_{\mathrm{KL}}(g\|\hat{g})} = O\left(\sqrt{\frac{d_{\mathcal{G}}}{n_{\mathrm{e}}}}\right).$$

Substitution into Eq. 4 shows that, whenever the Rademacher-complexity term satisfies

$$\sum_{y\in\mathcal{Y}} \mathfrak{R}_N(\ell_y \circ \mathcal{F}) = O\left(\sqrt{\frac{\mathrm{Comp}(\mathcal{F})}{N}}\right),$$

the resulting excess risk is

$$R(\hat{f}) - \inf_{f\in\mathcal{F}} R(f) = O\left(\sqrt{\frac{d_{\mathcal{G}}}{n_{\mathrm{e}}}} + \sqrt{\frac{\mathrm{Comp}(\mathcal{F})}{N}}\right).$$

If $\mathcal{G}$ is misspecified, let $g^\star_{\mathcal{G}} \in \arg\min_{q\in\mathcal{G}} \mathbb{E}[D_{\mathrm{KL}}(g\|q)]$ denote the population KL projection. Then

$$\mathbb{E}\|\hat{g} - g\|_1 \le \mathbb{E}\|\hat{g} - g^\star_{\mathcal{G}}\|_1 + \mathbb{E}\|g^\star_{\mathcal{G}} - g\|_1,$$

so the excess-risk bound additionally contains an approximation term of order

$$2\bar{\ell} \inf_{q\in\mathcal{G}} \mathbb{E}\|q - g\|_1.$$

□

A.4 Risk Estimation with Known Audit Propensities

The plug-in objective in Section 2.3 is the learning rule used by PCL. When π is known, the same observation model also admits standard IPW (Horvitz & Thompson, 1952) and doubly robust (DR) estimators (Bang & Robins, 2005) for a fixed predictor:

$$\hat{R}_{\mathrm{IPW}}(f) = \frac{1}{N}\sum_{i=1}^{N} \frac{o_i}{\pi(x_i,\tilde{y}_i)}\ell(f(x_i),y_i), \qquad \hat{R}_{\mathrm{DR}}(f) = \hat{R}_{\hat{g}}(f) + \frac{1}{N}\sum_{i=1}^{N} \frac{o_i}{\pi(x_i,\tilde{y}_i)}\left[\ell(f(x_i),y_i) - \sum_{y\in\mathcal{Y}} \hat{g}_y(x_i,\tilde{y}_i)\ell(f(x_i),y)\right]$$

Proposition 5 (Exact variance identities)**.** *Under EAR with known π, $\hat{R}_{\mathrm{IPW}}(f)$ is unbiased and $\hat{R}_{\mathrm{DR}}(f)$ is unbiased for every fixed $\hat{g}$. Let $\ell_Y := \ell(f(X),Y)$ and $\ell_{\hat{g}} := \sum_y \hat{g}_y(X,\tilde{Y})\ell(f(X),y)$. Then*

$$N\mathrm{Var}(\hat{R}_{\mathrm{IPW}}) = \mathrm{Var}(\ell_Y) + \mathbb{E}\left[\frac{1-\pi}{\pi}\ell_Y^2\right], \qquad N\mathrm{Var}(\hat{R}_{\mathrm{DR}}) = \mathrm{Var}(\ell_Y) + \mathbb{E}\left[\frac{1-\pi}{\pi}(\ell_Y - \ell_{\hat{g}})^2\right]. \quad (10)$$

With $\hat{g} = g$, the additional DR variance reduces to $\mathbb{E}[\frac{1-\pi}{\pi}\mathrm{Var}(\ell_Y \mid X,\tilde{Y})]$.

These estimators are not additional PCL training procedures; rather, they characterize clean-risk estimation when the audit mechanism is known. IPW uses only reviewed outcomes, whereas DR also exploits the full noisy pool through $\hat{g}$ and corrects its residual error on reviewed examples. Because this result concerns estimator bias and variance rather than downstream representation learning, we validate it separately in a controlled synthetic experiment in Appendix B.6, where the population clean risk is directly available.

Proof. Fix a predictor f and write

$$\ell_Y := \ell(f(X),Y), \qquad \ell_{\hat{g}} := \sum_{y\in\mathcal{Y}} \hat{g}_y(X,\tilde{Y})\ell(f(X),y).$$

Under EAR,

$$\mathbb{E}[O \mid X,\tilde{Y},Y] = \pi(X,\tilde{Y}).$$

Unbiasedness of IPW. The per-sample IPW contribution is

$$Z_{\mathrm{IPW}} = \frac{O}{\pi(X, \tilde{Y})} \ell_Y.$$

Conditioning on $(X, \tilde{Y}, Y)$ gives

$$\mathbb{E}[Z_{\mathrm{IPW}} \mid X, \tilde{Y}, Y] = \ell_Y.$$

Taking the outer expectation therefore yields

$$\mathbb{E}[\hat{R}_{\mathrm{IPW}}(f)] = \mathbb{E}[\ell_Y] = R(f).$$

Unbiasedness of DR. The corresponding DR contribution is

$$Z_{\mathrm{DR}} = \ell_{\hat{g}} + \frac{O}{\pi(X, \tilde{Y})}(\ell_Y - \ell_{\hat{g}}).$$

Again conditioning on $(X, \tilde{Y}, Y)$,

$$\mathbb{E}[Z_{\mathrm{DR}} \mid X, \tilde{Y}, Y] = \ell_{\hat{g}} + \frac{\mathbb{E}[O \mid X, \tilde{Y}, Y]}{\pi(X, \tilde{Y})}(\ell_Y - \ell_{\hat{g}}) = \ell_Y.$$

Thus $\mathbb{E}[\hat{R}_{\mathrm{DR}}(f)] = R(f)$ for every fixed $\hat{g}$.

Variance identities. By the law of total variance,

$$\begin{aligned}\mathrm{Var}(Z_{\mathrm{DR}}) &= \mathrm{Var}\left(\mathbb{E}[Z_{\mathrm{DR}} \mid X, \tilde{Y}, Y]\right) + \mathbb{E}\left[\mathrm{Var}(Z_{\mathrm{DR}} \mid X, \tilde{Y}, Y)\right] \\ &= \mathrm{Var}(\ell_Y) + \mathbb{E}\left[(\ell_Y - \ell_{\hat{g}})^2 \mathrm{Var}\left(\frac{O}{\pi} \,\middle|\, X, \tilde{Y}, Y\right)\right].\end{aligned}$$

Since O is Bernoulli with conditional mean π,

$$\mathrm{Var}\left(\frac{O}{\pi} \,\middle|\, X, \tilde{Y}, Y\right) = \frac{1-\pi}{\pi}.$$

Therefore

$$\mathrm{Var}(Z_{\mathrm{DR}}) = \mathrm{Var}(\ell_Y) + \mathbb{E}\left[\frac{1-\pi}{\pi}(\ell_Y - \ell_{\hat{g}})^2\right].$$

Because $\hat{R}_{\mathrm{DR}}$ averages N independent contributions,

$$N\mathrm{Var}(\hat{R}_{\mathrm{DR}}) = \mathrm{Var}(\ell_Y) + \mathbb{E}\left[\frac{1-\pi}{\pi}(\ell_Y - \ell_{\hat{g}})^2\right].$$

The IPW identity follows identically by setting the augmentation term $\ell_{\hat{g}}$ to zero:

$$N\mathrm{Var}(\hat{R}_{\mathrm{IPW}}) = \mathrm{Var}(\ell_Y) + \mathbb{E}\left[\frac{1-\pi}{\pi}\ell_Y^2\right].$$

When $\hat{g} = g$ and $\ell_g = \mathbb{E}[\ell_Y \mid X, \tilde{Y}]$ hold, and consequently

$$\mathbb{E}[(\ell_Y - \ell_g)^2 \mid X, \tilde{Y}] = \mathrm{Var}(\ell_Y \mid X, \tilde{Y}).$$

Substitution gives the stated variance expression.

Double robustness. Suppose first that π is correctly specified. The unbiasedness calculation above holds for any fixed $\hat{g}$, so consistency does not require a correct correction model.

Conversely, suppose the propensity used by the estimator is $\pi'(X, \tilde{Y})$, possibly different from the true π, but $\hat{g} = g$. Then

$$\begin{aligned}\mathbb{E}[Z_{\mathrm{DR}}] &= \mathbb{E}[\ell_g] + \mathbb{E}\left[\frac{O}{\pi'}(\ell_Y - \ell_g)\right] \\ &= R(f) + \mathbb{E}\left[\frac{\pi}{\pi'}\mathbb{E}[\ell_Y - \ell_g \mid X, \tilde{Y}]\right] = R(f),\end{aligned}$$

because $\mathbb{E}[\ell_Y - \ell_g \mid X, \tilde{Y}] = 0$. Thus the DR estimator remains consistent if either the propensity model is correct or the correction posterior is consistently estimated. □

A.5 Information Value of Paired Corrections

Consider a parametric model $\theta = (\alpha, \beta)$ with clean posterior $p_\alpha(y \mid x)$ and instance-dependent annotation mechanism $T_\beta(\tilde{y} \mid y, x)$.

Corollary 6 (Information in a correction pair)**.** *Under standard regularity conditions, conditional on $X = x$,*

$$I_{\mathrm{corr}}(\theta \mid x) = \begin{pmatrix} I_{\mathrm{clean}}(\alpha \mid x) & 0 \\ 0 & I_{\mathrm{ann}}(\beta \mid x) \end{pmatrix}, \qquad I_{\mathrm{fresh}}(\theta \mid x) = \begin{pmatrix} I_{\mathrm{clean}}(\alpha \mid x) & 0 \\ 0 & 0 \end{pmatrix}.$$

Thus a correction pair carries the same information about α as a fresh clean label, while additionally providing a supervised observation of the annotation mechanism β.

