HERO: Improving the Reliability and Sensitivity of Generative Model Evaluation Using Historical Data

Xinrui Ruan[1,2*] **Zhenyu Zhao**[2*] **Waverly Wei**[4]
Yueshan Zhang[1] **Zeyu Zheng**[1,3] **Sui Huang**[2] **Jingshen Wang**[1†]

[1]UC Berkeley [2]Roblox Corp
[3]Berkeley Artificial Intelligence Research (BAIR) Lab [4]University of Southern California

Abstract

Reliable generative AI models critically rely on expert human annotations to evaluate output quality, yet these "gold" labels are expensive to collect and limited in quantity. Organizations thus often turn to collecting vast but noisy "silver" labels from crowdsourced workers or vendor annotators as proxies for gold labels. Because gold remains the evaluation target, naively aggregating noisy silver labels may introduce bias, and estimators built on sparsely observed gold labels may have high variance to resolve the model performance gaps that guide practical decisions. Model evaluation has become an ongoing operational practice rather than a one-time exercise, with evaluation rounds repeating across model versions, releases, and content domains. A natural question is whether the previous historical evaluation data can be used to improve each new round of evaluation. We introduce HERO (History Enhanced RObust model evaluation), a novel framework that uses historical data to suppress bias (improve reliability) and reduce variance (improve sensitivity) in model performance evaluation. HERO calibrates silver labelers' performance learned from historical gold annotations, and stabilizes the resulting estimator by anchoring it to covariate information measured with high precision in the historical data. HERO can be broadly applied across multiple common evaluation tasks, and remains valid when only a subset of historical labelers appears in the current round. We establish conditions under which the bias and variance reductions hold, showcase HERO's performance in simulation studies, and demonstrate its effectiveness on real-world model evaluation benchmarking datasets.

1 Introduction

1.1 Motivation and Contribution

Human evaluation is at the heart of developing and deploying reliable generative AI models [1]. Many organizations rely heavily on human annotations to evaluate model output quality, assess safety, calibrate moderation systems, and compare candidate model versions before release [2–4]. Small differences in estimated model quality, even on the order of a few percentage points in win rate or defect rate, determine which model is shipped to hundreds of millions of users [5, 6].

As widely documented in natural language processing and related fields, a key challenge within an annotation task is that expert-annotated "gold" labels (a.k.a. ground truth labels) are expensive and scarce [2], so organizations often instead rely on judgments from a pool of less specialized

*Xinrui Ruan and Zhenyu Zhao contributed equally. This work was done while XR was an intern at Roblox Corporation.
†Correspondence: jingshenwang@berkeley.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

"silver" annotators whose labeling reliability varies. The so-called *gold* labels may be produced by a single expert annotator or by consensus among multiple experts within the organization, and they serve as the best available reference standard. In contrast, large-scale evaluations typically rely on *silver* labels collected from broader, less specialized sources, including crowd or vendor annotators and automated judges such as LLMs [7] or VLMs [8]. Throughout this paper, we view silver labels as the primary operational labels used in evaluation, and gold labels as a more accurate but limited supervision source used to assess and calibrate them. Because silver annotators differ in reliability, simply aggregating their labels without adjustment can reduce the reliability of evaluation and introduce bias in the estimated model performance.

Many methods have been developed to account for annotator variability [2, 9–11], but most operate within a single evaluation round and do not fully leverage the unique structure of modern model evaluation settings. We note that two novel features distinguish modern AI model evaluations from classical annotation tasks. *First*, organizations that evaluate generative models accumulate large-scale historical annotation data across prior evaluation rounds. These datasets often include repeated observations of annotator behavior and subsets of instances with gold labels, which together reveal patterns in labeler reliability. *Second*, the historical and current model evaluation tasks may share content that transfers across rounds, such as prompt features. Generation prompts often persist across rounds because organizations evaluate successive model versions on a stable prompt distribution drawn from real user traffic or curated benchmarks [1], so the same prompt features that appeared in earlier rounds also characterize the current evaluation set. With these two features, in principle, labeler reliability learned from historical gold labels can calibrate current annotations, and the shared generation prompts available in a large historical sample can stabilize current model evaluation estimates. We defer a more complete comparison with existing literature to Appendix F.

To fully exploit historical data for reliable model evaluation, we propose HERO (**H**istory **E**nhanced **RO**bust model evaluation), a framework that adjusts noisy annotation-based estimators with historical evaluation data to reduce both bias and variance in model performance estimates. The key contributions of our work include:

- We elucidate two unique features that distinguish modern AI evaluation from classical annotation literature (Figure 2), namely the accumulation of historical annotation data with gold-labeled subsets, and prompt feature transfer from earlier rounds to the current evaluation round.
- We identify bias sources that arise in this evaluation setting through real-world benchmarking data, isolating heterogeneous silver labeler reliability and labeler pool shift between rounds (Figure 1).
- We propose a theoretically grounded yet simple method, HERO, that combines bias reduction from labeler reliability estimation with variance reduction from control variates (Section 3). Our control variate construction is customized at the score level, departing from the existing literature that operates on raw outcomes.
- We show that the proposed method applies across multiple common model evaluation use cases (Table 1) and validate our method with comprehensive empirical evidence (Sections 4 and 5).

1.2 Model Evaluation Target Parameters and Tasks

We organize common model evaluation scenarios into a unified observational structure. Consider n instances indexed by $i = 1, \ldots, n$. For each instance, Y_i^* denotes the *operational ground-truth gold label*, the best available adjudicated assessment of the generated output, and Z_i denotes the noisy *silver labels* collected from less specialized annotators or automated AI systems as proxies for Y_i^*. The target parameter is a function of Y_i^*, and the task-specific outcome Y_i is constructed from Y_i^* according to the scenario. Table 1 summarizes four evaluation scenarios, listing the outcome variable Y_i, the corresponding target parameter, and the available covariates for each. We summarize each case in Table 1 below and defer its detailed explanation to the appendix.

2 Bias and Variance in Model Evaluation

Estimating the target parameter $\mathbb{E}[Y_i]$ from gold labels alone is rarely feasible in practice. This is because many evaluation tasks provide no gold labels at all. Even when a subset of instances is annotated by a gold labeler, the labeled fraction is typically small. The sample mean over this subset is unbiased, but its variance remains too large to resolve the differences in model performance

Use Case	Outcome Y_i	Target Parameter
1. Single Model Evaluation	Gold label for Model A: $Y_i = Y^*_{i,\mathtt{A}}$	Average performance: $\mu_{\mathtt{A}} = \mathbb{E}[Y_i]$
2. Model Comparative Evaluation	Difference in gold labels: $Y_i = Y^*_{i,\mathtt{A}} - Y^*_{i,\mathtt{B}}$	Average performance difference: $\alpha = \mathbb{E}[Y_i]$
3. Side-by-Side Model Comparison	Binary preference indicator: $Y_i = \mathbb{1}(Y^*_{i,\mathtt{A}} \succ Y^*_{i,\mathtt{B}})$	Win rate for Model A: $\pi_{\mathtt{A}\succ\mathtt{B}} = \mathbb{E}[Y_i]$
4. Safety Moderation Calibration	True positive flag: $Y_i = Y_i^* Y_i^{\mathtt{AI}}$	Precision $= \mathbb{E}[Y_i]/\mathbb{E}[Y_i^{\mathtt{AI}}]$; Recall $= \mathbb{E}[Y_i]/\mathbb{E}[Y_i^*]$

Table 1: Summary of model evaluation use cases.

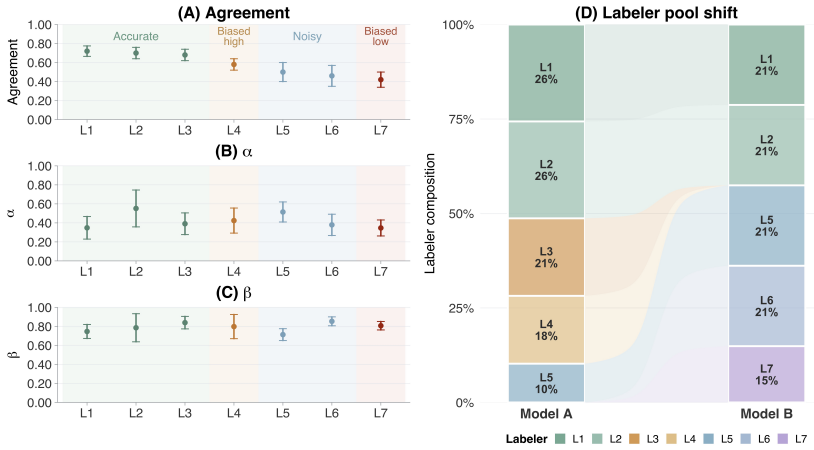


Figure 1: Panels (A)–(C) show heterogeneous silver labeler agreement rates, sensitivity α_j, and specificity β_j, with 95% confidence intervals in our benchmarking data (Appendix F, Asset detect detection data). Panel (D) shows the labeler pool shift between the two model evaluation rounds.

that guide practical decisions. For these reasons, practitioners often construct a proxy for Y_i by aggregating the available silver labels Z_i. Let $\widehat{\mu}$ denote an estimator of $\mathbb{E}[Y_i]$ constructed from a silver-labeler-based proxy. To assess estimator quality, we decompose the mean-squared error (MSE) into bias and variance:

$$\underbrace{\mathbb{E}\Big[\big(\widehat{\mu} - \mathbb{E}[Y_i]\big)^2\Big]}_{\text{MSE relative to target}} = \underbrace{\big(\mathbb{E}[\widehat{\mu}] - \mathbb{E}[Y_i]\big)^2}_{\text{Bias}^2} + \underbrace{\mathrm{Var}(\widehat{\mu})}_{\text{Variance}}. \tag{1}$$

Both components introduce distinct challenges. We start by examining the bias component through two sources that arise in generative model evaluation. The first source of bias is annotator heterogeneity. The inter-rater reliability literature consistently finds variation in annotator accuracy across domains from medical imaging [12, 13] to natural language processing [14], even under shared guidelines [9, 15–17]. Differences in expertise, interpretation standards, and decision thresholds cause some labelers to favor precision over recall, while others show the opposite tendency. Aggregation without calibration, such as the majority vote-based estimators, allows these systematic tendencies, particularly from lower-quality labelers, to shift the estimate of $\mathbb{E}[Y_i]$ away from the true target.

