

Collaborative Adaptive Labeling with Imperfect Labelers and Selective Expert Escalation

Anonymous Authors¹

Abstract

High-quality labels are essential for training, evaluation, benchmarking, and the reliable deployment of modern machine learning (ML) and artificial intelligence (AI) systems. Increasingly, labeling pipelines in practice involve collaborative “gold-silver” supervision, where all instances may receive multiple inexpensive, imperfect *silver labels* (e.g., from crowdsourcing platforms or automated AI judges), while a limited number of costly *gold labels* provided by experts can be selectively acquired for difficult cases (such as those that have many silver labels that differ). This setting differs from classical labeling formulations in that an instance can receive multiple silver labels, while expert labeling is applied selectively, and has become increasingly common in the evaluation of modern ML/AI systems.

Yet, a key challenge in this setup is determining when and how to allocate labeling effort across silver and gold labels under a fixed budget, while simultaneously supporting reliable evaluation of ML/AI systems. Escalating from silver to expert labeling too late may propagate incorrect labels from imperfect annotators, whereas escalating too early wastes scarce expert resources. Moreover, because labeling decisions depend on previously observed labels, the resulting data are adaptively sampled, inducing dependencies between labels and the sampling process. This adaptivity complicates both the design of systematic labeling algorithms and the validity of downstream statistical inference used for ML/AI system evaluation. To address these challenges, we propose a *cost-efficient collaborative adaptive labeling framework* in which each instance may receive multiple imperfect silver labels and, when war-

ranted, an expert-provided gold label. To support valid downstream statistical evaluation of ML/AI systems, we propose an estimator that systematically combines expert-provided gold labels and imperfect silver labels collected under adaptive labelling, and establishes their consistency under mild conditions. Across multiple benchmarking datasets, our method substantially improves labeling quality and the reliability of downstream statistical evaluation compared to existing baselines.

1. Introduction

1.1. Motivation and Our Contributions

Modern ML and AI systems increasingly rely on large volumes of high-quality labels for downstream tasks such as fine-tuning, benchmarking, and performance evaluation. While expert annotations remain the gold standard, they are costly and limited in scale. At the same time, there is growing access to inexpensive but imperfect labelers that, in many tasks, can produce labels comparable to expert quality. These imperfect sources range from crowdsourced human workers on online platforms to automated evaluation systems such as “LLM-as-a-judge.” This shift naturally gives rise to a general problem that we refer to as *collaborative labeling*: how can we effectively combine label sources of different quality and cost to achieve high labeling accuracy under a limited budget? Recent works have explored related ideas under the themes of human–AI collaboration [cite], learning to defer [cite], and language model routing [cite].

Despite recent progress, the application of collaborative labeling for ML/AI is limited by two key practical and theoretical challenges.

First, most existing collaborative labeling frameworks assume that each instance ultimately receives a single label. In practice, however, an instance may receive multiple intermediate labels from inexpensive sources, which can often be combined to produce a higher-quality signal before deciding whether expert annotation is necessary. This substantially complicates the labeling decision, as we must determine not only *which* source to query, but also *how many* labels to

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

acquire from each source.

Second, there exists a systematic gap between collaborative labeling and valid post-hoc statistical evaluation using the collected labels. Because labeling decisions typically depend on previously observed labels and model updates, the resulting data collection process is adaptive and history-dependent. This violates the independence assumptions underlying standard statistical methods designed for i.i.d. datasets. A principled framework for multi-source labeling that also guarantees valid post-hoc statistical inference remains largely unexplored.

To address these two limitations,

1.2. Related Literature

1. *Adaptive label acquisition.*
2. *Learning with noisy labels.*
3. *Recent advances in statistical inference: adaptive sampling and predicted outcomes.*

2. Problem Setup

In this section, we introduce the problem setup for collaborative adaptive labeling, as well as the statistical target for downstream tasks for the evaluation ML/AI systems based on adaptively collected labels.

Let $X \in \mathcal{X}$ denote the information of an instance to be labeled, such as a pair consisting of a prompt P and an image I generated from an AI model. Let $Y \in \{0, 1\}$ denote the gold label of the instance. We assume binary labels in the main text for illustration and defer the multiclass extension to the appendix. Due to the high cost of expert labeling, the gold label is *not observed* for all instances. We use $D \in \{0, 1\}$ to denote whether the gold label is observed for an instance ($D = 1$ if observed, and $D = 0$ otherwise).

In addition to the gold label, we (may) observe *multiple* silver labels for each instance. We first consider the setting where each silver label is an anonymous draw from a population of crowd workers, which reflects the common situation where silver labels are collected from third-party contractors. In this case, the silver labels for an instance are interchangeable, and we denote them as a sequence

$$\mathbf{Z} \in \{0, 1\}^m,$$

where the length $m \in \mathbb{Z}_{\geq 0}$ is the number of silver labels, which may differ across instances. In the Appendix, we extend this to a *channel-aware* setting, where silver labels may come from different channels with heterogeneous labeling criteria, such as from different countries.

We now describe the data structure, which resembles the “pool-based” setting in active learning. Suppose we observe

a historical dataset consisting of N instances with multiple silver labels and possibly a gold label, denoted by

$$\mathcal{D} = \{(X_i, \mathbf{Z}_i, D_i, D_i Y_i)\}_{i=1}^N.$$

We then receive a new dataset of n unlabeled instances, denoted by

$$\mathcal{D}^* = \{(X_i)\}_{i=N+1}^{N+n}.$$

In this paper, we pursue two objectives. Our *first objective* is to obtain high-quality labels for the unlabeled dataset $\mathcal{D}^*$ via a *collaborative adaptive labeling* (CAL) algorithm. By *adaptive*, we mean that labeling decisions are made sequentially over T stages, where at each stage the algorithm may take actions on multiple instances based on all labels observed so far. By *collaborative*, we mean that the algorithm jointly leverages inexpensive, imperfect silver labelers and costly expert (gold) labelers, allocating labeling effort across them in a coordinated manner. At each stage and for each currently unlabeled instance, the algorithm selects an action from $\{\text{GOLD}, \text{SILVER}, \text{NONE}\}$. The action GOLD escalates the instance to an expert to obtain a gold label, SILVER queries one additional silver label, and NONE defers labeling at the current stage. Let $\hat{Y}_i$ denote the final label assigned to instance i after T stages. Querying a silver label incurs a cost c_s , while obtaining a gold (expert) label incurs a higher cost c_g . The goal of CAL is to design a labeling policy that minimizes overall labeling error under a total budget constraint B .