Proof. Condition on $X = x$. For a correction pair $(Y, \tilde{Y})$, the joint likelihood factorizes as

$$p_{\alpha,\beta}(y, \tilde{y} \mid x) = p_\alpha(y \mid x) T_\beta(\tilde{y} \mid y, x),$$

so the log-likelihood is

$$\ell_{\mathrm{corr}}(\theta) = \log p_\alpha(y \mid x) + \log T_\beta(\tilde{y} \mid y, x).$$

Write the corresponding score as

$$s_\theta = \begin{pmatrix} s_\alpha \\ s_\beta \end{pmatrix}, \qquad s_\alpha = \nabla_\alpha \log p_\alpha(Y \mid x), \qquad s_\beta = \nabla_\beta \log T_\beta(\tilde{Y} \mid Y, x).$$

Because the first term depends only on α and the second only on β, the cross-information term satisfies

$$\begin{aligned} \mathbb{E}[s_\alpha s_\beta^\top \mid X = x] &= \mathbb{E}_Y \left[s_\alpha \, \mathbb{E}[s_\beta^\top \mid Y, X = x] \right] \\ &= 0, \end{aligned}$$

where the last equality follows from the zero-mean score identity under the stated regularity conditions. Hence

$$I_{\mathrm{corr}}(\theta \mid x) = \begin{pmatrix} I_{\mathrm{clean}}(\alpha \mid x) & 0 \\ 0 & I_{\mathrm{ann}}(\beta \mid x) \end{pmatrix}.$$

For a fresh clean label, only (X, Y) is observed, with conditional log-likelihood $\ell_{\mathrm{fresh}}(\theta) = \log p_\alpha(Y \mid x)$, which does not depend on β. Therefore,

$$I_{\mathrm{fresh}}(\theta \mid x) = \begin{pmatrix} I_{\mathrm{clean}}(\alpha \mid x) & 0 \\ 0 & 0 \end{pmatrix}.$$

Finally, consider a differentiable path $\theta(t) = (\alpha(t), \beta(t))$ along which the noisy-label marginal

$$q_{\theta(t)}(\tilde{y} \mid x) = \sum_y p_{\alpha(t)}(y \mid x) T_{\beta(t)}(\tilde{y} \mid y, x)$$

is constant in t. Then

$$\frac{d}{dt} \log q_{\theta(t)}(\tilde{Y} \mid x) = 0,$$

so the noisy-only score has zero projection along the tangent direction $\dot{\theta}(t)$. Hence the noisy-only Fisher information has a null direction along such an observationally equivalent path, providing the local counterpart of the non-identifiability in Theorem 2. □

A.6 Selective Review under MNAR

When review depends on the unknown clean label, Eq. 5 relates the observed reviewed posterior $\tilde{g}$ to the population correction posterior g.

Proposition 7 (Selective corrections under MNAR)**.** *If the review propensity may depend on y, then:*

(i) *Non-identification. Let*

$$q(x, \tilde{y}) := \mathbb{P}(O = 1 \mid X = x, \tilde{Y} = \tilde{y}).$$

At any $(x, \tilde{y})$ with $0 < q(x, \tilde{y}) < 1$, the observed law is compatible with a continuum of population correction posteriors $g(x, \tilde{y})$, absent further restrictions on $\pi(x, \tilde{y}, y)$. If $q(x, \tilde{y}) = 0$, $g(x, \tilde{y})$ is unconstrained by reviewed outcomes.

(ii) *Known selection mechanism. If $\pi(x, \tilde{y}, y) > 0$ and the relative propensities*

$$\frac{\pi(x, \tilde{y}, y)}{\pi(x, \tilde{y}, y')}$$

are known, then

$$g_y(x, \tilde{y}) \propto \frac{\tilde{g}_y(x, \tilde{y})}{\pi(x, \tilde{y}, y)},$$

with normalization over y. If the propensities themselves are known, fitting g on reviewed examples with inverse-propensity weights $\frac{1}{\pi(x_i, \tilde{y}_i, y_i)}$ is consistent.

(iii) *Sensitivity bounds. In the binary case, suppose $\Lambda^{-1} \leq \frac{\pi(x,\tilde{y},1)}{\pi(x,\tilde{y},0)} \leq \Lambda$ for some $\Lambda \geq 1$. Then*

$$\frac{\tilde{g}_1}{\tilde{g}_1 + \Lambda(1 - \tilde{g}_1)} \leq g_1(x, \tilde{y}) \leq \frac{\Lambda \tilde{g}_1}{\Lambda \tilde{g}_1 + (1 - \tilde{g}_1)}. \tag{11}$$

Substituting these pointwise bounds into Eq. 2 yields corresponding bounds on $R(f)$ whose width vanishes as $\Lambda \to 1$.

Proof. We prove the three claims separately.

(i) Non-identification. Fix $(x, \tilde{y})$ and write

$$q := \mathbb{P}(O = 1 \mid X = x, \tilde{Y} = \tilde{y}), \qquad \tilde{g}_y := \mathbb{P}(Y = y \mid X = x, \tilde{Y} = \tilde{y}, O = 1).$$

Both quantities are determined by the observed law whenever $q > 0$.

Suppose first that $0 < q < 1$. Let h be any probability vector with strictly positive coordinates, and define

$$g_y := q\tilde{g}_y + (1 - q)h_y, \qquad \pi(x, \tilde{y}, y) := \frac{q\tilde{g}_y}{g_y}.$$

Since $g_y \geq q\tilde{g}_y$, we have $0 \leq \pi(x, \tilde{y}, y) \leq 1$. Moreover, $\pi(x, \tilde{y}, y)g_y = q\tilde{g}_y$, and hence

$$\sum_y \pi(x, \tilde{y}, y)g_y = q.$$

Therefore the constructed model reproduces the observed review probability. It also reproduces the reviewed posterior:

$$\mathbb{P}(Y = y \mid x, \tilde{y}, O = 1) = \frac{\pi(x, \tilde{y}, y)g_y}{\sum_{y'} \pi(x, \tilde{y}, y')g_{y'}} = \tilde{g}_y.$$

Thus every choice of h yields the same observed law of the review indicator and reviewed outcomes, while generally inducing a different population correction posterior

$$g = q\tilde{g} + (1 - q)h.$$

For a label space with at least two classes, varying h gives a continuum of observationally equivalent population posteriors. Hence g is not identified.

If $q = 0$, no clean outcome is observed at $(x, \tilde{y})$, so the observed law places no restriction on $g(x, \tilde{y})$.

(ii) Known selection mechanism. From Eq. 5, $\tilde{g}_y(x,\tilde{y}) \propto \pi(x,\tilde{y},y)g_y(x,\tilde{y})$, and therefore

$$g_y(x,\tilde{y}) = \frac{\tilde{g}_y(x,\tilde{y})/\pi(x,\tilde{y},y)}{\sum_{y'}\tilde{g}_{y'}(x,\tilde{y})/\pi(x,\tilde{y},y')}.$$

Any multiplicative factor common to all values of y cancels in the normalization, so only the relative propensities across y are required for this identification formula.

When the propensities themselves are known, the same correction admits an inverse-propensity-weighted formulation. For any integrable function $h(X,\tilde{Y},Y)$,

$$\begin{aligned}\mathbb{E}\left[\frac{O}{\pi(X,\tilde{Y},Y)}h(X,\tilde{Y},Y)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{O}{\pi(X,\tilde{Y},Y)}h(X,\tilde{Y},Y) \mid X,\tilde{Y},Y\right]\right] \\ &= \mathbb{E}[h(X,\tilde{Y},Y)],\end{aligned}$$

because

$$\mathbb{E}[O \mid X,\tilde{Y},Y] = \pi(X,\tilde{Y},Y).$$

Thus inverse-propensity weighting recovers population expectations from the selectively reviewed sample. In particular, weighting the conditional log-likelihood of Y given $(X,\tilde{Y})$ by $1/\pi(X,\tilde{Y},Y)$ yields a consistent objective for estimating g.

(iii) Sensitivity bounds. Consider the binary case and write $t := \tilde{g}_1(x,\tilde{y})$, $r := \frac{\pi(x,\tilde{y},1)}{\pi(x,\tilde{y},0)}$. From Eq. 5,

$$t = \frac{\pi_1 g_1}{\pi_1 g_1 + \pi_0(1-g_1)}.$$

For $t \in (0,1)$, solving for g_1 gives

$$g_1(x,\tilde{y}) = \frac{t}{t + r(1-t)}.$$

The cases $t \in \{0,1\}$ follow by continuity. For fixed $t \in (0,1)$, this expression is decreasing in r. Hence if $r \in [\Lambda^{-1},\Lambda]$, then

$$\frac{t}{t+\Lambda(1-t)} \le g_1(x,\tilde{y}) \le \frac{t}{t+\Lambda^{-1}(1-t)} = \frac{\Lambda t}{\Lambda t + (1-t)}.$$

Substituting $t = \tilde{g}_1(x,\tilde{y})$ gives Eq. 11. Finally, for binary labels the integrand in Eq. 2 is

$$g_1(x,\tilde{y})\ell(f(x),1) + (1-g_1(x,\tilde{y}))\ell(f(x),0),$$

which is affine in $g_1(x,\tilde{y})$. Hence its minimum and maximum over the sensitivity interval are attained at the interval endpoints, with the relevant endpoint determined by the sign of

$$\ell(f(x),1) - \ell(f(x),0).$$

Integrating these pointwise extrema gives corresponding lower and upper bounds on $R(f)$. As $\Lambda \to 1$, the two bounds on g_1 coincide with $\tilde{g}_1$, so the resulting risk interval collapses to a point. □

B Controlled Testbed and Additional Analyses

B.1 Testbed construction and design choices

We now give the full specification of the controlled testbed (Figure 3) used in Section 3.1. We draw

$$X \sim \mathrm{Unif}([-3,3]), \qquad \eta(x) = \sigma\big(3(x^2-1)\big),$$

and define the localized corruption functions as

$$\begin{aligned}e_1(x) &= 0.01 + 0.60b_{-1}(x) + 0.05b_{+1}(x),\\ e_0(x) &= 0.01 + 0.03b_{-1}(x) + 0.30b_{+1}(x),\end{aligned}$$

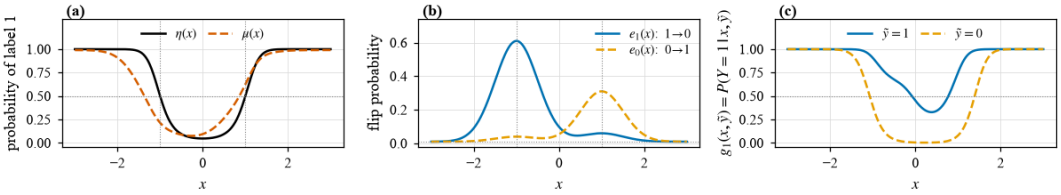


Figure 3: **Controlled testbed. (a)** Clean posterior $\eta(x) = \Pr(Y = 1 \mid x)$ and noisy marginal $\mu(x) = \Pr(\tilde{Y} = 1 \mid x)$. Vertical guides indicate the clean Bayes boundaries at $x = \pm 1$. **(b)** Instance-dependent flip probabilities. Their class-conditional averages are nearly identical despite strong local asymmetry. **(c)** Correction posterior $g_v(x) = \Pr(Y = 1 \mid X = x, \tilde{Y} = v)$, the quantities directly estimable from reviewed correction pairs under EAR.

where

$$b_c(x) = \exp\left(-\frac{(x-c)^2}{2(0.5)^2}\right).$$

The noisy-label posterior is

$$\mu(x) = \eta(x)(1 - e_1(x)) + (1 - \eta(x))e_0(x).$$

The construction is deliberately chosen to distinguish instance-dependent correction from class-conditional correction. Despite strong local asymmetry, the population class-conditional flip rates are nearly identical:

$$\Pr(\tilde{Y} = 0 \mid Y = 1) = 0.1141, \qquad \Pr(\tilde{Y} = 1 \mid Y = 0) = 0.1137.$$

A method that collapses the corruption mechanism to a single class-conditional transition therefore sees an almost symmetric average channel even though the true transition changes substantially around the two decision boundaries.