As shown in Figure 1 Panels A-C, the seven silver labelers in our benchmarking data span agreement rates from below 0.4 to above 0.8, and the sensitivity α_j and specificity β_j panels reveal asymmetric error profiles behind that spread, with some labelers favoring positives, others favoring negatives, and a third group noisy on both axes. A majority-vote estimator weights these labelers equally and therefore inherits the mixture rather than the gold-labeler target.

A second source of bias arises when the composition of the labeler pool changes between evaluation rounds. For example, in comparative evaluation tasks (Use Case 2 in Table 1), organizations may evaluate a newer model against an earlier version at different times. The labeler pools that annotate each model may only partially overlap. A reliable annotator who contributes a large fraction of Model A labels may complete only a small fraction of Model B labels, and vice versa. When the proportion of reliable annotators differs between the two rounds, the observed label difference reflects

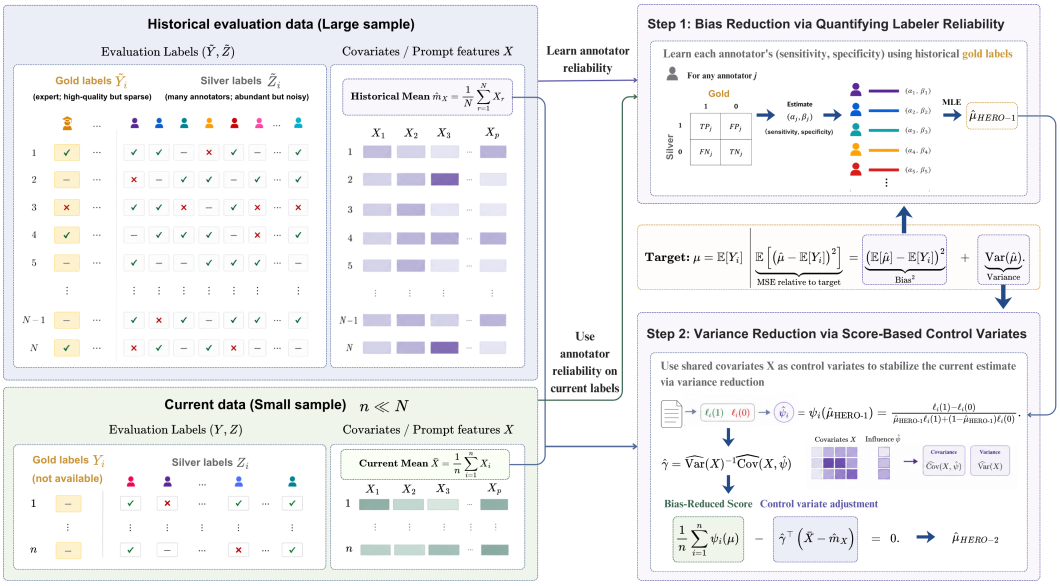


Figure 2: Overview of the History-Enhanced Robust (HERO) evaluation framework. HERO combines a large historical dataset $(\widetilde{Y}_i, \widetilde{Z}_i, \widetilde{X}_i)_{i=1}^{N}$ with a small current sample $(Z_i, X_i)_{i=1}^{n}$ $(n \ll N)$ to estimate the target parameter $\mu = \mathbb{E}[Y_i]$ with Y_i unobserved on current evaluation data.

both true differences in model quality and shifts in the composition of the labeler pool. An estimator that treats the two label sets as exchangeable confounds these two sources of variation. Figure 1 Panel D illustrates this shift in our benchmarking data. Two labelers active in the Model A round do not appear in the Model B round, and the labelers active in both rounds contribute different proportions of labels across the two models.

The variance component of the MSE in Equation (1) remains a challenge even after addressing the bias issues mentioned above. Even with a well-calibrated silver label based proxy, a small labeled set or high labeling outcome heterogeneity across instances can produce confidence intervals too wide to resolve the performance gaps that guide practical decisions. A variance reduction technique that halves the estimator variance is equivalent to doubling the effective sample size, so that an evaluation previously requiring 1,000 labeled instances can reach the same precision with 500, cutting annotation cost and turnaround time proportionally.

3 Method: History Enhanced RObust (HERO) Model Evaluation

In this paper, we propose a unified framework to reduce both bias and variance for an *accurate* (improved reliability) and *efficient* (improved sensitivity) estimation of the target parameter μ. Achieving both goals requires an external source of information beyond the only current silver labels. As organizations that evaluate generative models often maintain historical evaluation datasets from prior rounds, our framework leverages these historical data to construct an estimator of μ with reduced MSE. We term the proposed framework History Enhanced RObust (HERO) model evaluation and organize it as a two-step procedure that first reduces the bias and then controls the variance.

At a very high level, HERO works as follows. When we analyze current evaluation data with only silver labels, HERO first constructs calibrated gold proxies that weight each silver labeler by the reliability learned from historical data, yielding a bias-reduced estimator $\widehat{\mu}_{\text{HERO-1}}$. HERO then incorporates historical covariates (e.g., prompt features) as control variates to further reduce the variance of $\widehat{\mu}_{\text{HERO-1}}$, yielding the final estimator $\widehat{\mu}_{\text{HERO-2}}$. This estimator satisfies:

- $\widehat{\mu}_{\text{HERO-1}}$ and $\widehat{\mu}_{\text{HERO-2}}$ both have reduced bias compared to the majority-vote-based estimator.
- The variance of $\widehat{\mu}_{\text{HERO-2}}$ is reduced compared to $\widehat{\mu}_{\text{HERO-1}}$ when shared covariates between the current and historical data are available.

3.1 Historical and Current Evaluation Data

To distinguish historical data from the current evaluation task, we introduce a consistent notation. We use a tilde to denote quantities associated with historical data. The historical dataset consists of N instances collected from earlier evaluation rounds and is defined as follows:

- Noisy silver labels $\widetilde{Z}_i$ from a pool of m silver labelers. Not all labelers annotate every instance. The availability indicator $\widetilde{A}_{ij} \in \{0, 1\}$ records whether labeler j provided a label for instance i.
- A set of instances together with their associated covariates $\{\widetilde{X}_i\}_{i=1}^N$, such as prompt features, metadata, or contextual information. These prompts may *reappear* in future evaluation tasks.
- Gold labels $\widetilde{Y}_i$ are collected on a subset of historical instances to audit silver labeler performance and identify which labelers are consistently accurate. These gold labels are obtained under a randomized routing mechanism with known selection probability given $\widetilde{X}_i$. The routing indicator $\widetilde{R}_i \in \{0, 1\}$ equals 1 when instance i is sent to a gold labeler and 0 otherwise. Gold labels are observed only when $\widetilde{R}_i = 1$.

Meanwhile, the current evaluation dataset contains n instances, where $n \ll N$:

- A set of instances together with their associated covariates $\{X_i\}_{i=1}^n$, including prompt features, metadata, and contextual information.
- No gold labels are observed in the current task.
- Noisy silver labels Z_i with availability indicators $A_{ij} \in \{0, 1\}$.

Throughout this section, all outcomes and labels are binary, so that $Y_i \in \{0, 1\}$ and $\widetilde{Y}_i \in \{0, 1\}$. The target parameter remains $\mu = \mathbb{E}[Y_i]$ as defined in Section 1.2, but gold label Y_i is unobserved.

Scope of historical data. Our framework allows two kinds of mismatch between current and historical data. First, the current annotator pool can differ, but it is a subset of the labelers who contributed to the historical data. Second, the historical task itself may differ from the current task, since HERO calibrates labeler reliability rather than item-level labels.

3.2 HERO Step 1: Bias Reduction via Quantifying Labeler Reliability

To account for model evaluation bias introduced by silver labeler heterogeneity, historical data provide a natural source for revealing persistent patterns in labeler behavior. HERO Step 1 extracts these reliability patterns from the historical data and transfers them to the current evaluation task, where no gold labels are available. More concretely, we characterize each silver labeler j by two parameters that describe intrinsic properties of that labeler's decision process,

$$\alpha_j = P(Z_{ij} = 1 \mid Y_i = 1), \quad \beta_j = P(Z_{ij} = 0 \mid Y_i = 0).$$

Here α_j is the sensitivity and β_j is the specificity. Because these parameters describe the labeler rather than the task, they are expected to transfer from the historical dataset to the current dataset even when the underlying evaluation content/task differs.