Our *second objective* is to perform a statistically valid evaluation of an ML/AI system that is aligned with expert (gold) labelers, using all labels—both silver and gold—collected under collaborative adaptive labeling. In practice, expert-provided gold labels define the evaluation metric of interest but are observed for only a small, adaptively selected subset of instances, while inexpensive silver labels are abundant yet potentially biased and noisy. Naively averaging silver labels can therefore lead to biased evaluation, whereas relying exclusively on gold labels can be statistically inefficient. Accordingly, our evaluation target is

$$\mu := \mathbb{E}_{X \sim P_X^*}[Y],$$

where P_X^* denotes the instance distribution underlying the evaluation pool $\mathcal{D}^*$. This quantity represents the expected gold-label evaluation score over the population of instances. Our goal is to construct efficient estimators and valid confidence intervals for μ by leveraging adaptively collected gold labels together with noisy silver labels, thereby enabling statistically sound comparison of ML/AI systems under a fixed labeling budget. Extensions to more general empirical risk minimization targets are deferred to the appendix.

3. Collaborative Adaptive Labeling

3.1. Learning from Historical Data

What does CAL learn from historical data? CAL leverages the historical dataset $\mathcal{D}$ to initialize two models that guide later adaptive labeling on the new pool $\mathcal{D}^*$. The first is a *gold prediction model* (G-model),

$$p_0(X) = \hat{\mathbb{P}}(Y = 1 | X),$$

trained using historical instances with observed gold labels (i.e., $D_i = 1$). The second is an *item difficulty model* (D-model),

$$e_0(X) = \hat{\mathbb{P}}(Z \neq Y | X),$$

which estimates the instance-specific error rate of an individual silver label Z relative to the gold label Y . The D-model is learned from historical instances for which both gold labels and (possibly multiple) silver labels are available. Intuitively, $p_0(X)$ provides a *prior belief* about the gold label, while $e_0(X)$ determines how informative each additional silver label should be for that instance.

A simple way to fit the D-model is via a binomial likelihood. For a historical instance with gold label Y_i and m_i silver labels, let $r_i = \sum_{j=1}^{m_i} \mathbf{1}\{Z_{i,j} \neq Y_i\}$ denote the number of silver-gold disagreements. Then one may model

$$r_i | X_i \sim \text{Binomial}(m_i, e_0(X_i)),$$

enabling flexible regression (e.g., a lightweight head on embeddings) to learn instance-dependent difficulty. We defer full implementation details to Section 3.3.

How will these models used on $\mathcal{D}^*$? On the evaluation pool $\mathcal{D}^*$, CAL maintains for each instance i a stage-dependent belief

$$p_t(X_i) \approx \hat{\mathbb{P}}(Y_i = 1 | X_i, \mathbf{Z}_{i,t}),$$

where $\mathbf{Z}_{i,t}$ denotes the silver labels collected for instance i up to stage t . This belief is obtained by combining the G-model prior $p_0(X_i)$ with the observed silver labels through a difficulty-aware update governed by the D-model, so that each silver label is weighted according to its estimated reliability. CAL then converts the updated belief into an *uncertainty score*

$$s_t(X_i) = f(p_t(X_i)) \in [0, 1],$$

where $f(\cdot)$ can be entropy, margin, or variance. As a result, the uncertainty score reflects information from both instance features (via the G-model) and observed labeling disagreement (via the D-model-weighted silver labels). This score serves as the *primary signal* for routing labeling effort: instances that remain uncertain are prioritized for additional

labeling, either by querying more silver labels or by escalating to expert annotation.

In a greedy labeling strategy, gold labels are assigned to the instances with the largest scores until the stage budget is exhausted. Alternatively, for probabilistic sampling, these scores are converted into softmax sampling weights

$$w_{i,t} = \frac{\exp(s_{t-1}(X_i)/\tau)}{\sum_j \exp(s_{t-1}(X_j)/\tau)},$$

where τ is the temperature parameter controlling the randomness of the sampling. Given stage budget B_t , the sampling probability for querying a gold label is

$$q_{i,t} = \min\left(1, \frac{B_t}{c_g} w_{i,t}\right).$$

A gold label is queried according to $\delta_{i,t} \sim \text{Bernoulli}(q_{i,t})$. Once an instance receives a gold label, it is removed from the unlabeled pool. After collecting new gold labels, the model is updated from $p_{t-1}(x)$ to obtain $p_t(x)$.

3.2. CAL collects silver labels early

A key feature of CAL is that it allocates a larger share of its early-stage labeling budget to querying silver labels, using them as a low-cost mechanism for screening before selectively escalating to gold. Before presenting the proposed algorithm, we believe it might be helpful to intuitively understand why CAL naturally prioritizes early silver labeling for two complementary reasons. See Figure 6 for an illustration.

First, early silver labels provide cheap, instance-specific evidence that corrects miscalibrated priors and sharpens uncertainty estimates for routing. Concretely, on $\mathcal{D}^*$ we initially observe only X_i and hence start from the G-model prior $p_0(X_i) = \hat{\mathbb{P}}(Y_i = 1 | X_i)$ together with an estimated silver error rate $\hat{e}_0(X_i) = e_0(X_i)$. After collecting $m_{i,t}$ silver labels $\mathbf{Z}_{i,t}$ with $k_{i,t}$ ones, CAL forms an updated belief

$$p_t(X_i) \approx \hat{\mathbb{P}}(Y_i = 1 | X_i, \mathbf{Z}_{i,t}),$$

which combines the G-model prior with D-model-weighted silver evidence (Section 3.3.2). CAL then computes an uncertainty score $s_t(X_i) = f(p_t(X_i))$; in particular, early disagreement between a newly collected silver label and the current prediction tends to move $p_t(X_i)$ toward 0.5 and increase $s_t(X_i)$, flagging the instance as potentially difficult and improving downstream routing.