The clean Bayes classifier achieves accuracy 0.9162, whereas thresholding the noisy posterior, $\mathbf{1}\{\mu(x) > \frac{1}{2}\}$, achieves clean accuracy 0.8807. The resulting 3.55 percentage-point gap makes instance-dependent correction visible in downstream classification. The overall population corruption rate is 0.1140.

The additive 0.01 corruption floor ensures that both observed noisy-label branches retain support over the domain. This concerns support of $\tilde{Y}$, rather than review coverage in Assumption 1; the latter is varied explicitly in the coverage experiment below.

For $v \in \{0, 1\}$, define

$$g_v(x) = \Pr(Y = 1 \mid X = x, \tilde{Y} = v).$$

Bayes' rule gives

$$g_0(x) = \frac{\eta(x)e_1(x)}{1 - \mu(x)}, \qquad g_1(x) = \frac{\eta(x)(1 - e_1(x))}{\mu(x)}.$$

The identity

$$\eta(x) = (1 - \mu(x))g_0(x) + \mu(x)g_1(x)$$

provides a population consistency check.

B.2 Model classes and population error

We compare three fixed model classes. The quadratic model uses the joint feature representation

$$\phi(x, v) = (x, x^2, v, xv, x^2v)$$

with an intercept, giving six fitted coefficients. The spline and RBF estimators fit the two noisy-label branches separately. The cubic B-spline model uses degree three with eight knots and has 20 coefficients in total. The RBF estimator uses 300 Gaussian random Fourier features per branch with $\gamma = 1$ and has 602 coefficients including intercepts. The spline and RBF logistic regressions use $C = 10$. No hyperparameter is tuned across edit budgets or review mechanisms.

		Free fit: $cn_{\mathrm{e}}^{-\alpha} + \varepsilon_{\mathcal{G}}$		Fixed $\alpha = \frac{1}{2}$	Error at $n_{\mathrm{e}} = 12{,}800$
Model class $\mathcal{G}$	$d_{\mathcal{G}}$	$\hat{\alpha}$ [95% CI]	$\hat{\varepsilon}_{\mathcal{G}}$ [95% CI]	$\hat{c}$	Vector-L^1 error
Quadratic logistic	6	$0.570\,[0.485, 0.670]$	$0.0128\,[0.0087, 0.0163]$	0.9067	0.0204 ± 0.0008
Cubic spline	20	$0.505\,[0.470, 0.590]$	$0.0000\,[0.0000, 0.0050]$	1.1064	0.0098 ± 0.0015
RBF random features	602	$0.480\,[0.435, 0.585]$	$0.0000\,[0.0000, 0.0050]$	0.9353	0.0093 ± 0.0016

Table 3: Rate fits for correction-posterior estimation. Intervals are 95% percentile bootstrap intervals over complete seed-level curves. All three exponent intervals contain the reference value $\frac{1}{2}$. The fitted floor is separated from zero for the quadratic model but not for the spline or RBF families. The fitted floor is a descriptive extrapolation over the observed budget range rather than a direct measurement of asymptotic approximation error.

For $v \in \{0, 1\}$, define the scalar posterior

$$g_v(x) = \Pr(Y = 1 \mid X = x, \tilde{Y} = v),$$

and its corresponding class-probability vector

$$\mathbf{g}_v(x) = \big(1 - g_v(x),\, g_v(x)\big).$$

Define $\widehat{\mathbf{g}}_v(x)$ analogously from $\widehat{g}_v(x)$. The rate experiment reports the population vector-L^1 error

$$\mathcal{E}_{L^1}(\widehat{\mathbf{g}}) = \mathbb{E}_{X,\tilde{Y}}\left[\|\widehat{\mathbf{g}}_{\tilde{Y}}(X) - \mathbf{g}_{\tilde{Y}}(X)\|_1\right].$$

For binary labels,

$$\mathcal{E}_{L^1}(\widehat{\mathbf{g}}) = 2\mathbb{E}_{X,\tilde{Y}}\left[|\widehat{g}_{\tilde{Y}}(X) - g_{\tilde{Y}}(X)|\right].$$

Hence the scalar posterior MAE reported in the coverage and MNAR experiments is exactly one half of the vector-L^1 metric used for the rate experiment. All population errors are evaluated numerically using composite Simpson's rule rather than a finite clean test set.

B.3 Edit-budget sweep and rate fitting

For each

$$n_{\mathrm{e}} \in \{50, 100, 200, 400, 800, 1600, 3200, 6400, 12800\},$$

we directly draw n_{e} independent correction triples from the population and repeat the experiment over 20 seeds. The empirical mean error is fitted to

$$\overline{\mathcal{E}}_{L^1}(n_{\mathrm{e}}) \approx cn_{\mathrm{e}}^{-\alpha} + \varepsilon_{\mathcal{G}}.$$

The free exponent is searched over the grid $[0.05, 1.5]$ in increments of 0.005. We additionally fit a constrained model with $\alpha = \frac{1}{2}$. Confidence intervals are percentile intervals obtained from 2,000 bootstrap resamples of complete seed-level curves, thereby preserving the dependence of a seed across edit budgets.

The constrained $\alpha = \frac{1}{2}$ fits yield coefficients 0.9067, 1.1064, and 0.9353 for the quadratic, spline, and RBF families. We do not interpret these constants as implying that raw parameter count is a sharp empirical complexity measure: the theoretical finite-dimensional rate suppresses model- and distribution-dependent constants, while the spline and random-feature estimators are regularized and structurally constrained.

To complement the aggregate rate analysis, we inspect the representative budget $n_{\mathrm{e}} = 800$ shown in Figure 1(a,b). At this budget, the most visible mismatch occurs for the quadratic estimator in regions where the true correction posterior changes rapidly, whereas the spline and RBF estimators more closely track the population target. A separate finite-sample effect appears in the far tails of the $\tilde{Y} = 0$ branch. Under the population model above,

$$\Pr(|X| > 2, \tilde{Y} = 0) = \mathbb{E}[\mathbf{1}\{|X| > 2\}(1 - \mu(X))] \approx 0.0064.$$

Thus, the event

$$\{|X| > 2, \tilde{Y} = 0\}$$

accounts for only about 0.64% of the population. Among n_{e} reviewed observations, this corresponds to an expected count of approximately $0.0064\, n_{\mathrm{e}}$, or about five observations when $n_{\mathrm{e}} = 800$. Because this region has little population mass, its contribution to the population-weighted error is correspondingly small and decreases as n_{e} grows.

Region / Metric	Exploration Rate (ϵ)			
	0.00	0.01	0.02	0.05
Overall MAE ($\mathcal{E}_{\text{all}}$)	1.78 ± 0.47	1.33 ± 0.24	$\mathbf{1.30 \pm 0.21}$	1.55 ± 0.32
Inside $\mathcal{A}$ ($\mathcal{E}_{\text{in}}$)	$\mathbf{1.99 \pm 0.38}$	2.04 ± 0.41	2.24 ± 0.45	3.26 ± 0.79
Outside $\mathcal{A}$ ($\mathcal{E}_{\text{out}}$)	1.68 ± 0.69	0.98 ± 0.24	0.83 ± 0.21	$\mathbf{0.69 \pm 0.19}$

Table 4: Coverage experiment under a fixed expected 5% review budget. Entries are scalar class-1 posterior MAE ($\times 10^{-2}$), reported as mean $\pm$ one standard deviation over 20 seeds. Regional errors are conditional on the corresponding population region. Best performing overall and regional allocations are highlighted in bold.

B.4 Coverage under a fixed expected audit budget

Let

$$\mathcal{A} = (-1.5, -0.5) \cup (0.5, 1.5), \qquad \Pr(X \in \mathcal{A}) = \frac{1}{3},$$

denote neighborhoods of the two clean decision boundaries. We fix the target expected review probability at $\rho_\star = 0.05$ and define

$$\pi_\epsilon(x) = \begin{cases} a_\epsilon, & x \in \mathcal{A}, \\ \epsilon, & x \notin \mathcal{A}, \end{cases} \qquad a_\epsilon = 0.15 - 2\epsilon.$$

For

$$\epsilon \in \{0, 0.01, 0.02, 0.05\},$$

this gives

$$a_\epsilon \in \{0.15, 0.13, 0.11, 0.05\}, \qquad \mathbb{E}[\pi_\epsilon(X)] = 0.05.$$

Using pools of size $N = 20{,}000$, every policy therefore has the same expected audit count,

$$N\rho_\star = 1{,}000,$$

while reallocating that budget differently over the domain. The equality holds in expectation rather than as an enforced equality of realized audit counts.

For each seed, the same population pool, audit uniforms, and RBF feature map are shared across the four policies. The correction-posterior estimator is fitted using reviewed observations only; no propensity weighting is used in this experiment.

Define

$$h_{\widehat{g}}(x) = \sum_{v=0}^{1} \Pr(\tilde{Y} = v \mid X = x) \left| \widehat{g}^{(v)}(x) - g^{(v)}(x) \right|.$$

We report the overall population error

$$\mathcal{E}_{\text{all}} = \mathbb{E}[h_{\widehat{g}}(X)],$$

together with the regional errors

$$\mathcal{E}_{\text{in}} = \mathbb{E}[h_{\widehat{g}}(X) \mid X \in \mathcal{A}], \qquad \mathcal{E}_{\text{out}} = \mathbb{E}[h_{\widehat{g}}(X) \mid X \notin \mathcal{A}].$$

Table 4 quantifies the finite-sample allocation trade-off. As randomized exploration increases, the outside-region error decreases monotonically, while the inside-region error increases because fewer reviews remain concentrated near the decision boundaries. Among the evaluated policies, the overall error is lowest at the intermediate allocation $\epsilon = 0.02$.

Figure 4 illustrates the corresponding identification issue more directly. At $\epsilon = 0$, the RBF estimator can extrapolate smoothly outside $\mathcal{A}$, producing finite prediction error despite the absence of reviewed observations there. This should not be interpreted as identification without coverage: the correction data themselves impose no constraints on the population correction posterior outside the reviewed region. With positive exploration ($\epsilon = 0.05$), reviewed observations are supplied over the full population support, although this need not minimize finite-sample error under a fixed audit budget.