HERO estimates (α_j, β_j) for each labeler j from the historical dataset $\{(\widetilde{X}_i, \widetilde{R}_i\widetilde{Y}_i, \widetilde{Z}_i, \widetilde{R}_i, \widetilde{A}_{ij})\}_{i=1}^N$ using an EM algorithm that treats unobserved gold labels as latent variables, and use observed gold labelers as anchors. Given a candidate value $y \in \{0, 1\}$ for the gold label, the likelihood of the observed silver labels for historical instance i is

$$\widetilde{\ell}_i(y) = \prod_{j\,:\,\widetilde{A}_{ij}=1} P(\widetilde{Z}_{ij} \mid \widetilde{Y}_i = y), \tag{2}$$

where $P(\widetilde{Z}_{ij} \mid \widetilde{Y}_i = 1) = \alpha_j^{\widetilde{Z}_{ij}} (1 - \alpha_j)^{1-\widetilde{Z}_{ij}}$ and $P(\widetilde{Z}_{ij} \mid \widetilde{Y}_i = 0) = (1 - \beta_j)^{\widetilde{Z}_{ij}} \beta_j^{1-\widetilde{Z}_{ij}}$. An instance without a gold label ($\widetilde{R}_i = 0$) contributes the marginal $\widetilde{\mu}\,\widetilde{\ell}_i(1) + (1 - \widetilde{\mu})\,\widetilde{\ell}_i(0)$, where $\widetilde{\mu} = \mathbb{E}[\widetilde{Y}] \neq \mathbb{E}[Y]$. The full historical log-likelihood is

$$\widetilde{\mathcal{L}}\big(\widetilde{\mu},\, \{\alpha_j, \beta_j\}_{j=1}^m\big) = \sum_{i=1}^{N} \Big[\widetilde{R}_i\, \log \widetilde{\ell}_i(\widetilde{Y}_i) + (1 - \widetilde{R}_i)\, \log\big(\widetilde{\mu}\,\widetilde{\ell}_i(1) + (1 - \widetilde{\mu})\,\widetilde{\ell}_i(0)\big)\Big]. \tag{3}$$

To stabilize the EM updates, HERO initializes the parameters using only the gold-labeled subset $\mathcal{G} = \{i : \widetilde{R}_i = 1\}$. Specifically, we set

$$\widetilde{\mu}^{(0)} = |\mathcal{G}|^{-1} \sum_{i \in \mathcal{G}} \widetilde{Y}_i, \qquad \alpha_j^{(0)} = \frac{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} \widetilde{Y}_i \widetilde{Z}_{ij}}{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} \widetilde{Y}_i}, \qquad \beta_j^{(0)} = \frac{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} (1 - \widetilde{Y}_i)(1 - \widetilde{Z}_{ij})}{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} (1 - \widetilde{Y}_i)}.$$

HERO then maximizes $\widetilde{\mathcal{L}}$ via the EM algorithm (see Appendix Algorithm 1, Stage A), producing estimates $(\widehat{\alpha}_j, \widehat{\beta}_j)$ that are treated as fixed when estimating μ on the current data.

The current-data likelihood $l_i(y)$ follows the same structure as defined in (2), with (Z_{ij}, A_{ij}) replacing $(\widetilde{Z}_{ij}, \widetilde{A}_{ij})$ and the estimates $(\widehat{\alpha}_j, \widehat{\beta}_j)$ replacing the unknown parameters. Because no gold labels are observed, every instance contributes through the marginal form. The HERO Step 1 estimator maximizes the log-likelihood over $\mu \in (0, 1)$,

$$\widehat{\mu}_{\text{HERO-1}} = \arg \max_{\mu \in (0,1)} \sum_{i=1}^{n} \log \big[\mu \, \ell_i(1) + (1 - \mu) \, \ell_i(0) \big]. \tag{4}$$

The first-order condition is

$$\sum_{i=1}^{n} \psi_i(\mu) := \sum_{i=1}^{n} \frac{\ell_i(1) - \ell_i(0)}{\mu \, \ell_i(1) + (1 - \mu) \, \ell_i(0)} = 0, \tag{5}$$

which can be solved by any standard one-dimensional optimizer because μ is scalar. Here, $\psi_i(\mu)$ denotes the individual score contribution from instance i.

Remark 1 (Comparison with majority vote)**.** *The instance-level ratio $\ell_i(1)/\ell_i(0)$ controls how each instance influences $\widehat{\mu}_{\text{HERO-1}}$. When most high-sensitivity labelers agree on a positive label, this ratio is large, pulling the estimate toward the value implied by accurate annotators. Conversely, agreement among labelers with low specificity on a negative label carries less weight. When all labelers share identical accuracy, $\ell_i(1)/\ell_i(0)$ reduces to a monotone function of the majority-vote count, and the estimator recovers a majority-vote-type solution as a special case.*

3.3 HERO Step 2: Variance Reduction via Score-Based Control Variates

On top of bias reduction, HERO can further reduce estimator variance by leveraging auxiliary covariates that are shared between the historical and current datasets. Prompt features are a natural shared covariate, since organizations typically evaluate successive model versions on a stable prompt distribution drawn from real user traffic or curated benchmarks, and the same prompt features that appeared in earlier rounds also characterize the current evaluation set.

The key observation is that the historical dataset ($N \gg n$), being much larger than the current evaluation sample, provides a substantially more precise estimate of the covariate mean. This historical mean can therefore serve as a stable reference in a control-variate adjustment without changing the target parameter.

Yet, the classical control variate adjustment does not transfer directly in our task, because the Step 1 estimator is a calibrated functional of EM posteriors rather than a sample mean. We thus propose to apply the control variate adjustment to the estimating equation that defines the Step 1 HERO estimator, with a shared covariate function whose population mean is supplied by the historical sample. Let X_i denote a covariate observed in the current evaluation task, and let $\widehat{m}_X = \frac{1}{N} \sum_{r=1}^{N} \widetilde{X}_r$ denote the sample mean of the same covariate in the historical data. We construct a control-variate estimator based on the score equation in (5). After obtaining the pilot HERO estimator $\widehat{\mu}_{\text{HERO-1}}$ from Step 1, define the estimated score contribution for instance i as

$$\widehat{\psi}_i := \psi_i(\widehat{\mu}_{\text{HERO-1}}) = \frac{\ell_i(1) - \ell_i(0)}{\widehat{\mu}_{\text{HERO-1}} \ell_i(1) + \big(1 - \widehat{\mu}_{\text{HERO-1}}\big) \ell_i(0)}. \tag{6}$$

We then estimate the control-variate coefficient by regressing the pilot score $\widehat{\psi}_i$ on X_i:

$$\widehat{\gamma} = \widehat{\text{Var}}(X)^{-1} \widehat{\text{Cov}}(X, \widehat{\psi}), \tag{7}$$

where $\widehat{\text{Var}}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top$ and $\widehat{\text{Cov}}(X, \widehat{\psi}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) \widehat{\psi}_i$.

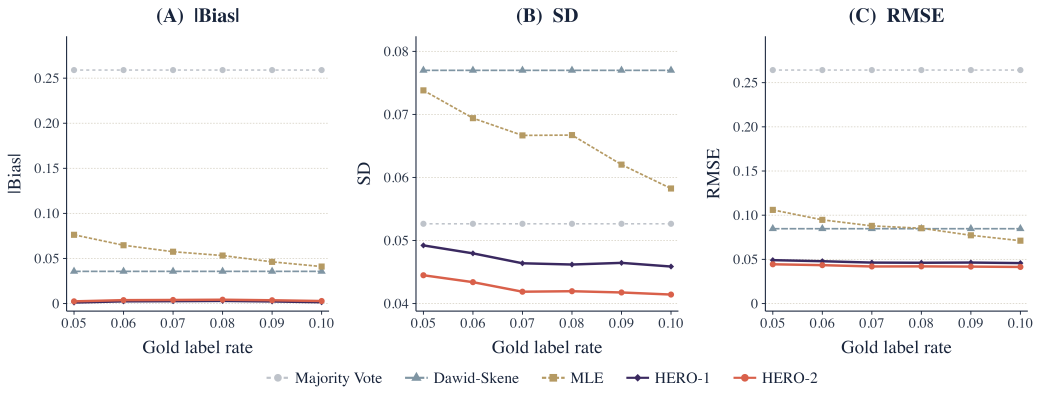


Figure 3: Experiment 1: (A) absolute bias, (B) Monte Carlo standard deviation, and (C) RMSE.

Whenever the covariate mean is shared between the historical and current datasets, the difference $\bar{X} - \widehat{m}_X$ is asymptotically mean zero. Therefore, the adjustment term $\widehat{\gamma}^\top(\bar{X} - \widehat{m}_X)$ can be used to reduce variance without changing the asymptotic expectation of the estimator. We define the control-variate adjusted estimator $\widehat{\mu}_{\text{HERO-2}}$ as the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i(\mu) - \widehat{\gamma}^\top(\bar{X} - \widehat{m}_X) = 0. \tag{8}$$

As in Step 1, this equation can also be solved using any standard one-dimensional optimizer.

This adjustment reduces variance whenever the covariate X_i is correlated with the score contribution ψ_i, which quantifies how each observation influences the estimator. If X_i carries little information about the outcome, then $\widehat{\gamma}$ is close to zero, and the estimator effectively reduces to $\widehat{\mu}_{\text{HERO-1}}$. Consequently, HERO Step 2 improves efficiency when informative auxiliary covariates are available, while leaving the estimator essentially unchanged when they are not.

3.4 Theoretical Guarantees

On top of being practical and easy to implement, HERO enjoys theoretical guarantees along three dimensions: reliability (Theorem 1), sensitivity (Theorem 2), and valid statistical inference (Theorem 3). Reliability means that HERO accurately targets the gold-label estimand, rather than a proxy estimand distorted by heterogeneous silver labelers. Sensitivity means that HERO reduces variance, which may improve the detection of small but practically meaningful model-performance differences. Valid statistical inference means that the estimator admits an asymptotic distribution, providing the basis for credible confidence intervals. Due to space limit, we defer the formal theorem statements and proofs to Appendix C.

4 Simulation Study

In this section, we evaluate the performance of HERO method against several benchmark methods. We summarize our takeaway as follows: First, HERO-1 remains accurate even when the historical gold-label rate is small, whereas the sample-proportion calibration baseline is unstable when only a few gold-routed examples are available per labeler. This shows HERO-1 can extract useful reliability information from the full historical label matrix, not only from the gold-labeled subset. Second, HERO-2 further improves sensitivity by reducing variance and RMSE relative to HERO-1 when the shared covariate is informative. Third, under increasing labeler heterogeneity, HERO-1 and HERO-2 remain substantially more robust than baseline methods, confirming that historical reliability calibration and score-level control variates reduce estimation errors. Due to the page limit, we defer the data generating process details to Appendix. In what follows, we introduce the methods in comparison, evaluation metrics, and simualtion study results.