Under limited labeling resources and a pretrained prediction model $p_0(x)$, the Bayes-optimal strategy is to prioritize instances where the model is likely to be wrong, characterized by the *inaccuracy score* $W_i := \mathbf{1}\{\hat{Y}_{i,0} \neq Y_i\}$. Since the gold label Y_i is unobserved, the practical surrogate used in Section ?? is an *uncertainty score* s_i and instances with

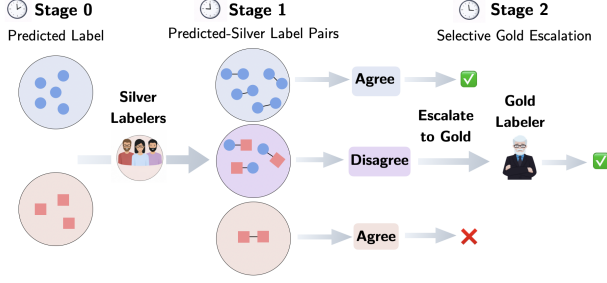


Figure 1. Caption

the largest s_i are labeled. However, the ranking induced by $\{s_i\}$ can be poorly aligned with the true ranking of $\{W_i\}$, causing standard uncertainty sampling to miss instances that genuinely require labeling.

Silver labels provide a cheap diagnostic signal for this misalignment. In particular, if a silver label disagrees with the model prediction, then even when s_i is small, there is a high probability that $W_i = 1$. Assigning higher score to such disagreement cases yields a sampling score that better approximates the true inaccuracy score W_i , leading to improved labeling efficiency once escalated to gold. A similar argument applies to agreement cases.

Second, the predictive value of additional silver labels typically exhibits *diminishing returns*, so an expected-uncertainty-reduction policy naturally favors collecting silver labels early for broad screening and reserving scarce gold labels for later stages when only genuinely hard instances remain uncertain. This decay can be seen directly from the posterior update: under the symmetric noise model with error rate $\hat{e}_0(X_i)$, each additional silver label shifts the posterior *log-odds* by a constant amount,

$$\text{logit}(p_t(X_i)) = \text{logit}(p_0(X_i)) + (2k_{i,t} - m_{i,t}) \log \frac{1 - \hat{e}_0(X_i)}{\hat{e}_0(X_i)},$$

so once $p_t(X_i)$ is already close to 0 or 1, the same $\pm \log \frac{1 - \hat{e}_0(X_i)}{\hat{e}_0(X_i)}$ increment induces only a small change in $p_t(X_i)$ (and hence in $s_t(X_i)$), analogous to the diminishing marginal gains of majority voting. Accordingly, the one-step expected uncertainty reduction from querying an additional silver label,

$$u_{i,t,\text{SILVER}} = \frac{s_t(X_i) - \hat{s}_{t+1}(X_i)}{c_s},$$

tends to decrease as $m_{i,t}$ grows, whereas escalating to gold removes uncertainty entirely with per-cost gain $u_{i,t,\text{GOLD}} = s_t(X_i)/c_g$ (Section 3.3). This structure encourages querying more silver early (cheap screening across many instances) and shifting toward gold later for the remaining uncertain instances.

3.3. Collaborative Adaptive Labeling

In this section, we present the proposed labeling framework. Before describing the full algorithm, we introduce its key building blocks.

3.3.1. BASE MODELS

The framework relies on two base models that support the diagnostic and prediction roles of silver labels. The first is gold label prediction model (*G-model*) $p(X; \theta) = \mathbb{P}(Y = 1|X)$, which predicts the gold label using only instance information X , without incorporating any silver label information. It plays the similar role as the predictor used in standard uncertainty sampling (Section ??).

The second is the item-level difficulty model (*D-model*). For each instance, we introduce a latent item difficulty random variable $e_i \in [0, 1]$, interpreted as the probability that a silver label Z disagrees with the corresponding gold label. To capture heterogeneity across instances, we model $\mathbb{P}[Z \neq Y|X] = e(X; \phi)$. We assume $e_i \leq \bar{e} < 0.5$ almost surely, since otherwise the label can be flipped.

Both models are trained on the historical dataset $\mathcal{D}$, yielding initial estimates $\hat{\theta}_0$ and $\hat{\phi}_0$. The framework does not impose restrictions on model classes, and these models can be implemented using any suitable architecture, such as lightweight heads on top of pre-trained embeddings.

3.3.2. SILVER-GUIDED PREDICTION UPDATE

We next describe the prediction update rule after collecting silver labels in the T -stage sequential framework. For instance i , suppose that up to stage $t \leq T - 1$ we have observed a sequence of silver labels $\mathbf{Z}_{i,t}$ of length $m_{i,t}$, among which $k_{i,t}$ labels take value 1. Combining information from the G-model and the D-model, we refine the prediction of the gold label by

$$p_t(X_i) \propto p(X_i; \hat{\theta}) (1 - e(X_i; \hat{\phi}))^{k_{i,t}} e(X_i; \hat{\phi})^{m_{i,t} - k_{i,t}}.$$

In log-odds form, this update can be interpreted as a difficulty-aware weighted majority vote between the G-model prediction and the collected silver labels, where each silver label contributes a weight $\log \frac{1 - e(X_i; \hat{\phi})}{e(X_i; \hat{\phi})}$ determined by the estimated item difficulty.

As in Section ??, we compute an uncertainty score $s_t(X_i) = f(p_t(X_i))$ from the updated prediction. For standard uncertainty measures that attain their maximum at $p_t(X_i) = 0.5$, observing a new silver label that disagrees with the current prediction moves $p_t(X_i)$ toward 0.5 and thus increases the uncertainty score, while agreement decreases it. This behavior naturally reflects the diagnostic role of silver labels.

3.3.3. ONE-STAGE-AHEAD SAMPLING SCORE

We now define the sampling score for each action $\{\text{GOLD}, \text{SILVER}, \text{NONE}\}$ used in adaptive labeling. To evaluate the benefit of querying an additional silver label, we compute the one-stage-ahead expected uncertainty score

$$\begin{aligned}\hat{s}_{t+1}(X_i) &= \hat{\mathbb{E}}[f(p_{t+1}(X_i)) \mid X_i, \mathbf{Z}_{i,t}] \\ &= \sum_{z \in \{0,1\}} \hat{\mathbb{P}}(Z_{i,t+1} = z \mid X_i, \mathbf{Z}_{i,t}) f(p_{t+1}^{(z)}(X_i)),\end{aligned}$$

where $p_{t+1}^{(z)}(X_i)$ is the updated prediction after observing a new silver label $z \in \{0, 1\}$. The probability of the next silver label depends on the current prediction and the estimated difficulty $\hat{e}_i = e(X_i; \hat{\phi})$:

$$\hat{\mathbb{P}}(Z_{i,t+1} = 1 \mid X_i, \mathbf{Z}_{i,t}) = \hat{e}_i + (1 - 2\hat{e}_i) p_t(X_i),$$

$$\hat{\mathbb{P}}(Z_{i,t+1} = 0 \mid X_i, \mathbf{Z}_{i,t}) = (1 - \hat{e}_i) - (1 - 2\hat{e}_i) p_t(X_i).$$

The sampling score for SILVER is defined as the expected uncertainty reduction per unit cost c_s :

$$u_{i,t,\text{SILVER}} = (s_t(X_i) - \hat{s}_{t+1}(X_i)) / c_s.$$

Although querying a silver label may have multi-stage effects (e.g., increasing uncertainty and triggering future escalation to gold), this long-term impact is difficult to quantify. We therefore adopt a one-stage-ahead approximation.