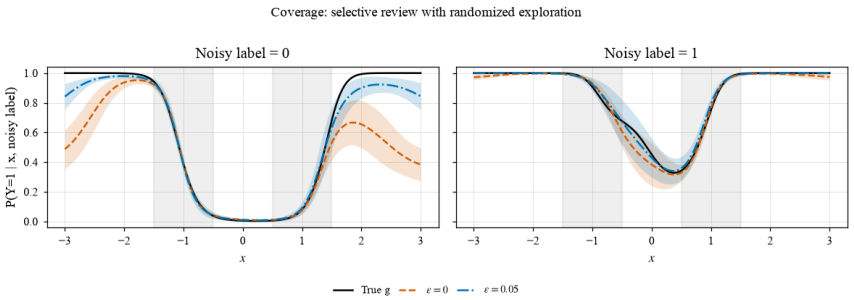


Figure 4: **Correction-posterior recovery with and without randomized exploration.** True and fitted correction posteriors under $\epsilon = 0$ and uniform $\epsilon = 0.05$ auditing. Shaded regions indicate the selected boundary neighborhoods $\mathcal{A}$. Without exploration, predictions outside $\mathcal{A}$ arise entirely from model extrapolation; positive exploration supplies reviewed observations over the full support. Curves show means with one-standard-deviation bands over 20 seeds.

B.5 MNAR REVIEW AND PROPENSITY WEIGHTING

We next violate EAR through error-triggered review. An observation is reviewed according to

$$\pi_m(Y, \tilde{Y}) = \begin{cases} 0.05, & Y = \tilde{Y}, \\ 0.05m, & Y \neq \tilde{Y}, \end{cases} \qquad m \in \{1, 2, 3, 5\}.$$

At $m = 1$, review is uniform and satisfies EAR. For $m > 1$, mislabeled examples are more likely to be reviewed, so the review decision depends on the unobserved clean label even after conditioning on $(X, \tilde{Y})$.

Consequently, unweighted fitting on the reviewed subset targets the selected correction posterior rather than the population correction posterior. For $v \in \{0, 1\}$, define

$$\widetilde{g}^{(v)}(x) = \Pr(Y = 1 \mid X = x, \tilde{Y} = v, O = 1).$$

Under the review mechanism above,

$$\widetilde{g}^{(0)}(x) = \frac{m g^{(0)}(x)}{m g^{(0)}(x) + 1 - g^{(0)}(x)}, \qquad \widetilde{g}^{(1)}(x) = \frac{g^{(1)}(x)}{g^{(1)}(x) + m(1 - g^{(1)}(x))}.$$

Thus error-triggered review shifts the two observed-label branches in opposite directions.

We compare unweighted fitting with propensity-weighted correction-posterior fitting using the known review probabilities,

$$w_i = \frac{1}{\pi_m(y_i, \tilde{y}_i)}.$$

At each m, the two methods use exactly the same reviewed observations and the same RBF feature representation. Because mislabeled examples are reviewed more frequently as m increases, the expected number of reviewed examples also increases, from approximately $1{,}000$ at $m = 1$ to $1{,}456$ at $m = 5$.

Table 5 quantifies the resulting selection effect. As error-triggered review becomes stronger, the error of the unweighted estimator increases substantially even though the expected number of reviewed examples grows. In contrast, propensity-weighted fitting remains close to the population correction posterior, indicating that the deterioration of the unweighted estimator is driven by the composition of the reviewed sample rather than by insufficient supervision. At $m = 1$, all inverse-propensity weights are constant; because the implemented logistic estimator is regularized and uses raw, unnormalized sample weights, this constant rescaling slightly changes the effective regularization strength. We therefore do not interpret the small difference between the unweighted and weighted fits at $m = 1$ as a selection-correction effect. For $m > 1$, however, the relative weights correct the clean-label-dependent selection tilt characterized by Proposition 7. When these relative review propensities are unknown, Proposition 7 implies non-identification, motivating sensitivity analysis rather than a generic point-correction procedure.

	m			
	1	2	3	5
Expected Reviews	1000	1114	1228	1456
MAE ($\times 10^{-2}$)				
Unweighted	$\mathbf{1.55 \pm 0.32}$	3.38 ± 0.52	5.35 ± 0.53	8.11 ± 0.55
Weighted (IPW)	1.74 ± 0.37	$\mathbf{1.37 \pm 0.27}$	$\mathbf{1.20 \pm 0.26}$	$\mathbf{1.12 \pm 0.20}$

Table 5: **Correction-posterior error under MNAR review.** Population class-1 posterior MAE, reported as mean $\pm$ one standard deviation over 20 seeds. Best-performing methods are highlighted in bold.

Audit frac. ρ	Estimator	Bias	95% CI	SD	RMSE
0.01	Plug-in	0.0210	[0.0194, 0.0227]	0.0118	0.0241
	IPW	-0.0001	[−0.0031, 0.0030]	0.0221	0.0221
	DR	0.0011	[−0.0013, 0.0034]	0.0172	**0.0172**
0.02	Plug-in	0.0119	[0.0107, 0.0131]	0.0085	0.0146
	IPW	0.0010	[−0.0011, 0.0031]	0.0151	0.0151
	DR	0.0003	[−0.0011, 0.0017]	0.0100	**0.0100**
0.05	Plug-in	0.0058	[0.0050, 0.0066]	0.0061	0.0084
	IPW	0.0016	[0.0002, 0.0031]	0.0103	0.0104
	DR	0.0007	[−0.0003, 0.0016]	0.0067	**0.0067**
0.10	Plug-in	0.0031	[0.0025, 0.0037]	0.0043	0.0052
	IPW	0.0001	[−0.0008, 0.0011]	0.0070	0.0070
	DR	0.0001	[−0.0006, 0.0007]	0.0048	**0.0048**
0.20	Plug-in	0.0014	[0.0010, 0.0019]	0.0032	**0.0035**
	IPW	0.0003	[−0.0004, 0.0010]	0.0049	0.0049
	DR	-0.0001	[−0.0006, 0.0004]	0.0036	0.0036

Table 6: **Clean-risk estimation under partial auditing.** Results are averaged over 200 independent pool-and-audit repetitions for each audit fraction ρ. Plug-in and doubly robust (DR) estimators use five-fold cross-fitted correction posteriors. Bias denotes the mean estimation error, with normal-approximation 95% confidence intervals; SD is the across-repetition standard deviation. Bold indicates the lowest RMSE for each audit fraction.

B.6 Clean-risk estimation under known propensities

Unlike Appendix B.5, where propensity weighting is used to correct selection bias in correction-posterior estimation, we now study estimation of the clean risk $R(f)$ of a fixed predictor. Proposition 5 characterizes inverse-propensity and doubly robust estimators under known review propensities. We examine their finite sample bias-variance behavior using the controlled testbed defined in Appendix B.1. This experiment evaluates risk estimation only; it does not train an additional downstream PCL predictor.

We fix

$$f(x) = \mathbf{1}\{|x| > 1.2\}, \qquad \ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}.$$

The threshold differs from the clean Bayes boundary at $|x| = 1$, making the probe deliberately imperfect. Its population clean risk, computed by numerical integration, is $R(f) = 0.1037$.

For each

$$\rho \in \{0.01, 0.02, 0.05, 0.10, 0.20\},$$

we generate 200 independent pools of size $N = 20{,}000$ and draw uniform EAR audits

$$O_i \sim \text{Bernoulli}(\rho).$$

The reported variation therefore includes both population-pool sampling and audit sampling.

To prevent an observation's own reviewed clean label from entering its nuisance prediction, we estimate the correction posterior by five-fold cross-fitting. For each fold, the RBF correction model is trained on reviewed observations in the remaining four folds and evaluated on the held-out fold. Writing $k(i)$ for the fold containing observation i, define

$$\widehat{g}_y^{\text{OOF}}(x_i, \tilde{y}_i) = \widehat{g}_y^{(-k(i))}(x_i, \tilde{y}_i), \qquad y \in \{0, 1\}.$$

Thus every observation receives an out-of-fold correction-posterior prediction from a model that excludes its own clean outcome.

Let

$$L_i = \ell(f(x_i), y_i), \quad \text{and} \quad \widehat{L}_{g,i} = \sum_{y \in \{0,1\}} \widehat{g}_y^{\mathrm{OOF}}(x_i, \tilde{y}_i)\ell(f(x_i), y).$$

We evaluate

$$\widehat{R}_{\text{Plug-in}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{L}_{g,i},$$

$$\widehat{R}_{\mathrm{IPW}} = \frac{1}{N}\sum_{i=1}^{N} \frac{O_i}{\rho} L_i,$$

and

$$\widehat{R}_{\mathrm{DR}} = \frac{1}{N}\sum_{i=1}^{N} \left[\widehat{L}_{g,i} + \frac{O_i}{\rho}\left(L_i - \widehat{L}_{g,i}\right)\right].$$

Conditional on the nuisance-training data outside observation i's fold and on $(x_i, y_i, \tilde{y}_i)$, the out-of-fold prediction does not depend on O_i. Hence under the known uniform review mechanism,

$$\mathbb{E}_{O_i}\left[\widehat{L}_{g,i} + \frac{O_i}{\rho}(L_i - \widehat{L}_{g,i})\right] = L_i,$$

and analogously

$$\mathbb{E}_{O_i}\left[\frac{O_i}{\rho} L_i\right] = L_i.$$

Cross-fitting therefore preserves the relevant unbiasedness mechanism while allowing the nuisance posterior to be estimated from the same overall pool. Because out-of-fold terms can share nuisance-training data, we interpret this experiment as a finite-sample validation of the bias-variance behavior in Proposition 5, rather than as a direct numerical verification of its fixed-$\widehat{g}$ variance identity.

Table 6 reports empirical bias, standard deviation, and RMSE over the 200 independent repetitions, while Figure 5 summarizes the corresponding bias-variance trade-off across audit fractions. At low audit rates, the plug-in estimator has relatively low variance but substantial bias from correction-posterior estimation, whereas IPW remains approximately unbiased but incurs a large variance penalty. DR combines the two sources of information: its empirical bias remains small while its RMSE is lower than that of IPW across all reported audit fractions. As the audit fraction increases, the plug-in bias shrinks and its RMSE approaches that of DR; at $\rho = 0.20$, the plug-in estimator attains slightly lower RMSE, illustrating that unbiased augmentation need not minimize finite-sample mean squared error in every regime.

All five DR bias confidence intervals contain zero. One finite Monte Carlo interval for IPW, at $\rho = 0.05$, does not contain zero despite the estimator's theoretical unbiasedness under the specified audit mechanism. We report the realized simulations without altering seeds or estimators to enforce nominal coverage.

C Real-Noise Benchmarks and Large-Label-Space Analysis

C.1 Real-noise benchmark setup

Datasets and backbones. We evaluate on five real-noise classification settings: CIFAR-10N Aggregate, CIFAR-10N Worst, CIFAR-100N, AlleNoise, and a paired-training adaptation of Clothing1M.