Methods in comparison. We compare five estimators: (1) Majority vote. (2) Dawid–Skene method with current data only: The current-only Dawid–Skene baseline fits the standard latent-class

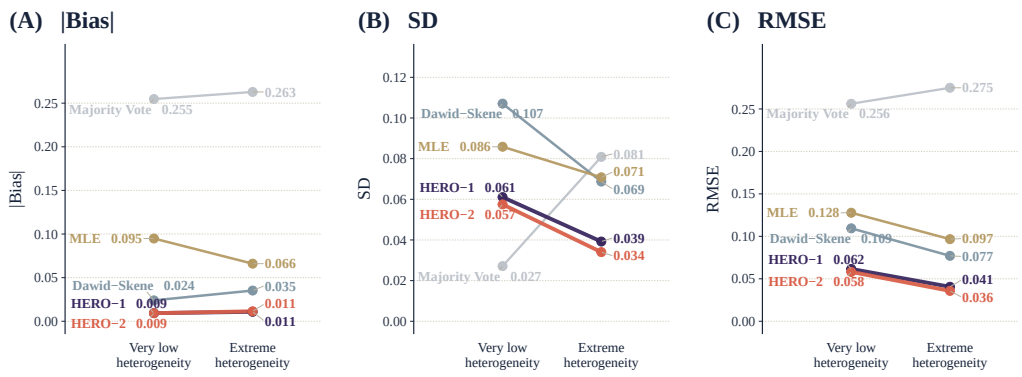


Figure 4: Experiment 2: Comparison of (A) absolute bias, (B) Monte Carlo standard deviation, and (C) RMSE as the Beta concentration k decreases, increasing the dispersion of labeler-specific sensitivities and specificities

model using the current silver labels alone, jointly estimating μ and the labeler-specific parameters $\{(\alpha_j, \beta_j)\}_{j=1}^{J}$. (3) Historical sample-proportion calibration. We estimate each labeler's sensitivity and specificity using only the gold-labeled subset of the historical data. With Laplace smoothing. We then maximize the current-data likelihood over μ with these calibration estimates fixed. (4) HERO-1 only use historical data for bias reduction. (5) HERO-2 applies the score-anchored control-variate adjustment to HERO-1. In this simulation, the population mean $\mathbb{E}[X] = 0$ is known, but we use $\bar{X}_{\text{hist}}$ to mirror the practical setting in which HERO relies on a high-precision historical estimate rather than oracle knowledge. Appendix Table 4 summarizes the simulation settings. The target μ_0 is computed from the data-generating model, $\mu_0 = \mathbb{E}_X\{\sigma(\beta_0 + \beta_1 X)\}$, and is fixed within each experimental setting.

Experiment 1: Historical gold-label rate. The first experiment varies the amount of historical gold supervision. In Setting (ii), we use a larger labeler pool and a stronger covariate signal: $n = 300$, $N = 800$, $J = 10$, and $\beta_1 = 3$. We vary $\rho \in \{0.05, 0.06, \ldots, 0.10\}$. **Experiment 2: Labeler heterogeneity.** The second experiment studies robustness to heterogeneity in silver-labeler accuracy. We use the same baseline configuration as Experiment 1 Setting (ii), with $n = 300$, $N = 800$, $J = 10$, $\beta_0 = -0.2$, and $\beta_1 = 3$. We hold the mean reliability fixed at $\mu_\alpha = 0.85$ and $\mu_\beta = 0.55$, and vary only the Beta concentration parameter k. Under the parametrization $\alpha_j \sim \text{Beta}\{k\mu_\alpha, k(1-\mu_\alpha)\}$, $\beta_j \sim \text{Beta}\{k\mu_\beta, k(1-\mu_\beta)\}$, the means satisfy $\mathbb{E}[\alpha_j] = \mu_\alpha$ and $\mathbb{E}[\beta_j] = \mu_\beta$ for every k, while $\text{Var}(\alpha_j) = \frac{\mu_\alpha(1-\mu_\alpha)}{k+1}$, $\text{Var}(\beta_j) = \frac{\mu_\beta(1-\mu_\beta)}{k+1}$. Thus, decreasing k increases labeler heterogeneity without changing the average sensitivity or specificity. We consider the five heterogeneity levels in Appendix Table 4. For each level, we evaluate historical gold-label rates $\rho \in \{0.05, 0.10\}$.

Evaluation metrics. For each configuration, we run $B = 200$ Monte Carlo replications. We report absolute bias, Monte Carlo standard deviation, and root mean squared error (RMSE). All EM algorithms use convergence tolerance 10^{-6} and a maximum of 200 iterations.

Simulation study results. Figure 3 shows that Majority vote has nearly constant bias as ρ varies, since it does not use historical gold labels. The current-only Dawid–Skene baseline also does not improve with ρ and remains limited by the identifiability and finite-sample challenges of estimating all labeler parameters from the small current round alone. The historical sample-proportion baseline improves as ρ increases, but it is unstable when the number of gold-routed examples per labeler is small. In contrast, HERO-1 is substantially less sensitive to low gold-label rates because its anchored EM step uses both the gold-routed and ungolded historical instances. HERO-2 further reduces SD and RMSE by exploiting the correlation between X_i and the score contribution, with larger variance reduction in the stronger-signal setting. We provide additional results in Appendix Figure 6.

Figure 4 shows that majority vote becomes increasingly unreliable as labeler heterogeneity grows, because it assigns equal weight to labelers with different error profiles. The current-only Dawid–Skene baseline partially adapts to heterogeneous reliability but remains unstable because the current sample

Estimator	Mean estimate	|Bias|	SD	RMSE
Majority Vote	0.110	0.130	**0.019**	0.131
Dawid–Skene	0.120	0.120	0.027	0.123
MLE (hist. sample mean)	0.069	0.171	0.121	0.209
HERO-1	0.302	0.062	0.070	0.093
HERO-2	0.288	**0.048**	0.064	**0.080**

Table 2: Case study I (image defect detection), gold-label rate 0.10.

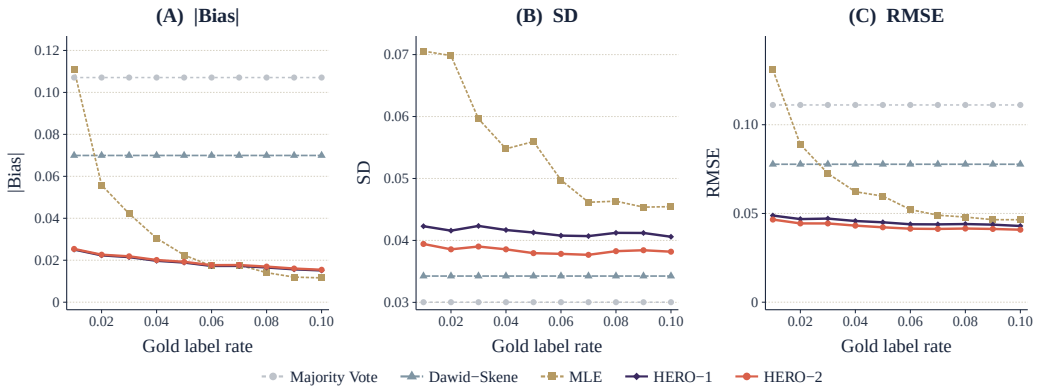


Figure 5: Case study II: Results in model safety evaluation dataset.

is small and unanchored by gold labels. Historical sample-proportion calibration improves when enough gold labels are available, but it can still have high variance when each labeler receives few gold-routed examples. HERO-1 maintains low bias across heterogeneity levels by pooling information from the full historical label matrix, and HERO-2 achieves the lowest RMSE by combining this calibration with score-level variance reduction.

5 Results in Real-World Benchmarking Data

In this section, we evaluate the proposed HERO method on two real-world datasets provided by a technology company that routinely conducts large-scale generative model evaluations. The datasets were collected to assess the performance of 3D asset generation models and contain both historical and current evaluation data. The case studies cover two representative tasks: generation defect evaluation and safety evaluation. In each task, every prompt–asset pair is evaluated by a small number of randomly selected silver labelers according to the task-specific evaluation objective, and a random subset of pairs is additionally evaluated by a gold labeler, who is a well-trained expert. To assess statistical performance and robustness, we use a resampling-based approach to report absolute bias, SD, and RMSE for the five estimators described in Section 4. Additional details on the real-world datasets are provided in Appendix E.

Table 5 and Figure 5 show that HERO method reduces absolute bias and RMSE relative to the baseline methods. This improvement is driven by the presence of labeler heterogeneity and labeler-pool shift between the historical and current datasets, as illustrated in Figure 1. HERO is designed to exploit precisely these sources of information by estimating historical labeler reliabilities and transferring them to the current evaluation task. Furthermore, HERO-2 achieves additional variance reduction through control variates. In Case Study I, shown in Table 5, HERO-2 uses GPT-5-predicted labels as the control variate, as all generated contents were evaluated by the same model. In Case Study II, shown in Figure 5, HERO-2 uses TF-IDF features as the control variate. In both cases, HERO-2 reduces the SD of HERO-1 by approximately 5%–10%, demonstrating that the proposed variance-reduction framework is effective across different choices of control variates.

In contrast, the classical Dawid–Skene baseline does not effectively learn labeler heterogeneity in either case. As a result, it performs worse than HERO and is often closer to Majority Vote in terms of bias and RMSE.

Limitation. Despite these promising real-world results, HERO Step 1 requires labeler reliability to transfer from historical data to the current evaluation task and Step 2 requires the shared covariates to have comparable distributions across the two periods. We view these requirements as mild, but they may not hold uniformly across model evaluation settings.

References

[1] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

[2] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y 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*, pages 254–263, 2008.

[3] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

[4] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[5] Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*, 2025.

[6] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

[7] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

[8] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[9] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 20–28, 1979.

[10] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22, 2009.

[11] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.

[12] Simon K. Warfield, Kelly H. Zou, and William M. Wells. An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[13] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[14] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. 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 (EMNLP)*, pages 254–263. Association for Computational Linguistics, 2008.

[15] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[16] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[17] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pages 2035–2043, 2009.

[18] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39 (1):1–22, 1977.

[19] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

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

[21] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

[22] 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*, pages 614–622, 2008.

[23] Art B. Owen. *Monte Carlo Theory, Methods and Examples*. Stanford University, November 2013. URL `https://artowen.su.domains/mc/`.

[24] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, pages 295–318, 2013.

[25] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132, 2013.

[26] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer, 2004.

[27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[28] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[29] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3356–3369, 2020.