For GOLD, the uncertainty after labeling becomes zero, so the one-stage uncertainty reduction equals $s_t(X_i)$. Its sampling score is

$$u_{i,t,\text{GOLD}} = s_t(X_i) / c_g.$$

3.3.4. TWO-STEP ADAPTIVE SAMPLING

We now describe a two-step hierarchical probabilistic sampling rule for selecting stage $t+1$ actions while respecting the stage budget B_{t+1} . A greedy version can be obtained by replacing the softmax steps with maximization. For each instance, define the action preference

$$q_{i,t,G} = \frac{\exp(u_{i,t,\text{GOLD}}/\tau)}{\exp(u_{i,t,\text{GOLD}}/\tau) + \exp(u_{i,t,\text{SILVER}}/\tau)},$$

and $q_{i,t,S} = 1 - q_{i,t,G}$. The expected utility and cost of acting on instance i are therefore

$$\tilde{u}_{i,t+1} = q_{i,t,G} u_{i,t+1,\text{GOLD}} + q_{i,t,S} u_{i,t+1,\text{SILVER}},$$

$$\tilde{c}_{i,t+1} = q_{i,t,G} c_g + q_{i,t,S} c_s.$$

Compute instance weights

$$\tilde{w}_{i,t+1} = \frac{\exp(\tilde{u}_{i,t+1}/\tau)}{\sum_j \exp(\tilde{u}_{j,t+1}/\tau)},$$

and scale them by

$$\alpha_{t+1} = \frac{B_t}{\sum_i \tilde{w}_{i,t+1} \tilde{c}_{i,t+1}}.$$

Each instance is independently selected for labeling with probability $\alpha_{t+1} \tilde{w}_{i,t+1}$. Conditional on being selected, GOLD is chosen with probability $q_{i,t,G}$, and SILVER otherwise. Similar to Section ??, once an instance receives a gold label, it is removed from the unlabeled pool, and the base models are updated using the data collected in the current stage. A complete pseudocode description is provided in the appendix.

4. Statistical Post-Hoc Evaluation

In this section, we present methods for the second objective, using the labels collected from the Section 3.3. We first describe the statistical challenges in estimating $\mu = \mathbb{E}_{X \sim P_X^*}[Y]$ in Section 4.1, and then introduce our proposed estimators in Section 4.2.

4.1. Statistical Challenges

Consistent and efficient estimation of the target μ faces the following challenges:

1. *Adaptive sampling with early stopping.* As described in Section 3.3, the probability that an instance receives a gold label depends on the full history of the labeling algorithm, including past labels and model updates. Moreover, once a gold label is collected, the instance leaves the unlabeled set and will never be labeled again. This creates a history-dependent sampling process with early stopping, which invalidates naive sample averages and is not addressed by standard i.i.d. assumptions or classical inverse propensity score weighting (IPW) formulations.
2. *Noisy silver labels.* Instances that do not receive gold labels typically obtain multiple silver labels that contain useful information about Y . However, these labels are noisy, and directly using them in estimation can introduce bias. A key challenge is how to leverage silver-label information to improve estimation efficiency while preserving consistency and valid inference.

4.2. Proposed Estimators

We now introduce two estimators for μ with strong statistical guarantees, whose properties are established in Section 5. Notably, our method allows the instance distribution in the historical dataset to differ from that in the unlabeled evaluation dataset; that is, $P_X \neq P_X^*$.

We first introduce additional notation for the label collection procedure based on the framework in Section 3.3. Define $\delta_{i,t} \in \{0, 1\}$ as the indicator of whether a gold label is collected at stage t for instance i . Let the information collected up to stage $t - 1$ be denoted by $\mathcal{F}_{t-1}$. According to the algorithm in Section 3.3, each $\delta_{i,t}$ equals 1 with known sampling probability

$$\pi_{i,t} := \mathbb{P}(\delta_{i,t} = 1 \mid \mathcal{F}_{t-1}) = \alpha_{t+1} \tilde{w}_{i,t+1} q_{i,t,G} \in (0, 1),$$

which depends on past information. We further define the stage at which a gold label is collected as τ_i , such that $1 \leq \tau_i \leq T$ if a gold label is collected within T stages, and $\tau_i = \infty$ otherwise. Once a gold label is collected, the instance is removed from the unlabeled pool, and the probability of collecting a gold label becomes 0, creating an early stopping mechanism that standard IPW does not directly handle.

To obtain an unbiased estimator of the target, we propose an adjusted IPW estimator, denoted by $\hat{\mu}_{\text{IPW,CAL}}$. Define

$$D_i = \frac{1}{T} \left(\sum_{t \leq \min\{\tau_i, T\}} \frac{\delta_{i,t}}{\pi_{i,t}} + \underbrace{(T - \tau_i)_+}_{\text{Early stopping adjustment}} \right).$$

The adjusted IPW estimator is then given by

$$\hat{\mu}_{\text{IPW,CAL}} = \frac{1}{n} \sum_{i=N+1}^{N+n} D_i Y_i.$$

Compared to standard IPW, this adjusted estimator explicitly accounts for the early stopping mechanism by incorporating the deterministic contribution of post-gold stages, thereby avoiding invalid weights and preserving unbiasedness under the adaptive sampling process. We can estimate the variance by

$$\hat{\mathbb{V}}_{\text{IPW,CAL}} = \frac{1}{n} \sum_{i=N+1}^{N+n} (D_i Y_i - \hat{\mu}_{\text{IPW,CAL}})^2,$$

and construct an asymptotic $(1 - \alpha)$ confidence interval as

$$\left[\hat{\mu}_{\text{IPW,CAL}} \pm z_{1-\alpha/2} \sqrt{\hat{\mathbb{V}}_{\text{IPW,CAL}}/n} \right].$$