The three CIFAR-N settings contain 50,000 training and 10,000 clean test images. Their empirical training-label noise rates are 9.03%, 40.21%, and 40.20%, respectively. We use the released `aggre_label`, `worse_label`, and `noisy_label` annotations. All CIFAR-N experiments use a PreAct ResNet-18 trained from scratch. Training images are randomly cropped to 32×32 after four-pixel reflective padding and randomly horizontally flipped.

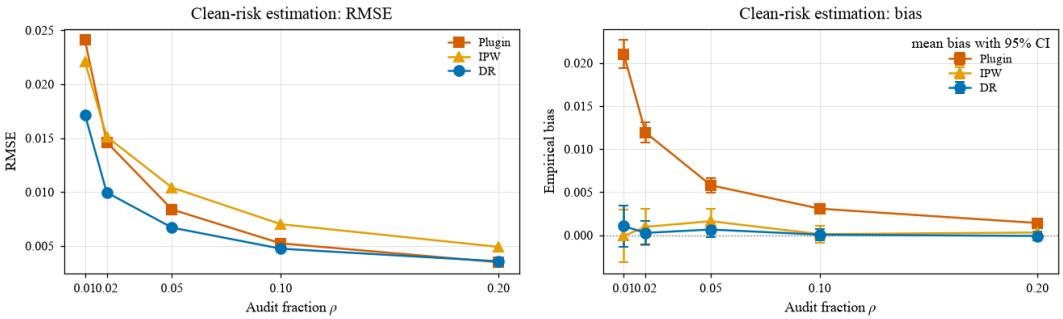


Figure 5: **Clean-risk estimation under known uniform audit propensities. Left:** RMSE of the plug-in, IPW, and doubly robust estimators. At low audit rates, IPW incurs a large variance penalty, whereas DR uses the estimated correction posterior to reduce RMSE. **Right:** Empirical bias with 95% confidence intervals for the mean. The plug-in bias decreases with the audit budget, while all reported DR intervals contain zero.

AlleNoise contains 502,310 e-commerce product titles from 5,692 leaf categories, with seller-assigned noisy labels and expert-corrected labels. We use a fixed, run-independent 90/10 split, yielding 452,079 training and 50,231 test examples. A lower-cased word-level vocabulary is constructed from the training split only; 1,024 hashed buckets are reserved for out-of-vocabulary tokens, and titles are truncated to 32 tokens. The classifier is a four-layer Transformer with hidden dimension 256, four attention heads, 1,024-dimensional feed-forward layers, dropout 0.1, learned positional embeddings, and masked-mean pooling.

The standard Clothing1M noisy-training pool does not contain a clean label for every noisy training image and therefore does not directly instantiate our paired-observation setting. We construct a paired training pool using only official clean-training images for which an original noisy annotation is also available. We do not synthesize missing clean or noisy labels. Evaluation is performed on the official clean test set, and the clean validation set is not used for training, early stopping, or model selection. All methods use an ImageNet-1K-V1-pretrained ResNet-50 with a newly initialized 14-class output layer. Images are resized to 256 pixels; training uses random 224×224 crops and horizontal flips, and evaluation uses center crops. This paired-training adaptation is distinct from the standard full-million Clothing1M protocol, and the corresponding results should not be directly compared with numbers reported under the standard protocol.

Table 7 summarizes the dataset-level model configurations and optimization settings used throughout the real-noise experiments.

Audit construction and information regimes. For a noisy training pool of size N and audit rate ρ, we reveal the clean labels of

$$n_{\mathrm{e}} = \mathrm{round}(\rho N)$$

examples. Audits are sampled without replacement and stratified by the observed noisy label. The resulting selection probability may depend on $\tilde{Y}$ but not on the unobserved clean label, satisfying the edits-at-random condition.

Every method evaluated on the same dataset, audit rate, and random seed receives exactly the same audited subset whenever the method uses trusted supervision. Noisy-only methods do not receive the audited labels. Different audit rates are sampled independently rather than being nested.

The audit-fraction study on the three CIFAR-N settings uses

$$\rho \in \{0.5, 1, 2, 5, 10, 20\}\%$$

and compares Noisy CE, Audit-only, Replace-corrected, Fresh-clean, PCL, and the All-clean skyline.

For Fresh-clean, the clean label budget is matched to PCL, but the clean labels are taken from instances disjoint from the audited correction set. Thus Fresh-clean observes the same number of clean labels but does not observe same-instance $(\tilde{y}, y)$ correction pairs.

The broader benchmark comparison uses

$$\rho \in \{1, 5, 10\}\%$$

and compares Noisy CE, GLC, L2RW, DivideMix, BLTM, PCL, PCL+DivideMix, and All-clean.

Configuration	CIFAR-10N	CIFAR-100N	AlleNoise	Clothing1M paired
Backbone	PreAct-RN18	PreAct-RN18	Transformer	RN50-ImageNet
Classes	10	100	5,692	14
Final training epochs	100	100	20	10
Training batch size	128	128	256	32
Evaluation batch size	512	512	1,024	128
Optimizer	SGD	SGD	AdamW	SGD
Initial learning rate	0.02	0.02	3×10^{-4}	0.002
Momentum	0.9	0.9	–	0.9
Weight decay	5×10^{-4}	5×10^{-4}	0.01	0.001
Nesterov momentum	yes	yes	–	yes
LR schedule	cosine	cosine	cosine	cosine
LR warm-up	0	0	1 epoch	0
Gradient clipping	none	none	1.0	none
Evaluation interval	5	5	2	1
PCL noisy-label warm-up	30	30	6	3
PCL channel epochs	60	60	12	6
PCL channel learning rate	0.02	0.02	3×10^{-4}	0.002
PCL channel weight decay	5×10^{-4}	5×10^{-4}	0.01	0.001
PCL channel head	per-label	embedding	embedding	per-label
PCL cross-fitting folds	2	2	2	2
PCL corrected targets	dense	dense	top-16	dense
GLC warm-up epochs	30	30	6	3
L2RW meta learning rate	0.02	0.02	3×10^{-4}	0.002
BLTM warm-up epochs	30	30	6	3
BLTM transition epochs	30	30	6	3
DivideMix warm-up epochs	10	30	4	1
DivideMix λ_u	25	150	25	0
DivideMix Beta parameter α	4	4	4	0.5
DivideMix temperature	0.5	0.5	0.5	0.5
DivideMix ramp-up epochs	16	30	4	4

Table 7: Dataset-level optimization settings for the real-noise experiments. "Final training epochs" excludes the additional warm-up and channel estimation stages of PCL, GLC, and BLTM. DivideMix includes its warm-up within its configured total training schedule.

PCL implementation. PCL first trains a noisy-label classifier to initialize the representation used by the correction model. The correction posterior

$$\widehat{g}_y(x, \tilde{y}) \approx P(Y = y \mid X = x, \tilde{Y} = \tilde{y})$$

is then fitted on audited correction pairs using two-fold cross-fitting.

For each fold, the correction model is trained on audited examples outside the fold and is used to construct corrected targets for examples inside the fold. Audited examples use a point mass on the observed clean label. A newly initialized classifier is subsequently trained on the complete training pool using

$$\widehat{R}_{\mathrm{PCL}}(f) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y} \widehat{g}_y(x_i, \tilde{y}_i) \ell(f(x_i), y).$$

For CIFAR-10N and paired Clothing1M, the correction posterior uses a separate linear output head for each conditioning noisy label. CIFAR-100N and AlleNoise instead use an embedding-conditioned head in order to share parameters across noisy labels and avoid the $O(K^2 d)$ parameter cost of a fully interacted design. The channel additionally contains a residual-to-identity bias reflecting the fact that most observed labels are correct.

For AlleNoise, storing the full $452{,}079 \times 5{,}692$ corrected-target matrix would require approximately 10 GB in single precision. We therefore retain the top-16 posterior entries for each training example and renormalize them before downstream training.

Baselines. Noisy CE trains on the complete noisy training pool. Audit-only trains only on the n_{e} audited clean examples. Replace-corrected replaces the noisy label of each audited example by its clean label and otherwise performs ordinary cross-entropy training. Fresh-clean performs an analogous clean-label replacement on an equally sized set of instances disjoint from the correction set, thereby matching the clean-label budget without observing same-instance edit pairs. All-clean trains using the clean label of every training example and serves only as a skyline.

GLC first trains a noisy-label classifier and estimates a class-conditional corruption matrix from the audited examples. It then applies forward loss correction, while audited examples use their clean labels directly.

L2RW treats audited examples as a clean meta-validation stream and differentiates through a virtual noisy-data update to obtain non-negative per-example training weights.

DivideMix trains two networks that alternately fit a two-component Gaussian mixture model to the peer network's per-example losses. In our trusted-data adaptation, audited examples are always assigned to the labeled partition, use their clean labels, and receive clean probability one. The labeled-partition size is calibrated using the disagreement rate observed on the audited subset. For AlleNoise, the two textual input views are identical and input MixUp is disabled because integer token IDs cannot be convexly interpolated; label refinement and the unlabeled consistency objective are retained.

BLTM estimates an instance-dependent forward transition

$$T_{yj}(x) = P(\tilde{Y} = j \mid Y = y, X = x).$$

The transition model is initialized from a Laplace-smoothed class-conditional confusion matrix estimated from audited pairs. The noisy-label-pretrained backbone is frozen while the transition model is fitted. No distilled pseudo-labeled examples are used. PCL and BLTM use the same channel-head family and the same capacity rule. For AlleNoise, BLTM approximates the forward mixture using the top-16 candidate clean labels rather than summing over all 5,692 classes.

PCL+DivideMix uses the same cross-fitted estimated correction posterior as PCL. It replaces DivideMix's one-hot noisy-label refinement prior by $\widehat{g}(x, \tilde{y})$ while retaining the remaining two-network DivideMix procedure.

Evaluation and compute. Every configuration is run with seeds $\{0, 1, 2\}$. We report final-epoch clean-test top-1 accuracy as the mean and one sample standard deviation over the three seeds. Test accuracy is evaluated periodically for diagnostics, but best-test checkpoint selection is never used. AlleNoise top-5 accuracy is additionally logged for diagnostics.

Experiments run on a single compute node with eight NVIDIA A100 GPUs and 96 allocated CPU cores. Eight independent workers are launched, each assigned one GPU and a disjoint, cost-balanced subset of the experiment grid. Individual models do not use distributed data parallelism. Standard CUDA phases use automatic mixed precision; L2RW disables mixed precision because it requires double backward through a virtual update, and the custom DivideMix loop runs in full precision.