[30] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

Technical appendices and supplementary material

A Use cases

Following the setup in Section 2 in the main manuscript, we provide more details on the four use cases for model evaluations:

Use Case 1. Single Model Evaluation. This scenario focuses on evaluating the performance of a single model (such as defect rate), Model A. We define the outcome as the gold label for the model, $Y_i = Y^*_{i,\mathtt{A}}$. The target parameter, which represents the average performance or expected quality, is defined as $\mu_{\mathtt{A}} = \mathbb{E}[Y_i]$.

Use Case 2. Model Comparative Evaluation. This scenario independently evaluates two models, Model A and Model B, on the same prompts to calculate the absolute difference in their performance. We define the outcome as the difference in gold labels between the two models, $Y_i = Y^*_{i,\mathtt{A}} - Y^*_{i,\mathtt{B}}$. The target parameter for the average difference in performance is expressed as $\alpha = \mathbb{E}[Y_i]$.

Use Case 3. Side-by-Side Model Comparison. This scenario involves a direct, simultaneous comparison between outputs from multiple models to establish a relative preference. We define the outcome Y_i as a binary indicator of preference rather than an absolute score, such that it equals 1 if the gold labeler prefers Model A over Model B, and 0 otherwise. This can be formulated as $Y_i = \mathbb{1}(Y^*_{i,\mathtt{A}} \succ Y^*_{i,\mathtt{B}})$. The target parameter, representing the win rate or expected preference for Model A, is $\pi_{\mathtt{A}\succ\mathtt{B}} = \mathbb{E}[Y_i]$.

Use Case 4. Safety Moderation Calibration. This scenario treats the online safety moderation system as an AI automated rater and the offline human review as the gold truth. We let $Y_i^{\mathtt{AI}} \in \{0, 1\}$ represent the online model's evaluation, where $Y_i^{\mathtt{AI}} = 1$ indicates the content is flagged as unsafe. We let $Y_i^* \in \{0, 1\}$ represent the offline human-labeled ground truth, where $Y_i^* = 1$ indicates an actual safety violation. We define the outcome as the intersection of the gold and AI labels, representing a true positive flag, $Y_i = Y_i^* Y_i^{\mathtt{AI}}$. The target parameters include the actual violation percentage or prevalence base rate, $p = \mathbb{E}[Y_i^*]$. Additionally, we estimate the online model precision, which is the probability of a true violation given the model flagged it, calculated as Precision $= \frac{\mathbb{E}[Y_i]}{\mathbb{E}[Y_i^{\mathtt{AI}}]}$. The online model recall, which is the probability the model flagged the content given it is a true violation, is defined as Recall $= \frac{\mathbb{E}[Y_i]}{\mathbb{E}[Y_i^*]}$.

B HERO Algorithmic Presentation

Please see Algorithm 1.

C Theoretical Guarantees of HERO

This section establishes the theoretical guarantees of HERO. The theoretical results rely on three assumptions on the labeling mechanism, the transportability of labeler reliability between historical and current data, and the sampling structure. Lemma 1 shows that the reliability estimates $\widehat{\alpha}_j$ and $\widehat{\beta}_j$ converge at rate $O_p(N^{-1/2})$ when historical gold labels anchor the EM, and Lemma 2 shows that the same EM algorithm without a gold anchor may fail to recover the true labeler parameters. Three theorems then state the main results. Theorem 1 shows that HERO-1 attains the parametric rate $O_p(n^{-1/2})$ for the gold-label target, while the majority-vote and unanchored Dawid–Skene baselines are not guaranteed to converge to that target. Theorem 2 shows that HERO-2 reduces the asymptotic variance of HERO-1 whenever the shared covariate correlates with the score function, and preserves the bias of HERO-1 up to higher-order terms. Theorem 3 establishes asymptotic normality for both estimators, providing the basis for standard error estimation and Wald-type confidence intervals.

Assumption 1 (Randomized labeling)**.** *The silver-label availability indicators A_{ij} and $\widetilde{A}_{ij}$ are completely randomized, satisfying*

$$P(A_{ij} = 1) \geq \pi_{\min} > 0, \quad P(\widetilde{A}_{ij} = 1) \geq \pi_{\min} > 0,$$

and the historical gold-routing indicators $\widetilde{R}_i$ are completely randomized, satisfying

$$P(\widetilde{R}_i = 1) = \rho \geq \rho_{\min} > 0,$$

Algorithm 1: HERO

Input: Historical data $\{(\widetilde{Z}_i, \widetilde{A}_i, \widetilde{R}_i, \widetilde{Y}_i, \widetilde{X}_i)\}_{i=1}^{N}$;
Current data $\{(Z_i, A_i, X_i)\}_{i=1}^{n}$
Output: $\widehat{\mu}_{\text{HERO}}$

/* Stage A. Initialize from the gold-routed historical subset and maximize $\widetilde{\mathcal{L}}(\widetilde{\mu}, \{\alpha_j, \beta_j\})$ via EM */

Let $\mathcal{G} = \{i : \widetilde{R}_i = 1\}$. Initialize, for $j = 1, \ldots, m$,

$$\widetilde{\mu} \leftarrow |\mathcal{G}|^{-1} \sum_{i \in \mathcal{G}} \widetilde{Y}_i, \qquad \alpha_j \leftarrow \frac{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} \widetilde{Y}_i \widetilde{Z}_{ij}}{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} \widetilde{Y}_i}, \qquad \beta_j \leftarrow \frac{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} (1 - \widetilde{Y}_i)(1 - \widetilde{Z}_{ij})}{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} (1 - \widetilde{Y}_i)}.$$

Clip all initial values to $[\delta, 1 - \delta]$ for a small $\delta > 0$

repeat

  E-step:

  foreach $i = 1, \ldots, N$ **do**

    if $\widetilde{R}_i = 1$ **then** $\widetilde{\gamma}_i \leftarrow \widetilde{Y}_i$

    else $\widetilde{\gamma}_i \leftarrow \widetilde{\mu}\, \widetilde{\ell}_i(1) / \big[\widetilde{\mu}\, \widetilde{\ell}_i(1) + (1 - \widetilde{\mu})\, \widetilde{\ell}_i(0)\big]$

  M-step:

  $\widetilde{\mu} \leftarrow N^{-1} \sum_{i=1}^{N} \widetilde{\gamma}_i$

  foreach $j = 1, \ldots, m$ **do**

$$\alpha_j \leftarrow \frac{\sum_i \widetilde{A}_{ij} \widetilde{\gamma}_i \widetilde{Z}_{ij}}{\sum_i \widetilde{A}_{ij} \widetilde{\gamma}_i}, \qquad \beta_j \leftarrow \frac{\sum_i \widetilde{A}_{ij} (1 - \widetilde{\gamma}_i)(1 - \widetilde{Z}_{ij})}{\sum_i \widetilde{A}_{ij} (1 - \widetilde{\gamma}_i)}.$$

    Clip α_j, β_j to $[\delta, 1 - \delta]$

  Clip $\widetilde{\mu}$ to $[\delta, 1 - \delta]$

until relative change in $\widetilde{\mathcal{L}} < \epsilon = o(N^{-1})$

Store $(\widehat{\alpha}_j, \widehat{\beta}_j)_{j=1}^{m}$

/* Stage B. Maximize $L(\mu)$ with $(\widehat{\alpha}_j, \widehat{\beta}_j)$ */

foreach $i = 1, \ldots, n$ **do**

  $\ell_i(1) \leftarrow \prod_{j : A_{ij} = 1} \widehat{\alpha}_j^{Z_{ij}} (1 - \widehat{\alpha}_j)^{1 - Z_{ij}}, \quad \ell_i(0) \leftarrow \prod_{j : A_{ij} = 1} (1 - \widehat{\beta}_j)^{Z_{ij}} \widehat{\beta}_j^{1 - Z_{ij}}$

$\widehat{\mu}_{\text{HERO-1}} \leftarrow \arg\max_{\mu \in (0,1)} \sum_{i=1}^{n} \log\big[\mu\, \ell_i(1) + (1 - \mu)\, \ell_i(0)\big]$

/* Stage C. Variance reduction via control variates */

if *no shared covariates available* **then return** $\widehat{\mu}_{\text{HERO-1}}$

$\widehat{m}_X \leftarrow N^{-1} \sum_{r=1}^{N} \widetilde{X}_r, \quad \bar{X} \leftarrow n^{-1} \sum_{i=1}^{n} X_i$

foreach $i = 1, \ldots, n$ **do**

  $\widehat{\psi}_i \leftarrow \frac{\ell_i(1) - \ell_i(0)}{\widehat{\mu}_{\text{HERO-1}}\, \ell_i(1) + (1 - \widehat{\mu}_{\text{HERO-1}})\, \ell_i(0)}$

$\widehat{\gamma} \leftarrow \Big[\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top\Big]^{-1} \Big[\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) \widehat{\psi}_i\Big]$

$\widehat{\mu}_{\text{HERO-2}} \leftarrow$ Find in $(0, 1)$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\ell_i(1) - \ell_i(0)}{\mu\, \ell_i(1) + (1 - \mu)\, \ell_i(0)} - \widehat{\gamma}^\top (\bar{X} - \widehat{m}_X) = 0$$

return $\widehat{\mu}_{\text{HERO-2}}$

where $\pi_{\min}$ and $\rho_{\min}$ are positive constants.