To leverage the collected silver labels to further improve asymptotic estimation efficiency without introducing bias, we propose an adjusted augmented IPW estimator:

$$\hat{\mu}_{\text{AIPW,CAL}} = \frac{1}{n} \sum_{i=N+1}^{N+n} \psi_i, \quad \hat{\mathbb{V}}_{\text{AIPW,CAL}} = \frac{1}{n} \sum_{i=N+1}^{N+n} (\psi_i - \bar{\psi})^2,$$

where

$$\psi_i = D_i Y_i + \underbrace{\frac{1}{T} \sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) p_{t-1}(X_i)}_{\text{Sequential prediction augmentation}}.$$

Here, $p_{t-1}(X_i)$ denotes the silver-guided gold prediction introduced in Section 3.3.2. Notably, this construction respects the sequential nature of the adaptive labeling procedure: at each stage, the augmented term is paired with the prediction model available at that time, without using any future information. The theoretical guarantees of both estimators are established in Section 5.

5. Theoretical Investigation

In this section, we provide theoretical support for the proposed methods in Sections 3 and 4. In Section 5.1, we formalize the diagnostic role of silver labels and selective gold escalation, providing a theoretical justification for the intuition in Section ???. In Section 5.2, we establish finite-sample and asymptotic properties of the estimators proposed in Section 4.2.

5.1. Diagnostic Role of Silver Labels

While a full analysis of the CAL framework is beyond the scope of this paper, we show in a simplified setting when the diagnostic role of silver labels (Section ??) can improve labeling quality over standard uncertainty sampling.

We first present a lemma showing that a sampling score more aligned with true labeling mistakes leads to higher labeling accuracy. Consider greedy uncertainty sampling with gold labels only (Section ??) under a single stage ($T = 1$), without model updating, and suppose the labeling budget is a proportion $b \in (0, 1)$. Recall the *inaccuracy score* $W_i := \mathbf{1}\{\hat{Y}_{i,0} \neq Y_i\}$ defined in Section ??, where $\hat{Y}_{i,0}$ is the predicted label from $p_0(x)$.

For any sampling score $U_i \in \mathbb{R}$ (larger values indicate higher priority for labeling), define its alignment with the true inaccuracy score via the AUC: $\text{AUC}(U, W) = \mathbb{P}(U^+ > U^-) + \frac{1}{2} \mathbb{P}(U^+ = U^-)$, where U^+ is drawn from $U \mid (W = 1)$ and U^- is drawn from $U \mid (W = 0)$. Let $S_b(U)$ denote the top- b fraction of instances ranked by U , and define the labeling gain as $G_b(U) := \mathbb{P}(W = 1 \mid i \in S_b(U))$. This quantity directly determines the improvement in final labeling accuracy, since gold labeling corrects exactly those mistakes.

Lemma 5.1 (Score alignment and accuracy). *Let S and U be two sampling scores. If $\text{AUC}(S, W) > \text{AUC}(U, W)$, then, averaged over all labeling budgets $b \in (0, 1)$,*

$$\int_0^1 G_b(S) db > \int_0^1 G_b(U) db.$$

In other words, a sampling score that is more aligned with the inaccuracy score (in the AUC sense) leads to higher expected labeling accuracy.

We next show that querying silver labels can produce an

updated sampling score that is better aligned with the inaccuracy score under mild conditions. Suppose each instance receives a silver label $S_i \in \{0, 1\}$ with error rate $e_i := \mathbb{P}(S_i \neq Y_i \mid X_i)$. We consider a general sampling score update rule after observing silver, which accommodates but goes beyond the update rule in Section 3.3.2.

Define the agreement indicator $M_i := \mathbf{1}\{S_i = \hat{Y}_i\}$, where $\hat{Y}_i$ is the predicted label. Let the baseline sampling score be U_i . We define the updated sampling score as an additive perturbation of the baseline score $T_i^{\text{sil}} := U_i + \Delta_i$, where the update term Δ_i depends on the agreement indicator and satisfies

$$\Delta_i = \begin{cases} \delta_i^+, & M_i = 0 \quad (\text{disagreement}), \\ -\delta_i^-, & M_i = 1 \quad (\text{agreement}), \end{cases}$$

for some nonnegative quantities $\delta_i^+, \delta_i^- \geq 0$, with $\delta_i^+ + \delta_i^- > 0$. That is, disagreement between the silver label and the model prediction increases the sampling score relative to agreement. We impose the following assumptions.

Assumption 5.2 (Independent silver noise). $S_i \perp \hat{Y}_i \mid (Y_i, X_i)$.

Assumption 5.3 (Silver quality). There exists $\bar{e} < \frac{1}{2}$ such that $e_i \leq \bar{e}$ for all i .

Assumption 5.2 states that the silver label provides an independent noisy observation of the ground truth relative to the model prediction, rather than being derived from the model itself. Assumption 5.3 requires that the silver labeler performs uniformly better than random guessing.

Proposition 5.4 (Silver disagreement improves score alignment). *Under Assumptions 5.2–5.3, the updated score T^{sil} satisfies*

$$\text{AUC}(T^{\text{sil}}, W) > \text{AUC}(U, W),$$

unless $\text{AUC}(U, W) = 1$.

Proposition 5.4 shows that the benefit of silver labels does not rely on an exact posterior update. Any update rule that increases the score upon disagreement injects an independent signal that perturbs the ranking in the correct direction, and is *robust* to misspecification of the D-model. Combined with Lemma 5.1, this implies that silver-based score updates lead to higher labeling accuracy on average across labeling budgets whenever the baseline uncertainty score is not already an optimal detector of its own mistakes.

5.2. Statistical Properties of Post-Hoc Estimators

A key advantage of the proposed CAL framework is that, despite the adaptive labeling process with early stopping, valid post-hoc statistical inference remains possible for evaluation. We establish finite-sample unbiasedness together with

asymptotic consistency and normality for the estimators introduced in Section 4.2. Notably, these guarantees hold for the *softmax probabilistic sampling* version of CAL, but not for the greedy version. The probabilistic design ensures that, before stopping, the probability of acquiring a gold label for any instance is bounded away from 0, which corresponds to the overlap condition in missing data problems.