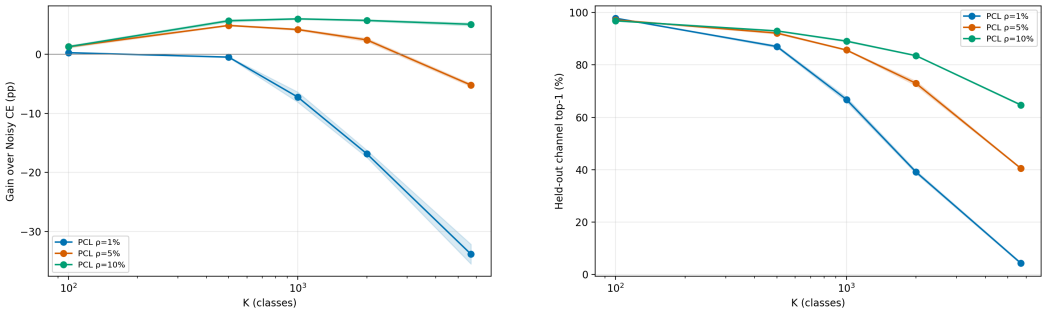
C.2 AlleNoise cardinality stress test

AlleNoise provides a stress test for correction-posterior estimation in an extreme label space. The full dataset contains $5{,}692$ classes, so even a substantial audit budget yields relatively sparse paired supervision for each observed label. As shown in Table 1, vanilla PCL performs poorly at very small audit fractions but improves sharply as additional correction pairs become available.

This behavior suggests a finite-sample estimation problem rather than a limitation caused simply by the parameterization of the correction model. Our AlleNoise implementation uses an embedding-conditioned correction head rather than a fully interacted $K \times K$ categorical transition matrix, allowing parameters to be shared across labels. Nevertheless, as the label space grows, each observed noisy label is supported by fewer audited corrections, making the instance-dependent correction posterior

$$g(y \mid x, \tilde{y})$$

increasingly difficult to estimate.



(a) Gain over Noisy CE **(b)** Correction-posterior accuracy

Figure 6: **Effect of label-space cardinality on paired correction.** PCL is evaluated on nested AlleNoise subsets with $K \in \{100, 500, 1000, 2000, 5692\}$ while keeping the architecture and training procedure fixed. **(a)** Gain in clean-test accuracy over Noisy CE. At low audit rates, the benefit of PCL deteriorates as the label space grows, whereas larger audit budgets substantially mitigate the effect. **(b)** Held-out top-1 accuracy of the estimated correction posterior $\widehat{g}(y \mid x, \tilde{y})$. The degradation in downstream performance closely tracks the difficulty of correction-posterior estimation. Shaded regions denote one standard deviation over three seeds.

To examine this effect directly, we construct nested AlleNoise subsets with

$$K \in \{100, 500, 1000, 2000, 5692\}.$$

Classes are ranked by frequency in the fixed training split, and each smaller benchmark contains the K most frequent classes. The subsets are therefore nested and deterministic across random seeds. The test split is restricted to the same label set, labels are remapped to contiguous indices, and no examples are synthesized.

For every value of K, we train Noisy CE and All-clean once per random seed and evaluate PCL at

$$\rho \in \{1, 5, 10\}\%.$$

The model architecture, optimization procedure, cross-fitting scheme, and top-16 posterior approximation are kept identical to the full AlleNoise experiment, except for the output-label dimension. No additional hyperparameter tuning is performed for any value of K.

Cardinality-dependent degradation. Figure 6 reports two quantities that directly characterize the failure mode: the downstream gain of PCL over Noisy CE and the held-out top-1 accuracy of the estimated correction posterior.

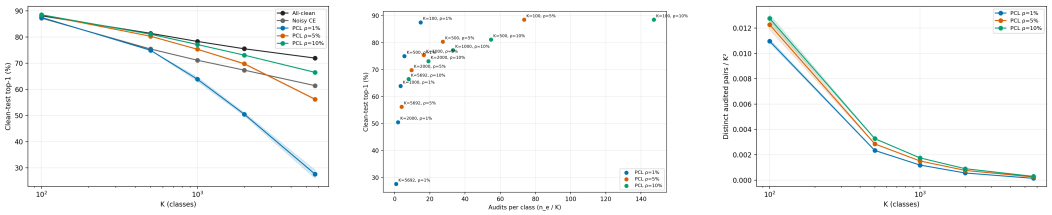
At low audit rates, PCL performs competitively when K is small but deteriorates as the number of classes grows. Larger audit budgets substantially mitigate this degradation: at $\rho = 10\%$, PCL remains above Noisy CE throughout the cardinality sweep. The correction-posterior diagnostic exhibits the same pattern. At $\rho = 1\%$, held-out top-1 accuracy of

$$\widehat{g}(y \mid x, \tilde{y})$$

falls sharply as K increases, reaching only about 4% on the full $5{,}692$-class problem. Increasing the audit fraction to 10% raises this accuracy to roughly 65%, alongside a corresponding recovery in downstream PCL performance.

These results support a finite-sample explanation for the low-audit AlleNoise regime. Positive audit coverage may ensure identification, but it does not guarantee accurate estimation of a high-cardinality correction posterior from a small number of correction pairs. As the label space grows, paired supervision becomes increasingly sparse relative to the estimation problem, and downstream performance deteriorates when the correction-posterior estimator becomes inaccurate. We do not interpret this experiment as establishing a K^2 sample-complexity law, since the correction structure need not be dense and our embedding-conditioned model shares information across labels.

Additional diagnostics. Figure 7 provides three complementary views of the same finite-sample effect.



(a) Absolute accuracy **(b)** Audit density **(c)** Correction-pair coverage

Figure 7: **Additional diagnostics for the AlleNoise cardinality stress test. (a)** Absolute clean-test accuracy of PCL, Noisy CE, and the All-clean skyline as the label space grows. The underlying classification problem becomes harder with increasing K, while PCL exhibits additional audit-dependent degradation at low audit rates. **(b)** PCL accuracy versus the mean number of audited examples per class, n_{e}/K. Effective paired supervision becomes increasingly sparse in the high-cardinality, low-audit regime. **(c)** Fraction of possible $(\tilde{y}, y)$ class pairs observed at least once in the audited subset. Pair coverage decreases rapidly with K; this quantity is used only as a descriptive sparsity diagnostic and does not imply a K^2 sample-complexity requirement. Shaded regions denote one standard deviation over three seeds where applicable.

First, absolute accuracy helps distinguish correction-posterior estimation difficulty from the increasing difficulty of the underlying classification problem. As shown in Figure 7(a), both Noisy CE and the All-clean skyline decline as K grows. PCL exhibits an additional audit-dependent degradation: at $\rho = 1\%$, its performance increasingly falls below Noisy CE as the label space grows, whereas at $\rho = 10\%$ it remains above Noisy CE throughout the sweep. At the largest audit rate, PCL also more closely tracks the All-clean trajectory.

Second, a fixed audit fraction need not correspond to a comparable correction-estimation problem across different values of K. Figure 7(b) relates downstream performance to the effective audit density $\frac{n_{\mathrm{e}}}{K}$, the mean number of audited examples per class. For the full $K = 5{,}692$ label space, $\rho = 1\%$ yields fewer than one audited example per class on average. At $\rho = 5\%$ and 10%, this increases to roughly four and eight audited examples per class, respectively. The same audit fractions provide substantially denser paired supervision on smaller label spaces.

We emphasize that $\frac{n_{\mathrm{e}}}{K}$ is only a coarse diagnostic. AlleNoise is class-imbalanced, and correction-posterior estimation depends on the joint distribution of $(X, \tilde{Y}, Y)$ rather than on the average number of audited examples per class alone. Nevertheless, the statistic illustrates how rapidly effective paired supervision becomes sparse in the extreme-cardinality regime.

Third, we measure the fraction of all possible class pairs $(\tilde{y}, y)$ that are observed at least once in the audited subset:

$$C_{\mathrm{pair}} = \frac{|\{(\tilde{y}_i, y_i) : O_i = 1\}|}{K^2}.$$

Figure 7(c) shows that this quantity decreases rapidly as the label space grows, even when the audit fraction is increased. The decline in C_{pair} is consistent with increasingly sparse direct supervision of the correction structure.

We do *not* interpret C_{pair} as implying that PCL requires observing a constant fraction of all K^2 class pairs, nor as evidence for a K^2 sample-complexity law. The true correction support may be highly sparse, and the embedding-conditioned correction model shares statistical strength across labels rather than estimating an independent parameter for every class pair. Accordingly, C_{pair} is used only as a descriptive sparsity diagnostic.

Taken together, the cardinality sweep and accompanying diagnostics indicate that the low-audit failure on full AlleNoise is a finite-sample limitation of correction-posterior estimation in an extreme-cardinality label space. As K grows, the available paired supervision becomes increasingly sparse, the held-out accuracy of $\widehat{g}(y \mid x, \tilde{y})$ deteriorates, and the downstream advantage of PCL over Noisy CE disappears. Increasing the audit budget reverses these trends without changing the PCL architecture or optimization procedure.



Figure 8: Review interface used in the prospective study (model-guided condition; the interface is in Japanese). The top banner shows the recommended action together with the three-class model probabilities. Below are the editable question, the fixed correct answer, the four editable distractors, and, on the right, the explanation and the collapsible reference sentences. The baseline condition uses the same interface without the banner.

D Additional Details for Real-World Human-in-the-Loop Curation

This appendix provides additional details for the real-world curation experiment in Section 3.3, including the historical prediction protocol, deployment model selection, complete user actions, editing behavior, participant-level results, and order-stratified analyses.

D.1 System and Pipeline Implementation

The curation system consists of an offline pre-processing pipeline, which converts a source corpus into a pool of MCQ candidates with supporting evidence, and a web-based review interface. The pre-processing pipeline and the historical review records (Appendix D.2) are shared with a concurrent study on MCQ dataset construction that is currently under review; we therefore summarize only the components needed to interpret the present experiment.

Source corpus and pre-processing. All candidates are generated from the Japanese Wikipedia article on the Sengoku period (Wikipedia, 2026), split into chunks at sentence boundaries. The pipeline combines standard components of automatic question generation (Kurdi et al., 2020): (i) *answer-candidate extraction*, in which an LLM extracts history-relevant entities (e.g., persons, places, years and eras, events) and only candidates that appear verbatim in the source text are retained, following the approach used in answer-aware question generation (Du et al., 2017); (ii) *over-generation and filtering* (Heilman & Smith, 2010), in which several questions are drafted for each entity and its source sentences, and LLM judges (Zheng et al., 2023) discard questions that are inconsistent with the source or duplicate existing ones; (iii) *distractor construction*, which selects same-category entities from the extracted candidate pool (Liang et al., 2018), complements them with LLM-generated distractors, and validates the resulting option set through an LLM critique step (Madaan et al., 2023; Yao et al., 2025); (iv) *evidence attachment*, which retrieves supporting sentences from the corpus (Lewis et al., 2020) and generates a short explanation that grounds the answer in them. All LLM calls use a model from the Sarashina series, which is trained for the Japanese language (SB Intuitions, 2026). Each resulting candidate Z consists of a question, one correct answer, and four distractors, accompanied by its reference sentences and explanation. The pipeline produces a pool of 559 Sengoku MCQ candidates (Appendix D.2).