Assumption 2 (Correct specification and transportable reliability)**.** *For each labeler j, there exist constants (α_j, β_j) satisfying $c < \alpha_j, \beta_j < 1 - c$ for some $c > 0$ such that the conditional reliabilities are identical in historical and current data:*

$$\Pr(\widetilde{Z}_{ij} = 1 \mid \widetilde{Y}_i = 1) = \Pr(Z_{ij} = 1 \mid Y_i = 1) = \alpha_j,$$
$$\Pr(\widetilde{Z}_{ij} = 0 \mid \widetilde{Y}_i = 0) = \Pr(Z_{ij} = 0 \mid Y_i = 0) = \beta_j,$$

for all instances i and all labelers j. The silver labels are independent across labelers conditional on the gold label:

$$\Pr\left(\widetilde{Z}_{i1} = z_1, \ldots, \widetilde{Z}_{iJ} = z_J \;\middle|\; \widetilde{Y}_i = y\right) = \prod_{j=1}^{m} \Pr(\widetilde{Z}_{ij} = z_j \mid \widetilde{Y}_i = y), \tag{9}$$

$$\Pr\left(Z_{i1} = z_1, \ldots, Z_{iJ} = z_J \;\middle|\; Y_i = y\right) = \prod_{j=1}^{m} \Pr(Z_{ij} = z_j \mid Y_i = y), \tag{10}$$

for all instances i and labelers $j = 1, \ldots, m$. The parameters $\mu = \mathbb{E}[Y]$ and $\widetilde{\mu} = \mathbb{E}[\widetilde{Y}]$ may differ, but are bounded away from 0 and 1:

$$c < \mu, \widetilde{\mu} < 1 - c.$$

Lemma 1 (Convergence of HERO historical reliability estimates)**.** *Under Assumptions 1 and 2, the estimators for α_j and β_j in HERO algorithm satisfy*

$$\widehat{\alpha}_j - \alpha_j = O_p(N^{-1/2}), \qquad \widehat{\beta}_j - \beta_j = O_p(N^{-1/2}), \quad \textit{for each labeler } j.$$

Proof. Let $\mathcal{G} = \{i : \widetilde{R}_i = 1\}$ denote the set of gold-routed historical instances. By the algorithm, the EM is initialized using the gold-only sample-mean estimators

$$\widetilde{\mu}^{(0)} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \widetilde{Y}_i, \quad \widehat{\alpha}_j^{(0)} = \frac{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} \widetilde{Y}_i \widetilde{Z}_{ij}}{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} \widetilde{Y}_i}, \quad \widehat{\beta}_j^{(0)} = \frac{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} (1 - \widetilde{Y}_i)(1 - \widetilde{Z}_{ij})}{\sum_{i \in \mathcal{G}} \widetilde{A}_{ij} (1 - \widetilde{Y}_i)}, \quad j = 1, \ldots, m.$$

Under Assumption 1, the gold-routing indicators $\widetilde{R}_i$ are completely randomized with $P(\widetilde{R}_i = 1) = \rho \geq \rho_{\min} > 0$ and the silver-label availability indicators $\widetilde{A}_{ij}$ are completely randomized with $P(\widetilde{A}_{ij} = 1) \geq \pi_{\min} > 0$, so by the law of large numbers and the delta method, for each j,

$$\widehat{\alpha}_j^{(0)} - \alpha_j = O_p\Big((N\rho\pi_j)^{-1/2}\Big), \qquad \widehat{\beta}_j^{(0)} - \beta_j = O_p\Big((N\rho\pi_j)^{-1/2}\Big), \qquad \widetilde{\mu}^{(0)} - \widetilde{\mu} = O_p\Big((N\rho)^{-1/2}\Big),$$

which implies that the initialization lies within a neighborhood of the gold-oriented population parameter $\eta_0 = (\widetilde{\mu}, \alpha_1, \beta_1, \ldots, \alpha_m, \beta_m)$. Under Assumption 2, the historical data follow the Dawid–Skene model with true gold-oriented parameters $\eta_0 = (\widetilde{\mu}, \alpha_1, \beta_1, \ldots, \alpha_m, \beta_m)$, so that for each i and j the conditional distributions of $\widetilde{Z}_{ij}$ given $\widetilde{Y}_i$ are correctly specified and independent across labelers. The observed-data log-likelihood for the historical dataset is

$$\widetilde{\mathcal{L}}_N(\widetilde{\mu}, \{\alpha_j, \beta_j\}) = \sum_{i=1}^{N} \Big[\widetilde{R}_i \log \widetilde{\ell}_i(\widetilde{Y}_i) + (1 - \widetilde{R}_i) \log\left(\widetilde{\mu}\widetilde{\ell}_i(1) + (1 - \widetilde{\mu})\widetilde{\ell}_i(0)\right)\Big].$$

By construction, this likelihood is twice continuously differentiable in a neighborhood of η_0, and the Fisher information matrix

$$\mathcal{I}_\eta = -E\Big[\frac{\partial^2}{\partial\eta\partial\eta^\top} \log P(\widetilde{Z}_i \mid \widetilde{Y}_i)\Big|_{\eta=\eta_0}\Big]$$

is nonsingular because each labeler contributes a positive amount of information and $\alpha_j, \beta_j \in (c, 1-c)$ for some $c > 0$, ensuring strong identifiability. By standard results on EM for latent-variable models [18, 19], the EM map is monotone in the log-likelihood, i.e.,

$$\widetilde{\mathcal{L}}_N(\eta^{(t+1)}) \geq \widetilde{\mathcal{L}}_N(\eta^{(t)}),$$

and converges to a stationary point of $\widetilde{\mathcal{L}}_N$. Since the initialization $\eta^{(0)}$ is obtained from the gold-only sample mean and satisfies $\|\eta^{(0)} - \eta_0\| = O_p((N\rho\pi_{\min})^{-1/2})$, it lies within the basin of attraction of the gold-oriented local maximizer $\widehat{\eta}_N$, and therefore the EM iterates converge to $\widehat{\eta}_N$. By the stopping criterion $\epsilon = o(N^{-1})$, the EM optimization error satisfies

$$\|\eta^{(T)} - \widehat{\eta}_N\| = o_p(N^{-1/2}),$$

so that it is negligible relative to the statistical variability of the maximizer. Finally, applying standard M-estimation theory for local maximizers of smooth likelihoods [20], we have

$$\widehat{\eta}_N - \eta_0 = O_p(N^{-1/2}),$$

and combining with the EM optimization error yields

$$\widehat{\alpha}_j - \alpha_j = O_p(N^{-1/2}), \quad \widehat{\beta}_j - \beta_j = O_p(N^{-1/2}), \quad \widehat{\mu} - \widetilde{\mu} = O_p(N^{-1/2}), \qquad j = 1, \ldots, m.$$

□

Lemma 2 (EM without gold anchor may fail)**.** *Consider the Dawid–Skene EM applied to historical data or current data without using any gold labels. Then the EM does not necessarily recover the true parameters α and β.*

Proof. Let the observed-data likelihood for the unsupervised DS problem be

$$\widetilde{\mathcal{L}}_N(\mu, \{\alpha_j, \beta_j\}) = \sum_{i=1}^{N} \log\big(\mu\widetilde{\ell}_i(1) + (1-\mu)\widetilde{\ell}_i(0)\big),$$

with

$$\widetilde{\ell}_i(1) = \prod_{j:\widetilde{A}_{ij}=1} \alpha_j^{\widetilde{Z}_{ij}}(1-\alpha_j)^{1-\widetilde{Z}_{ij}}, \quad \widetilde{\ell}_i(0) = \prod_{j:\widetilde{A}_{ij}=1} (1-\beta_j)^{\widetilde{Z}_{ij}}\beta_j^{1-\widetilde{Z}_{ij}}.$$

Consider the second derivative of the likelihood with respect to any α_j:

$$\frac{\partial^2}{\partial\alpha_j^2}\log\big(\mu\widetilde{\ell}_i(1) + (1-\mu)\widetilde{\ell}_i(0)\big) = \frac{-\mu^2\big(\partial_{\alpha_j}\widetilde{\ell}_i(1)\big)^2 + \mu\big(\mu\partial^2_{\alpha_j}\widetilde{\ell}_i(1) + (1-\mu)\partial^2_{\alpha_j}\widetilde{\ell}_i(0)\big)\widetilde{\ell}_i(1)}{(\mu\widetilde{\ell}_i(1) + (1-\mu)\widetilde{\ell}_i(0))^2},$$

which depends on both $\widetilde{\ell}_i(1)$ and $\widetilde{\ell}_i(0)$. Since $\widetilde{\ell}_i(0)$ depends nonlinearly on β_j and $\widetilde{\ell}_i(1)$ depends on α_j, the Hessian is not negative semidefinite in general. Therefore, $\widetilde{\mathcal{L}}_N$ is not concave in (μ, α_j, β_j).

Because EM is a local algorithm, it is only guaranteed to converge to a stationary point of $\widetilde{\mathcal{L}}_N$ [18, 19]. Hence, EM without a gold anchor does not guarantee consistent recovery of the true maxima. □

Theorem 1 (Estimation rate and comparison with baselines)**.** *Under Assumptions 1 and 2, the HERO-1 estimator satisfies*

$$\widehat{\mu}_{\textit{HERO-1}} - \mu = O_p(n^{-1/2}) + O_p(N^{-1/2}).$$

In contrast, the Majority Vote estimator and the classical Dawid–Skene EM estimator without gold anchor do not necessarily converge to the gold-oriented truth.

Proof. Decompose the HERO-1 estimation error as

$$\widehat{\mu}_{\text{HERO-1}} - \mu = (\widehat{\mu}_{\text{HERO-1}} - \mu^*) + (\mu^* - \mu),$$

where

$$\mu^* = \arg\max_{\mu\in(0,1)} \sum_{i=1}^{n} \log\big[\mu\ell_i(1;\alpha_j,\beta_j) + (1-\mu)\ell_i(0;\alpha_j,\beta_j)\big]$$

is the oracle MLE computed using the true labeler parameters (α_j, β_j).

By standard MLE theory [20], the oracle MLE satisfies

$$\mu^* - \mu = O_p(n^{-1/2}),$$

because the log-likelihood for the current data is smooth and concave in μ, and the n instances are independent.