Proposition 5.5 (Adjusted IPW estimator). *Under the adaptive labeling design in Section 3.3, for any finite number of stages T , the estimator $\hat{\mu}_{\text{IPW,CAL}}$ is unbiased and consistent for μ . Moreover, $\frac{\sqrt{n}(\hat{\mu}_{\text{IPW,CAL}} - \mu)}{\sqrt{\hat{V}_{\text{IPW,CAL}}}} \Rightarrow \mathcal{N}(0, 1)$.*

This result implies that valid confidence intervals for μ can be constructed directly using the estimated variance $\hat{V}_{\text{IPW,CAL}}$, despite the adaptive and sequential nature of the labeling process.

Assumption 5.6 (Stability of Base Models). Suppose the parameters of the D-model and G-model admit probability limits θ^* and ϕ^* , such that $\hat{\theta} \xrightarrow{P} \theta^*$ and $\hat{\phi} \xrightarrow{P} \phi^*$. In addition, the evaluation dataset size and the historical dataset size satisfy $n/N \rightarrow c$ for some constant $c \in (0, \infty)$.

Proposition 5.7 (Adjusted AIPW estimator). *Under the adaptive labeling design in Section 3.3, for any finite number of stages T , the estimator $\hat{\mu}_{\text{AIPW,CAL}}$ is unbiased for μ . Moreover, under Assumption 5.6, it is consistent for μ and asymptotically normal, i.e., $\frac{\sqrt{n}(\hat{\mu}_{\text{AIPW,CAL}} - \mu)}{\sqrt{\hat{V}_{\text{AIPW,CAL}}}} \Rightarrow \mathcal{N}(0, 1)$.*

The adjusted AIPW estimator further incorporates predictions from the base models to reduce variance while preserving validity under the adaptive labeling framework. Notably, the proposition does not rely on strong modeling assumptions on these base models, making the estimator robust to potential misspecification. This combination of robustness and efficiency is particularly valuable in settings where gold labels are limited and labeling decisions depend heavily on silver labels. Empirical results in Section 6.2 demonstrate the practical efficiency gains of this estimator in realistic ML evaluation scenarios.

6. Experiments

We first present a synthetic study in Section 6.1 to illustrate the diagnostic role of silver labels and how posterior updates improve the reliability of the uncertainty score. We then evaluate the performance of the proposed CAL framework on two real-world datasets in Section 6.2. Detailed implementation settings and additional experiments are deferred to the Appendix.

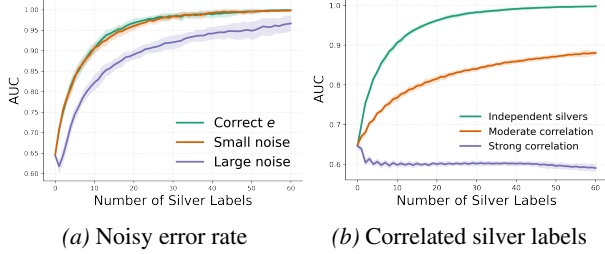


Figure 2. Evolution of the AUC between the uncertainty score and the realized labeling error as the number of silver labels increases under different practical conditions.

6.1. Synthetic Data

We construct a synthetic experiment to study how silver label updates improve the quality of the uncertainty score produced by an imperfect prediction model under several practical scenarios.

The initial prediction probabilities $p_0(X_i)$, the corresponding uncertainty scores, and the gold labels are randomly generated. Silver labels are generated with heterogeneous, instance-specific error rates. The prediction update rule follows the rule described in Section 3.3.2.

Scenarios. We consider two realistic situations that may weaken the diagnostic role of silver labels. First, we introduce noise to the true silver error rate when performing the posterior update. This mimics the case where the D-model is learned from data and its estimated difficulty variable deviates from the truth. Second, we introduce a latent shared bias term into the silver label generation process. This mimics correlated mistakes across silver labels, for example due to shared misconceptions or information leakage.

Results. Figure 2 illustrates how the alignment between the uncertainty score and the realized labeling error evolves as more silver labels are collected. When the error rate used in the update is noisy, the uncertainty score still becomes increasingly predictive of labeling mistakes, demonstrating robustness to misspecification. In contrast, when silver labels are correlated, the alignment improves more slowly and may even deteriorate as additional silver labels are added. This occurs because correlated errors reinforce the same incorrect belief, reducing the effective information gained from each query.

6.2. Real-World Data

Data and labels. We evaluate CAL on two real-world datasets. The first is CIFAR-10, consisting of 32×32 color images from 10 classes. Silver labels are synthetically generated according to class-dependent error rates to mimic heterogeneous annotator quality. The second is the Google Image Caption Quality (GICQ) dataset, where AI-generated

captions for Open Images are evaluated by crowd workers with binary quality labels. In this dataset, silver labels are sampled without replacement from real crowd annotations to preserve both heterogeneity and correlation. Gold labels are constructed by majority vote over the crowd labels with additional noise injected to reflect imperfect expert annotation. To mimic realistic labeling settings with limited data, we subsample 2000 instances for each run. Among them, 30% form a historical dataset for training the base models, and the remaining 70% form the evaluation pool for adaptive labeling. We fix labeling costs to $c_g = 1$ and $c_s = 0.1$.

Base models. For CIFAR-10, we use a convolutional neural network (CNN) trained on the historical dataset as the initial gold prediction model. For GICQ, each instance is an image-caption pair. We obtain pretrained CLIP embeddings for both the image and the caption, construct interaction features from the embeddings, and train a regularized linear classifier as the base model.

Benchmark methods. We compare CAL with three baselines: (i) *Active learning with gold*, which performs uncertainty-based sampling using only gold labels; (ii) *Random gold*, which randomly allocates the gold labeling budget; (iii) *Random silver*, which randomly queries silver labels and applies majority voting.

Uncertainty measures. When comparing CAL with baselines, we use margin score as the default uncertainty measure. We also evaluate entropy and variance scores to examine the robustness of CAL to the choice of uncertainty metric.

Results. Figure 4 shows that CAL consistently outperforms the baselines across budgets, especially when the budget is small. As the budget increases, CAL behaves similarly to active learning with gold on GICQ, since the budget is enough for most instances to be eventually escalated to gold. Figure 5 shows that CAL is robust to the choice of uncertainty measure, with margin sampling performing slightly better in both datasets. Figure 3 demonstrates that the proposed IPW and AIPW estimators are unbiased and consistent for estimating the population average, while the AIPW estimator achieves substantially smaller variance.