Review interface. We implemented a new review interface for the prospective study (Figure 8). Each candidate displays its question, correct answer, and four distractors. The question and distractors are editable, whereas the correct answer is fixed. The explanation and reference sentences are shown alongside the item for fact-checking. Annotators finalize each item with REJECT or ACCEPT; an acceptance after any modification is recorded as edit-and-accept. In the model-guided condition, a banner additionally shows the recommended action and the precomputed ACCEPT/EDIT/REJECT probabilities; the baseline condition hides the banner and is otherwise identical. Candidates are presented in batches of five. The interface logs the original and final state of every presented item together with its final decision, which form the analysis records in Appendices D.3–D.5.

Annotation guidelines. Both the historical records and the prospective study follow the same annotation guidelines. Each MCQ must have exactly one correct answer and four distractors that are definitively incorrect; in particular, other historically valid answers must not appear among the distractors. Annotators briefly verify each item against the displayed evidence, the source article, or a web search. The remaining rules enforce option consistency, e.g., the same calendar system for years, the same suffix type (such as castle or temple names), the same personal-name format, the same category as the question asks for, and a uniform level of abstraction across options. Annotators edit an item when it can be repaired and reject it when the question is ambiguous or cannot be fixed within a reasonable time.

D.2 HISTORICAL CURATION DATA AND EVALUATION PROTOCOL

Data construction. The historical records were produced by ten experienced professional annotators in an earlier annotation session. The historical annotation files contain 735 question presentations. We exclude 51 pending presentations with no final curation decision, leaving 684 completed judgments over 156 question, answer families and 167 exact MCQ versions. The three-way target distribution is 109 ACCEPT, 485 EDIT, and 90 REJECT judgments.

For auditing purposes, the original edit records can also be decomposed into 109 direct accepts, 346 distractor-only edits, 139 stem edits, and 90 rejections. A stem edit includes both stem-only changes and changes affecting both the stem and distractors.

A question *family* is defined by the normalized original question and answer. A *version* additionally includes the original distractor set. Unicode NFKC normalization is applied, leading and trailing whitespace are removed, and internal whitespace is collapsed.

Each model receives only

$$(\text{original question},\ \text{answer},\ \text{original distractors}).$$

The final edited question, edit subtype, participant identifier, item identifier, event history, reference material, and final human decision are excluded from the model input.

Family-disjoint evaluation. We use five-fold stratified group cross-validation with question family as the grouping variable and split seed 42. All judgments belonging to the same question family therefore remain in the same outer fold. Within each outer training fold, hyperparameters are selected using a family-disjoint inner validation split. Every judgment receives an out-of-fold prediction from a model that has never observed its family during training.

During training, each record receives inverse-family-size weight to prevent families with many repeated judgments from dominating the objective. Neural encoders are trained with seeds 13, 21, and 42.

Models and selection rationale. We compare models spanning lexical, discriminative pretrained, and generative in-context regimes. The historical curation records are non-public, so all model training and evaluation is performed locally rather than by sending candidate content or human curation labels to third-party inference APIs. The original annotation system used a Sarashina-based model (SB Intuitions, 2026), whereas for the reproducible generative baseline reported here we use the open-weight 9B-parameter `Qwen/Qwen3.5-9B` (Qwen Team, 2026), which can be evaluated locally under the same data-governance constraint.

Table 8: Five-fold family-disjoint out-of-fold results on the historical annotation data. Macro-F1 and Reject AP are computed over all 684 judgment-level predictions. P@10% and Cov.@10% are computed over 167 MCQ versions ranked by the version-level mean $p(\text{REJECT})$, using a 10% review budget. Results for BERT, DeBERTa V2, and BERT+TF-IDF are means $\pm$ sample standard deviations over three encoder seeds $\{13, 21, 42\}$; single-run methods are reported without dispersion. Bold values indicate the best result among evaluated models. The empirical content-only oracle uses all observed labels and is included only as a descriptive reference; it is neither an OOF result nor a trainable-model result.

	Model							
Metric	Constant	TF-IDF	Japanese BERT	Japanese DeBERTa V2	Qwen-ZeroShot	Qwen-History	BERT+TF-IDF	Empirical Oracle
Macro-F1	0.2766	0.3589	0.3974±0.0129	0.3013±0.0221	0.2086	0.3193	**0.4003±0.0127**	0.6734
Reject AP	0.1257	0.2170	0.2133±0.0016	0.1819±0.0236	0.1894	0.1721	**0.2231±0.0009**	–
P@10%	–	0.3794	**0.3904±0.0317**	0.2728±0.0166	0.2326	0.2245	0.3704±0.0129	–
Cov.@10%	–	0.2662	**0.2739±0.0222**	0.1914±0.0116	0.1632	0.1575	0.2599±0.0090	–

Specifically, we compare a family-weighted class-prior predictor, character 2-5-gram TF-IDF with multinomial logistic regression, Japanese BERT (`tohoku-nlp/bert-base-japanese-v3`) (Devlin et al., 2019; Tohoku NLP Group, 2023), Japanese DeBERTa V2 (`izumi-lab/deberta-v2-base-japanese`) (He et al., 2021; Suzuki et al., 2023), and Qwen3.5-9B. Qwen is evaluated both zero-shot and with deterministic demonstrations drawn from the corresponding outer training fold. We do not fine-tune the LLM: with only 684 historical judgments, our goal is to test whether historical curation records themselves provide transferable predictive signal without introducing an additional supervised adaptation stage whose cost and variance would be difficult to separate from the value of the history itself. No LLM parameters are updated in either Qwen condition. We additionally report a fixed equal-weight BERT+TF-IDF ensemble.

This action-prediction experiment is intentionally narrower than direct estimation of the full structured correction posterior. Fine-tuning an open-weight generative model to predict the reviewed structured outcome, and thereby approximate a structured $g(Y \mid X, \tilde{Y})$, is a natural extension beyond the present action-prediction formulation.

Version-level screening metrics. For each version v, predictions from repeated historical judgments are averaged to obtain a single rejection score. Let

$$r_v = \frac{\#\{\text{REJECT JUDGMENTS ON } v\}}{\#\{\text{all judgments on } v\}}$$

denote the empirical reject mass of version v. At a review budget selecting the top k versions, we compute

$$\text{Precision@}k = \frac{1}{k} \sum_{v \in \text{Top}_k} r_v, \tag{12}$$

$$\text{Coverage@}k = \frac{\sum_{v \in \text{Top}_k} r_v}{\sum_v r_v}, \tag{13}$$

$$\text{Enrichment@}k = \frac{\text{Precision@}k}{V^{-1} \sum_v r_v}. \tag{14}$$

Here $V = 167$. At a 10% budget, $k = \lceil 0.1V \rceil = 17$.

Table 8 showcases these metric values on historical family-disjoint out-of-fold prediction.

To quantify the limitation caused by conflicting human decisions for identical inputs, we partition the 684 historical judgments into 179 groups according to the exact model input string, preserving distractor order. A deterministic content-only predictor must assign the same class to every judgment in the same group. We then select one label from $\{\text{ACCEPT}, \text{EDIT}, \text{REJECT}\}$ for each group so as to maximize Macro-F1 over all 684 judgments. We solve this joint discrete optimization exactly using mixed-integer linear programming, obtaining a maximum Macro-F1 of 0.6734. Because the optimization uses all observed labels, this value is an empirical oracle on the current dataset, rather than out-of-fold performance, the result of a trainable model, or a population-level upper bound.

Uncertainty is estimated using 500 paired family-cluster bootstrap samples. Each bootstrap draw resamples complete question families together with all their judgments and versions. The intervals condition on the already fitted models.

Deployment model selection. Although the fixed BERT+TF-IDF ensemble achieves slightly higher macro-F1 and Reject AP in Table 8, Japanese BERT achieves the best P@10% and Cov.@10%. Because the deployment workflow uses the model primarily to prioritize candidates under a limited human-review budget, we therefore use Japanese BERT as the base deployment model.

The deployed model is trained from the same 684 historical judgments. The original Sengoku MCQ pool contains 559 questions. Of the 156 families in the historical training data, 150 matched families in the 559-question Sengoku pool. We excluded the corresponding 150 pool questions, leaving 409 family-disjoint questions for deployment inference. The remaining six training families did not occur in the pool but were identified as Sengoku through their session metadata.

We compare a three-class head with a four-class head that separately predicts direct acceptance, distractor editing, stem editing, and rejection. For task-level comparison, the two edit probabilities of the four-class model are collapsed as

$$p(\textsc{Edit}) = p(\textsc{DistractorEdit}) + p(\textsc{StemEdit}).$$

Both variants use the same Japanese BERT checkpoint, nested family-disjoint folds, hyperparameter grid, and random seeds. The three-class model reaches 0.3914 ± 0.0059 pooled three-way OOF macro-F1, compared with 0.3768 ± 0.0157 for the collapsed four-class model, and is therefore used for deployment.

The selected configuration uses learning rate 2×10^{-5}, eight epochs, family weighting without additional class-frequency weighting, weight decay 0.01, maximum sequence length 256, batch size 16, and 10% linear warmup. The final model is refit on all 684 historical judgments with seed 42 and exports uncalibrated ACCEPT/EDIT/REJECT softmax scores for the 409 unseen candidates. The scores are used as ranking and recommendation signals rather than as automated final decisions.

D.3 Prospective Study Protocol

Participants. Eight professional annotators participate in the deployed-system evaluation. All eight are employed in dataset annotation work, with a median of 17.5 months of annotation experience (range 6–21), and all had also contributed to the historical records (Appendix D.2). Six participants reported their age (mean 42.7, SD 12.2, range 27–60) and gender (five male, one female). Seven of the eight reported no professional experience in machine-learning or LLM development. The study followed a process approved by our institution's ethics review board.

Design and procedure. All sessions take place on a single day. Before the main task, participants review the annotation guidelines (Appendix D.1). Each participant then completes two annotation blocks of approximately 30 minutes each in a counterbalanced crossover design: P01∼P04 experience the model-guided condition followed by the baseline, whereas P05∼P08 experience the baseline followed by the model-guided condition. Both blocks draw from the 409 unseen candidates (Appendix D.2). Each block ends at the time limit, and items presented but not decided by then are recorded as pending.

The model-guided condition presents five-question chunks that alternate between candidates predicted as ACCEPT and as EDIT by the deployed three-class model, and displays the corresponding ACCEPT or EDIT recommendation. The baseline samples questions uniformly from the candidate pool and displays no recommendation. In both conditions, annotators retain complete control over the final action and may:

- accept the question without modification;
- edit and accept the question;
- reject the question; or
- leave the question pending.

Table 9: Number of distractors replaced among questions receiving an edit.