From Lemma 1, the historical HERO EM estimates satisfy

$$\widehat{\alpha}_j - \alpha_j = O_p(N^{-1/2}), \qquad \widehat{\beta}_j - \beta_j = O_p(N^{-1/2}), \quad j = 1, \ldots, m.$$

Since the current-data likelihood is smooth in (α_j, β_j), a first-order Taylor expansion gives

$$\ell_i(1;\widehat{\alpha}_j) - \ell_i(1;\alpha_j) = O_p(N^{-1/2}), \qquad \ell_i(0;\widehat{\beta}_j) - \ell_i(0;\beta_j) = O_p(N^{-1/2}),$$

uniformly in i. Applying the delta method to the MLE equation for $\widehat{\mu}_{\text{HERO-1}}$, we obtain

$$\widehat{\mu}_{\text{HERO-1}} - \mu^* = O_p(N^{-1/2}).$$

Combining the two contributions via the triangle inequality yields

$$\widehat{\mu}_{\text{HERO-1}} - \mu = O_p(n^{-1/2}) + O_p(N^{-1/2}).$$

□

Assumption 3 (Instance sampling and moments)**.** *The current instances $\{i = 1, \dots, n\}$ and historical instances $\{i = 1, \dots, N\}$ are independently and identically distributed from the same population, with sample size $N \gg n$. The HERO-1 score contributions $\psi_i(\mu_0)$ and covariates X_i have finite second moments:*

$$\mathbb{E}[\psi_i(\mu_0)^2] < \infty, \quad \mathbb{E}[X_i^2] < \infty.$$

Theorem 2 (Variance reduction of HERO-2 over HERO-1)**.** *Under Assumption 3, the asymptotic variances of the HERO-1 and HERO-2 estimators satisfy*

$$\mathbb{V}(\widehat{\mu}_{\text{HERO-2}}) = \mathbb{V}(\widehat{\mu}_{\text{HERO-1}}) - \mathcal{I}_0^{-2}\mathbb{C}\text{ov}\{\psi_i(\mu_0), X_i\}^\top \Sigma_X^{-1}\mathbb{C}\text{ov}\{X_i, \psi_i(\mu_0)\},$$

where $\mathcal{I}_0 = -\partial_\mu \mathbb{E}[\psi_i(\mu)]\big|_{\mu=\mu_0} > 0$. Consequently,

$$\mathbb{V}(\widehat{\mu}_{\text{HERO-2}}) \le \mathbb{V}(\widehat{\mu}_{\text{HERO-1}}),$$

with strict reduction whenever $\mathbb{C}\text{ov}\{X_i, \psi_i(\mu_0)\} \neq 0$. Moreover, HERO-2 preserves the bias reduction of HERO-1:

$$\text{Bias}(\widehat{\mu}_{\text{HERO-2}}) = \text{Bias}(\widehat{\mu}_{\text{HERO-1}}) + O(N^{-1/2}) + o(n^{-1/2}).$$

Proof. We first consider the oracle HERO-2 estimator using the true control-variate coefficient γ^* and the population mean m_X. Linearizing the score equation around the true parameter μ_0, we have

$$\widehat{\mu}_{\text{HERO-2}} - \mu_0 = I_0^{-1}\frac{1}{n}\sum_{i=1}^{n}\big(\psi_i(\mu_0) - \gamma^{*\top}(X_i - m_X)\big) + R_n,$$

where $I_0 = -\partial_\mu \mathbb{E}[\psi_i(\mu)]\big|_{\mu=\mu_0} > 0$ and $R_n = o_p(n^{-1/2})$ is the remainder from the Taylor expansion of the score function. By the standard projection property of linear regression, the variance decomposes as

$$\text{Var}\big[\psi_i(\mu_0) - \gamma^{*\top}(X_i - m_X)\big] = \text{Var}[\psi_i(\mu_0)] - \text{Cov}(\psi_i(\mu_0), X_i)^\top \Sigma_X^{-1}\,\text{Cov}(X_i, \psi_i(\mu_0)) \le \text{Var}[\psi_i(\mu_0)],$$

with strict inequality whenever $\text{Cov}(\psi_i(\mu_0), X_i) \neq 0$.

Using Assumption 3, the historical mean $\widehat{m}_X$ satisfies $\widehat{m}_X - m_X = O_p(N^{-1/2})$ and the estimated coefficient $\widehat{\gamma}$ satisfies $\widehat{\gamma} - \gamma^* = o_p(1)$. These contribute only a higher-order term of

$$O_p(N^{-1/2}) + o_p(n^{-1/2})$$

to the asymptotic expansion of $\widehat{\mu}_{\text{HERO-2}}$ since $N \gg n$, which does not affect the first-order variance reduction.

Therefore, replacing m_X by $\widehat{m}_X$ and γ^* by $\widehat{\gamma}$ preserves the variance reduction, and the bias satisfies

$$\text{Bias}(\widehat{\mu}_{\text{HERO-2}}) = \text{Bias}(\widehat{\mu}_{\text{HERO-1}}) + O(N^{-1/2}) + o(n^{-1/2}).$$

This proves that HERO-2 strictly reduces variance relative to HERO-1 while preserving its bias properties. □

Theorem 3 (Asymptotic normality of HERO-1 and HERO-2)**.** *Suppose Assumptions 1–3 hold, then HERO-1 satisfies*

$$\sqrt{n}(\widehat{\mu}_{\text{HERO-1}} - \mu_0) \xrightarrow{d} \mathcal{N}\left(0,\, \mathcal{I}_0^{-2}\mathbb{V}\text{ar}\{\psi_i(\mu_0)\}\right),$$

and HERO-2 satisfies

$$\sqrt{n}(\widehat{\mu}_{\text{HERO-2}} - \mu_0) \rightsquigarrow N\left(0,\, \mathcal{I}_0^{-2}\mathbb{V}\text{ar}\{\psi_i(\mu_0) - \gamma^{*\top}(X_i - m_X)\}\right).$$

Proof. For HERO-1, consider first the idealized MLE using the true labeler reliabilities α_j and β_j. The score equation for μ is smooth and strictly concave, so by the classical M-estimator theory, we have

$$\sqrt{n}\big(\widehat{\mu}_{\text{HERO-1}}(\alpha, \beta) - \mu_0\big) \xrightarrow{d} \mathcal{N}\Big(0, \mathcal{I}_0^{-2}\,\text{Var}\{\psi_i(\mu_0)\}\Big),$$

where $\psi_i(\mu_0)$ is the per-instance score and $\mathcal{I}_0 = -\partial_\mu \mathbb{E}[\psi_i(\mu)]|_{\mu=\mu_0} > 0$.

By Lemma 1, the estimated $\widehat{\alpha}_j, \widehat{\beta}_j$ from the historical data satisfy $\widehat{\alpha}_j - \alpha_j = O_p(N^{-1/2})$, $\widehat{\beta}_j - \beta_j = O_p(N^{-1/2})$. Using Assumption 3 that $N \gg n$, the contribution of the estimation error to the μ-MLE is

$$I_0^{-1} \frac{\partial \psi_i}{\partial(\alpha, \beta)} (\widehat{\alpha} - \alpha, \widehat{\beta} - \beta) = o_p(n^{-1/2}),$$

so $\sqrt{n}(\widehat{\mu}_{\text{HERO-1}} - \mu_0)$ has the same asymptotic distribution as the idealized MLE.

For HERO-2, consider first the oracle estimator using the true control-variate coefficient γ^* and population mean m_X. Linearizing the adjusted score equation, we have

$$\sqrt{n}\big(\widehat{\mu}_{\text{HERO-2}}(\gamma^*, m_X) - \mu_0\big) = \mathcal{I}_0^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\psi_i(\mu_0) - \gamma^{*\top}(X_i - m_X)\big) + o_p(1),$$

which immediately yields the asymptotic normality with variance $\mathcal{I}_0^{-2} \operatorname{Var}[\psi_i(\mu_0) - \gamma^{*\top}(X_i - m_X)]$.

Finally, by the argument in the proof of Theorem 2, replacing γ^* and m_X by their estimators $\widehat{\gamma}$ and $\widehat{m}_X$ contributes only $o_p(1)$ to the $\sqrt{n}$-scale expansion. Hence the asymptotic distribution of HERO-2 is unchanged. This completes the proof. □

D Additional simulation details and results

In this section, we shall provide additional simulation setup details in Section D.1 and additional simulation results in Section D.2.

D.1 Simulation setup additional details.

In Table 3, we shall provide the details of how we define heterogeneity level in Experiment 2. In Table 4, we summarize the simulation parameters used in the main manuscript. In what follows, we shall describe the data generating process. In each Monte Carlo replication, we generate two datasets: a large historical evaluation sample of size N and a smaller current evaluation sample of size n. The two samples share the same prompt-feature distribution and the same pool of silver labelers, but gold labels are observed only for a randomized subset of historical instances. This construction matches the intended deployment setting of HERO, where historical gold labels are used to learn persistent labeler reliability and historical covariates provide a high-precision control-variate anchor. For each instance i, we draw a scalar covariate $X_i \sim \mathcal{N}(0, 1)$, and generate the latent gold label from the logistic model $Y_i \mid X_i \sim \text{Bernoulli}\{p(X_i)\}$, $p(x) = \sigma(\beta_0 + \beta_1 x)$, $\sigma(t) = \frac{1}{1+\exp(-t)}$. The estimand is therefore $\mu_0 := \mathbb{E}_X\{\sigma(\beta_0 + \beta_1 X)\}$. We set $\beta_0 = -0.2$. The slope β_1 controls the correlation between the auxiliary covariate X_i and the latent label Y_i: larger β_1 makes X_i more informative and therefore increases the potential gain from the HERO-2 control-variate adjustment.

Each silver labeler $j \in \{1, \ldots, J\}$ has a sensitivity α_j and specificity β_j, $\alpha_j = \Pr(Z_{ij} = 1 \mid Y_i = 1), \beta_j = \Pr(Z_{ij} = 0 \mid Y_i = 0)$. To induce persistent labeler heterogeneity, we draw $\alpha_j \overset{\text{iid}}{\sim} \text{Beta}\{k\mu_\alpha, k(1-\mu_\alpha)\}, \beta_j \overset{\text{iid}}{\sim} \text{Beta}\{k\mu_\beta, k(1-\mu_\beta)\}$. Throughout Experiment 1 we use $\mu_\alpha = 0.85$, $\mu_\beta = 0.55$, and $k = 8$. Thus, the average silver labeler has high sensitivity but only moderate specificity, creating an asymmetric error pattern under which uncalibrated majority vote is expected to be biased. The concentration parameter k controls across-labeler heterogeneity: larger k yields a more homogeneous labeler pool around the same mean reliability, whereas smaller k yields more dispersed labeler-specific error rates. Each labeler j annotates instance i with probability $A_{ij} \sim \text{Bernoulli}(\pi_j)$, $\pi_j \sim \text{Uniform}(0.2, 0.6)$, independently across instances conditional on π_j. If $A_{ij} = 1$, the observed silver label is generated as $Z_{ij} \mid Y_i = \text{Bernoulli}(\alpha_j)$, if $Y_i = 1$, and $Z_{ij} \mid Y_i = \text{Bernoulli}(1 - \beta_j)$, if $Y_i = 0$. Instances with no observed silver labels are excluded from all estimators; this event is rare in the reported settings and affects all methods equally. The historical and current datasets are generated from the same mechanism, with independent instances but a common labeler pool. For the historical sample, a randomized routing indicator $R_i \sim \text{Bernoulli}(\rho)$ determines whether the latent gold label is revealed. The parameter ρ is the historical gold-label rate. For the current sample, $R_i = 0$ for all instances, so no current gold labels are observed.