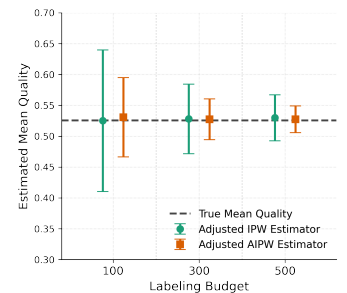


Figure 3. Confidence interval performance of IPW and AIPW under CAL.

References

Langley, P. Crafting papers on machine

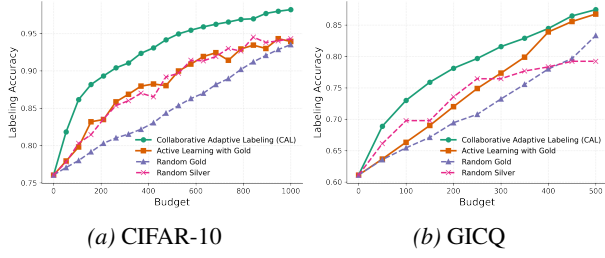


Figure 4. Labeling accuracy versus budget for CAL and baseline methods.

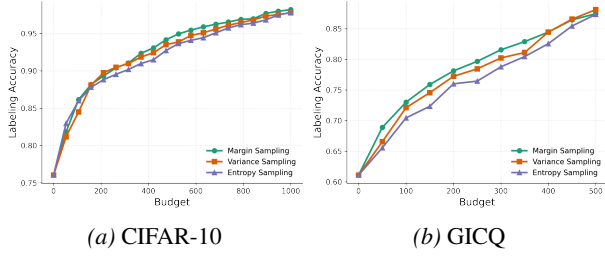


Figure 5. Effect of different uncertainty measures on CAL performance.

learning. In Langley,
P. (ed.), *Proceedings
of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000.
Morgan Kaufmann.

A. Extension

A.1. Collaborative Adaptive Labeling Pseudocode

Algorithm 1 Collaborative Adaptive Labeling (CAL)

Input: historical data $\mathcal{D}$, unlabeled pool $\mathcal{D}^*$, budgets $\{B_t\}_{t=1}^T$, costs c_g, c_s , temperature τ

Train G-model $p(X; \hat{\theta}_0)$ and D-model $e(X; \hat{\phi}_0)$ on $\mathcal{D}$

for $t = 1$ **to** T **do**

for each instance $i \in \mathcal{D}^*$ **do**

 Compute silver-guided prediction $p_t(X_i)$

 Compute uncertainty score $s_t(X_i)$

 Compute one-step-ahead uncertainty $\hat{s}_{t+1}(X_i)$

$u_{i,t,G} \leftarrow s_t(X_i)/c_g$

$u_{i,t,S} \leftarrow (s_t(X_i) - \hat{s}_{t+1}(X_i))/c_s$

$q_{i,t,G} \leftarrow \frac{\exp(u_{i,t,G}/\tau)}{\exp(u_{i,t,G}/\tau) + \exp(u_{i,t,S}/\tau)}$

$\tilde{u}_{i,t} \leftarrow q_{i,t,G}u_{i,t,G} + (1 - q_{i,t,G})u_{i,t,S}$

$\tilde{c}_{i,t} \leftarrow q_{i,t,G}c_g + (1 - q_{i,t,G})c_s$

end for

 Compute weights $\tilde{w}_{i,t} = \frac{\exp(\tilde{u}_{i,t}/\tau)}{\sum_j \exp(\tilde{u}_{j,t}/\tau)}$

$\alpha_t \leftarrow \frac{B_t}{\sum_i \tilde{w}_{i,t} \tilde{c}_{i,t}}$

for each instance $i \in \mathcal{D}^*$ **do**

 Sample i with probability $\alpha_t \tilde{w}_{i,t}$

if i is selected **then**

 Sample action: GOLD with prob. $q_{i,t,G}$, SILVER otherwise

if GOLD **then**

 Obtain gold label Y_i

 Remove i from $\mathcal{D}^*$

else

 Obtain silver label $Z_{i,t}$

 Update silver counts for i

end if

end if

end for

 Update G-model and D-model using all collected data: $\hat{\theta}_{t-1} \rightarrow \hat{\theta}_t, \hat{\phi}_{t-1} \rightarrow \hat{\phi}_t$.

end for

Output: collected labels and sampling probabilities

A.2. Channel-Aware Silver Labeling Extension

The CAL framework naturally extends to the setting where silver labels originate from multiple heterogeneous channels. Suppose there are K silver channels. For each channel k , let $Z_i^{(k)} \in \{0, 1\}$ denote a silver label, with channel-specific error model

$$e_k(X; \phi_k) = \mathbb{P}(Z^{(k)} \neq Y \mid X), \quad e_k(X) \leq \bar{e} < \frac{1}{2},$$

and cost $c_s^{(k)}$. If instance i has received $m_{i,t}^{(k)}$ silver labels from channel k , among which $k_{i,t}^{(k)}$ are 1's, the silver-guided prediction update in Section 3.3.2 generalizes to

$$p_t(X_i) \propto p(X_i; \hat{\theta}) \prod_{k=1}^K (1 - e_k(X_i; \hat{\phi}_k))^{k_{i,t}^{(k)}} e_k(X_i; \hat{\phi}_k)^{m_{i,t}^{(k)} - k_{i,t}^{(k)}}.$$

Each channel contributes a weight $\log \frac{1-e_k(X_i)}{e_k(X_i)}$ in the log-odds update. The one-step-ahead uncertainty for querying a silver label from channel k is

$$\hat{s}_{t+1}^{(k)}(X_i) = \sum_{z \in \{0,1\}} \hat{\mathbb{P}}\left(Z_{i,t+1}^{(k)} = z \mid X_i, \mathbf{Z}_{i,t}\right) f\left(p_{t+1}^{(k,z)}(X_i)\right),$$

where

$$\hat{\mathbb{P}}\left(Z_{i,t+1}^{(k)} = 1 \mid X_i, \mathbf{Z}_{i,t}\right) = e_k(X_i) + (1 - 2e_k(X_i))p_t(X_i).$$

The sampling score becomes

$$u_{i,t,\text{SILVER}(k)} = \frac{s_t(X_i) - \hat{s}_{t+1}^{(k)}(X_i)}{c_s^{(k)}}.$$

The labeling action space extends to

$$\{\text{GOLD}, \text{SILVER}(1), \dots, \text{SILVER}(K), \text{NONE}\},$$

and the softmax preference in Section 3.3 is applied over these actions. This extension enables CAL to automatically determine *which silver channel to query* based on its reliability and cost, without any modification to the core algorithm or theory.