Condition	n	0	1	2	3	4	Mean
ACCEPT recommendation	33	1	9	8	12	3	2.21
EDIT recommendation	63	7	12	19	8	17	2.25
No recommendation	132	2	70	29	14	17	1.80

The study contains 510 candidate presentations: 235 under the model-guided condition and 275 under the baseline. Among the 235 model-guided presentations, 85 carry an ACCEPT recommendation and 110 carry an EDIT recommendation. For the first five-question chunk of each participant (40 presentations in total), the recommendation-type label was not retained in the analysis record. These 40 observations are included in overall model-guided versus baseline comparisons, but excluded from recommendation-specific analyses.

Final-quality verification. We manually inspect the final state of each presented question and determine whether its final answer agrees with the reference evidence. One item is indeterminate and is excluded where answer correctness is required.

We define:

$$\text{Accepted-set accuracy} = \frac{\#\{\text{correct accepted final items}\}}{\#\{\text{all accepted final items}\}},$$
$$\text{Missed-error rate} = \frac{\#\{\text{incorrect unedited acceptances}\}}{\#\{\text{unedited acceptances}\}},$$
$$\text{Failed-edit rate} = \frac{\#\{\text{incorrect edited acceptances}\}}{\#\{\text{edited acceptances}\}}.$$

The quality verification targets final-answer correctness. Other dimensions of MCQ quality, such as distractor plausibility or wording clarity, are not included in this particular audit.

D.4 EDITING DEPTH

For edited questions, we characterize intervention depth by the number of distractors replaced. A value of zero denotes a stem-only edit.

Table 9 shows that edits made under model recommendations tend to involve more distractor replacements than those in the no-recommendation condition. Single-distractor replacements account for 70/132 (53.0%) baseline edits, compared with 9/33 (27.3%) under an ACCEPT recommendation and 12/63 (19.0%) under an EDIT recommendation. Conversely, replacing at least three distractors occurs in 45.5%, 39.7%, and 23.5% of edited items under ACCEPT, EDIT, and no recommendation, respectively.

D.5 PARTICIPANT-LEVEL AND ORDER-STRATIFIED RESULTS

Table 10 reports participant-level accepted-set accuracy and processing rate. All eight participants achieve 100% accepted-set answer accuracy under the model-guided workflow, whereas baseline accepted-set accuracy varies across participants.

Direct acceptance is higher under the model-guided workflow for seven of eight participants. The participant-level mean is 32.9% under model guidance and 24.9% under baseline presentation (paired Wilcoxon $p = 0.063$).

Mean processing rate is 0.79 items/minute under model guidance and 1.05 under the baseline. This participant-level difference is not significant (paired Wilcoxon $p = 0.38$). The baseline rate is also substantially more variable across participants (SD 0.66 vs. 0.18).

Model guidance reallocates human effort. The interaction logs reveal how the learned signal changes annotation behavior. When the system recommends ACCEPT, annotators accept a candidate without modification in 56.5% of cases, compared with 29.8% under the no-recommendation baseline.

Table 10: Participant-level results. "A→B" denotes model guidance followed by baseline; "B→A" denotes the reverse order.

	Presented		Accepted-set acc.		Items/min	
Participant	A	B	A	B	A	B
P01 (A→B)	35	50	100	88.9	1.06	1.75
P02 (A→B)	30	40	100	97.0	0.64	0.77
P03 (A→B)	20	35	100	100	0.60	1.30
P04 (A→B)	50	65	100	94.9	1.01	2.21
P05 (B→A)	30	30	100	100	0.93	0.61
P06 (B→A)	25	25	100	95.8	0.71	0.76
P07 (B→A)	20	15	100	100	0.61	0.51
P08 (B→A)	25	15	100	93.3	0.76	0.46

Table 11: Descriptive baseline behavior stratified by block order. Each group contains four participants.

Metric	Baseline first	Baseline after guidance
Editing rate	67.1%	39.5%
Rejection rate	4.7%	22.6%
Missed errors	1/16 (6.2%)	8/66 (12.1%)
Accepted-set accuracy	95.9%	92.7%
Throughput (items/min)	0.58	1.51

When the system recommends EDIT, direct acceptance falls to 13.6%; instead, 54.5% of candidates are edited and accepted, while the remaining cases are primarily rejected or deferred.

Importantly, this behavioral shift is not explained by simply inducing more editing. The overall edit rate is nearly unchanged between the model-guided and baseline workflows (49.8% vs. 48.0%). Rather, the model changes *where* human effort is spent. Among edited questions, replacing only one distractor accounts for 53.0% of baseline edits, compared with 27.3% under ACCEPT recommendations and 19.0% under EDIT recommendations. Conversely, edits affecting at least three distractors are substantially more frequent under model guidance.

Order-stratified baseline behavior. Because the study is counterbalanced, we can also inspect baseline behavior according to whether the baseline is experienced before or after the model-guided workflow.

The baseline becomes substantially faster when it follows the model-guided block, while editing decreases and rejection increases. Eight of the nine observed baseline missed errors also occur in this order group. We therefore report the order-stratified results as descriptive evidence rather than attributing them to a single mechanism.

D.6 STATISTICAL SCOPE

The study combines two levels of analysis. Final-item quality outcomes are summarized over individual reviewed items, for which the main-text comparisons use Fisher's exact tests. Participant-level behavioral summaries use paired Wilcoxon signed-rank tests because every participant experiences both conditions.

Because the study contains eight annotators and repeated observations within each participant, the item-level tests should be interpreted together with the participant-level and order-stratified results above rather than as independent evidence from 510 separate annotators. Accordingly, the main paper uses the prospective study primarily to evaluate the deployed *human-AI workflow* and the quality of its resulting dataset, rather than to estimate a population-level effect of the recommendation display in isolation.

The intervention also combines candidate curation and recommendation display. The study therefore evaluates their joint system-level effect. Disentangling these components would require a factorial design in which candidate selection and recommendation display are independently randomized.

E Additional Related work

Noise-transition estimation and identifiability. For class-conditional noise, transition matrices are estimable under assumptions such as irreducibility (Scott et al., 2013), anchor points (Liu & Tao, 2015; Patrini et al., 2017), transition revision (Xia et al., 2019), factorization through intermediate labels (Yao et al., 2020), or minimal-volume conditions (Zhang et al., 2021). Instance-dependent noise makes identification substantially harder: Liu et al. (2023) characterize identifiability through Kruskal-rank arguments and show that generic instance-level identification from noisy labels alone requires multiple noisy labels per instance; crowdsourcing exploits precisely such annotator redundancy (Zhang et al., 2016; Guo et al., 2023). Other approaches restore tractability by imposing structure on the corruption process, including part-dependent transitions (Xia et al., 2020), causal generative models (Yao et al., 2021), Bayes-label transitions (Yang et al., 2022), and confidence-scored labels (Berthon et al., 2021). Our route is different: rather than further restricting the instance-dependent channel, we exploit noisy-clean label pairs observed on a reviewed subset. We characterize when these observations identify the correction posterior and, consequently, the clean posterior, forward transition, and clean risk with only one original noisy label per instance.

Clean subsets, paired validation, and label audits. A practical literature combines noisy data with a small amount of trusted supervision. Gold Loss Correction estimates a class-conditional corruption matrix from model predictions on trusted examples (Hendrycks et al., 2018); learning-to-reweight tunes example weights using a clean validation set (Ren et al., 2018); FasTEN estimates a class-conditional transition from clean batches (Kye et al., 2022); Xiao et al. (2015) use both noisy and clean annotations on a subset of Clothing1M. Semi-verified learning similarly studies learning from a limited verified set alongside a larger untrusted sample (Charikar et al., 2017). More closely related, Tjandra & Wiens (2023) use an *alignment set* containing both observed and ground-truth labels to learn instance-dependent label correctness and improve downstream prediction, with a particular motivation from healthcare applications. Active label-correction methods similarly choose which existing noisy labels to verify or repair (Kremer et al., 2018; Kim et al., 2024b). These works demonstrate the algorithmic and empirical value of paired or trusted subsets. We instead take the correction record itself as the object of statistical study: under an unrestricted instance-dependent corruption process, we characterize what partial noisy-clean pairs identify, when information from the reviewed subset can be transported to the full population, and how identification changes under incomplete coverage or clean-label-dependent review. Closer to our data-annotation motivation, Li et al. (2026) introduce Human-Corrected Labels for VLM-based annotation, soliciting human corrections when VLM predictions disagree and deriving a risk-consistent learning objective. Their setting corresponds to a particular discrepancy-triggered correction policy. In our notation, such a deterministic policy has zero review propensity outside the selected region, so the correction posterior there is not identified without additional assumptions. PCL instead treats review and correction as a general observation process, encompassing randomized audits, selective review, and voluntary corrections; also it characterizes the statistical guarantees available under each regime.

Learning from human corrections in applications. Correction records arise naturally in human-in-the-loop systems. Post-edit streams adapt machine-translation systems online (Turchi et al., 2017; Peris et al., 2017); human edits and language feedback are used to fine-tune language models (Ouyang et al., 2022; Scheurer et al., 2023); interactive imitation learning treats interventions as corrective supervision (Kelly et al., 2019). Recent human-LLM annotation systems similarly use automated verification to identify machine-generated labels for human re-annotation (Wang et al., 2024). These approaches demonstrate the practical utility of corrections and selective verification, but primarily consume corrected examples as additional training signal or optimize which items should be reviewed. We instead treat the historical before/after record as an observation of the annotation channel itself and ask when information learned from the corrected subset can be transported to the remaining data. Related work on robust preference optimization handles noisy preferences using known or estimated flip rates (Chowdhury et al., 2024); our framework suggests that analogous instance-dependent correction mechanisms can in principle be learned directly from logged revisions.

Internal validation, missing data, and counterfactual estimation. Our observation structure is also related to classical *internal validation* or *double-sampling* designs, in which an inexpensive fallible measurement is collected broadly and an expensive gold-standard measurement is additionally

obtained on a subsample (Tenenbein, 1970). From this perspective, the review indicator is a missingness indicator on Y: edits at random correspond to a MAR condition (Rubin, 1976; Little & Rubin, 2019), while voluntary error-triggered corrections produce a nonignorable selection problem related to verification bias. Inverse-propensity and doubly robust estimators therefore have well-established antecedents (Robins et al., 1994), as do sensitivity analyses for departures from ignorable selection (Rosenbaum, 2005; Manski, 2003). More recently, prediction-powered inference (Angelopoulos et al., 2023) and design-based supervised learning (Egami et al., 2023) combine large collections of imperfect predictions or annotations with a smaller gold-standard sample to obtain valid downstream statistical inference. These literatures primarily target population parameters or estimating equations. Our focus is instead predictive learning under an instance-dependent annotation channel: paired corrections identify the correction posterior $g(y \mid x, \tilde{y})$, which in turn determines the clean posterior, transition mechanism, and predictive risk, yielding separate statistical dependence on the number of reviewed pairs n_{e} and the full pool size N.