Table 3: Heterogeneity levels used in Experiment 2. Smaller k corresponds to greater across-labeler dispersion while preserving the same mean sensitivity and specificity.

Category	k	$\mathbb{E}[\alpha_j]$	$\mathrm{SD}(\alpha_j)$	$\mathbb{E}[\beta_j]$	$\mathrm{SD}(\beta_j)$
Very low	150	0.85	0.029	0.55	0.040
Low	30	0.85	0.064	0.55	0.089
Moderate	12	0.85	0.099	0.55	0.138
High	6	0.85	0.135	0.55	0.188
Extreme	3	0.85	0.179	0.55	0.249

Table 4: Simulation parameters.

Simulation parameters used in the Monte Carlo experiments.

Parameter	Value	Description
n	150	Current instances in Experiment 1 Setting (i)
n	300	Current instances in Experiment 1 Setting (ii) and Experiment 2
N	400	Historical instances in Experiment 1 Setting (i)
N	800	Historical instances in Experiment 1 Setting (ii) and Experiment 2
J	5	Number of labelers in Experiment 1 Setting (i)
J	10	Number of labelers in Experiment 1 Setting (ii) and Experiment 2
β_0	-0.2	Logistic intercept
β_1	1.0	Logistic slope in Experiment 1 Setting (i)
β_1	3.0	Logistic slope in Experiment 1 Setting (ii) and Experiment 2
μ_α	0.85	Mean silver-labeler sensitivity
μ_β	0.55	Mean silver-labeler specificity
k	8	Baseline Beta concentration in Experiment 1
π_j	$\mathrm{Uniform}(0.2, 0.6)$	Labeler-specific annotation probability
ρ	$\{0.01, 0.02, \ldots, 0.10\}$	Historical gold-label rate in Experiment 1 Setting (i)
ρ	$\{0.05, 0.06, \ldots, 0.10\}$	Historical gold-label rate in Experiment 1 Setting (ii)
k	$\{150, 30, 12, 6, 3\}$	Beta concentration levels in Experiment 2
ρ	$\{0.05, 0.10\}$	Historical gold-label rates in Experiment 2
B	200	Monte Carlo replications

D.2 Additional simulation results

In what follows, we show another setting of Experiment 1. In this setting, we use a small current sample and a moderate covariate signal: $n = 150$, $N = 400$, $J = 5$, and $\beta_1 = 1$. We vary the historical gold-label rate over $\rho \in \{0.01, 0.02, \ldots, 0.10\}$.

E Additional details on real-world case studies

1. **Asset defect detection dataset.** This dataset is designed to evaluate generation defects in 3D generation models. It contains data from two versions of the model: the historical dataset consists of outputs generated by the older model version, while the current dataset consists of outputs generated by the newer model version. Both model versions generate 3D assets from the same set of 532 prompts. For each prompt–asset pair, silver labelers evaluate whether the generated asset contains a specific type of defect. The silver labels are binary, with Yes/No responses encoded as 1/0. There are 10 silver labelers in total, and each prompt–asset pair is evaluated by three randomly selected, distinct silver labelers. In addition, a completely random 10% subset of prompt–asset pairs is evaluated by a golden labeler, who is a well-trained expert. Most prompt–asset pairs are also evaluated by an AI labeler, GPT-5, which is prompted to provide the same binary defect label.

   Shared control variate. For the HERO-2 method, we use the AI labels generated by GPT-5 as control variates. We repeat this procedure 100 times and report the absolute bias, standard deviation, and root mean squared error (RMSE) based on the resampled results, following the similar evaluation protocol as in the Simulation Study.

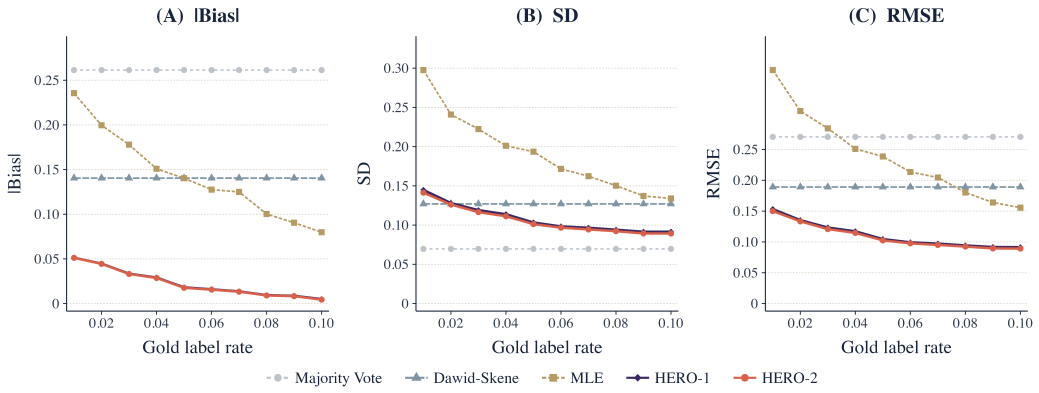


Figure 6: Experiment 1 Setting (ii): Comparison of (A) absolute bias, (B) variance, and (C) MSE with respect to various gold label rate. $\beta_1 = 1$, $J = 5$, $n = 150$, $N = 400$ Effect of historical gold-label rate in Setting (i). Panels report (A) absolute bias, (B) Monte Carlo standard deviation, and (C) RMSE for estimating $\mu_0 = \mathbb{E}[Y_i]$. The current sample is small and the covariate signal is moderate: $\beta_1 = 1$, $J = 5$, $n = 150$, and $N = 400$.

2. **Safety evaluation dataset.** This dataset contains both a historical dataset and a current dataset for evaluating the safety of a 3D generation model with respect to policy violation categories. The historical dataset corresponds to evaluations along one safety dimension, whereas the current dataset corresponds to evaluations along another safety dimension. The model generates 3D assets from 2,046 prompts. For each prompt–asset pair, silver labelers provide binary safety labels along the corresponding safety dimension. There are 11 silver labelers in total, and each prompt–asset pair is evaluated by a randomly selected set of 4–7 distinct silver labelers. In addition, a completely random 80% subset of prompt–asset pairs is evaluated by a golden labeler, who is a well-trained expert, to benchmark the performance of the silver labelers.

   Shared control variate. For HERO-2, we use TF-IDF features extracted from the prompts as control variates. Since TF-IDF (Term Frequency-Inverse Document Frequency) features are typically high-dimensional relative to the sample size, we apply a ridge penalty when estimating the coefficient γ in the score-based control-variate adjustment. As in the previous case study, we report the absolute bias, standard deviation, and RMSE for HERO and the baseline methods.

F Comparison with existing literature

In this section, we compare HERO with three related strands of work: learning from noisy annotators, variance reduction techniques, and generative AI evaluation.

Learning from crowds and noisy annotators has a long history in statistics and machine learning. The classical Dawid–Skene (DS) model [9] introduced an EM-based approach to estimate individual annotator error rates from noisy labels. Subsequent work extended this idea to more complex scenarios, including latent variable models for multi-dimensional annotation [21], probabilistic truth inference [10], and crowdsourced aggregation with worker reliability [11, 22]. Most of these studies assume that all labelers are noisy, and they do not exploit the existence of a small set of high-quality gold labels for calibration, leaving a gap for methods that systematically combine historical gold-labeled subset with current noisy annotations.

Variance reduction has been studied extensively in statistics and experimental design. Classical tools include control variates and regression adjustment, where auxiliary information is used to reduce estimation variance without bias [23, 24]. In applied fields, [25] uses pre-experiment measurements to improve sensitivity in controlled experiments, and Monte Carlo methods with control variates have a rich theoretical literature [26]. These techniques reduce variance by exploiting correlated auxiliary variables, but, to our knowledge, they have not been systematically adapted to the noisy-label setting.

Evaluation of generative AI raises additional challenges. Human preference judgments and pairwise comparison metrics have been used in practice to assess LLM quality [1, 27], and broad benchmarking frameworks such as HELM [28] aggregate diverse evaluation metrics. Safety and robustness evaluations, such as toxicity benchmarks, further illustrate the complexity of human annotations [29, 30]. However, there is not yet a fully statistical framework that jointly accounts for annotator heterogeneity, leverages historical gold supervision, and reduces variance in large-scale AI model evaluation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see Section 1

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please see Section 3.4 and Appendix Section C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see Section 4, Section 5 and Appendix Section D and E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code and detailed instructions for reproducing the simulation studies and all synthetic-data experiments reported in the paper. The real-world evaluation datasets are proprietary and cannot be publicly released at this stage; however, we describe the data structure, preprocessing steps, and experimental protocol in sufficient detail to support reproducibility on comparable datasets.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see Section 4, Section 5 and Appendix Section D and E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see Section 4 and 5

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix Section D and E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the potential positive social impacts in Section 1. The negative social impacts are not applicable.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly credited the original data source.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper uses pre-collected annotation data and does not conduct new crowdsourcing experiments or new human-subject data collection. Therefore, participant instructions, screenshots, and compensation details are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable, because this paper analyzes pre-collected annotation data and does not involve new interaction with study participants or new human-subject data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Although the paper studies generative model evaluation, LLMs are not used as an important, original, or non-standard component of the proposed method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.