A.3. Multiclass

yao b

A.4. General ERM

B. Proof

B.1. Proof of Lemma 5.1

Proof. For any score U , define the quantile threshold

$$t_b(U) := \inf\{t : \mathbb{P}(U \geq t) \leq b\}.$$

Then $S_b(U) = \{U \geq t_b(U)\}$ and

$$G_b(U) = \mathbb{P}(W = 1 \mid U \geq t_b(U)) = \frac{\mathbb{P}(U \geq t_b(U) \mid W = 1) \mathbb{P}(W = 1)}{\mathbb{P}(U \geq t_b(U))}.$$

Since $\mathbb{P}(U \geq t_b(U)) = b$, we obtain

$$G_b(U) = \frac{\mathbb{P}(W = 1)}{b} \mathbb{P}(U \geq t_b(U) \mid W = 1).$$

Hence

$$\int_0^1 G_b(U) db = \mathbb{P}(W = 1) \int_0^1 \frac{\mathbb{P}(U \geq t_b(U) \mid W = 1)}{b} db.$$

Let $F_{U|W=1}$ denote the CDF of U conditional on $W = 1$, and write $Q_U(b)$ for the $(1 - b)$ -quantile of U , so that $t_b(U) = Q_U(b)$. A change of variables yields

$$\int_0^1 G_b(U) db = \mathbb{P}(W = 1) \int_{\mathbb{R}} \mathbb{P}(U \geq u \mid W = 1) dF_U(u).$$

An analogous identity holds for any score S .

Observe that

$$\text{AUC}(U, W) = \mathbb{P}(U^+ > U^-) + \frac{1}{2}\mathbb{P}(U^+ = U^-) = \int_{\mathbb{R}} \mathbb{P}(U \geq u \mid W = 1) dF_{U|W=0}(u).$$

Thus, a larger AUC means that, in an average sense over thresholds u drawn from the distribution of $U \mid W = 0$, the tail probability $\mathbb{P}(U \geq u \mid W = 1)$ is larger.

Since F_U is a mixture of $F_{U|W=1}$ and $F_{U|W=0}$, a larger AUC implies that the integral

$$\int_{\mathbb{R}} \mathbb{P}(U \geq u \mid W = 1) dF_U(u)$$

is larger. Therefore,

$$\int_0^1 G_b(S) db > \int_0^1 G_b(U) db.$$

□

B.2. Proof of Proposition 5.4

Proof. Recall that

$$\text{AUC}(Z, W) = \mathbb{P}(Z^+ > Z^-) + \frac{1}{2}\mathbb{P}(Z^+ = Z^-),$$

where $Z^+ \sim Z \mid (W = 1)$ and $Z^- \sim Z \mid (W = 0)$ are independent. Let (U^+, M^+) and (U^-, M^-) be independent draws from the conditional distributions given $W = 1$ and $W = 0$, respectively. Define

$$D := U^+ - U^-, \quad \Gamma := \Delta^+ - \Delta^-,$$

so that

$$T^{\text{sil},+} - T^{\text{sil},-} = D + \Gamma.$$

Under Assumptions 5.2–5.3, conditioning on (Y, X) gives

$$\mathbb{P}(M = 0 \mid W = 1) = 1 - e(X), \quad \mathbb{P}(M = 0 \mid W = 0) = e(X),$$

and since $e(X) \leq \bar{e} < 1/2$, we have

$$\mathbb{P}(M = 0 \mid W = 1) > \mathbb{P}(M = 0 \mid W = 0).$$

Because disagreement yields a positive increment δ^+ while agreement yields a negative decrement $-\delta^-$ with $\delta^+ + \delta^- > 0$, it follows that

$$\mathbb{E}[\Gamma] > 0.$$

Moreover, Γ takes both positive and negative values with positive probability. Hence, for any fixed value d ,

$$\mathbb{P}(d + \Gamma > 0) \geq \mathbf{1}\{d > 0\},$$

with strict inequality whenever $d \leq 0$. Integrating over the distribution of D yields

$$\mathbb{P}(D + \Gamma > 0) > \mathbb{P}(D > 0)$$

whenever $\mathbb{P}(D \leq 0) > 0$, i.e., whenever $\text{AUC}(U, W) < 1$.

The same argument applies to equality events, which gives

$$\text{AUC}(T^{\text{sil}}, W) > \text{AUC}(U, W),$$

unless $\text{AUC}(U, W) = 1$.

□

B.3. Proof of Proposition 5.5

B.3.1. UNBIASEDNESS

Proof. We first note that the stopping time τ_i is measurable with respect to the filtration $\{\mathcal{F}_t\}$ since

$$\{\tau_i = t\} = \{\delta_{i,1} = 0, \dots, \delta_{i,t-1} = 0, \delta_{i,t} = 1\}$$

is determined entirely by the adaptive sampling rule based on past information.

Next, observe the identity

$$(T - \tau_i)_+ = T - \min(\tau_i, T),$$

so that

$$\begin{aligned} D_i &= \frac{1}{T} \left(\sum_{t \leq \min(\tau_i, T)} \frac{\delta_{i,t}}{\pi_{i,t}} + (T - \tau_i)_+ \right) \\ &= 1 + \frac{1}{T} \left(\sum_{t \leq \min(\tau_i, T)} \frac{\delta_{i,t}}{\pi_{i,t}} - \min(\tau_i, T) \right). \end{aligned}$$

We now compute the expectation of the summation term. Using iterated expectation and the fact that

$$\mathbb{E}[\delta_{i,t} \mid \mathcal{F}_{t-1}] = \pi_{i,t},$$

we have

$$\begin{aligned} \mathbb{E} \left[\sum_{t \leq \min(\tau_i, T)} \frac{\delta_{i,t}}{\pi_{i,t}} \right] &= \sum_{t=1}^T \mathbb{E} \left[\mathbf{1}\{t \leq \tau_i\} \frac{\delta_{i,t}}{\pi_{i,t}} \right] \\ &= \sum_{t=1}^T \mathbb{E} \left[\mathbf{1}\{t \leq \tau_i\} \frac{\mathbb{E}[\delta_{i,t} \mid \mathcal{F}_{t-1}]}{\pi_{i,t}} \right] \\ &= \sum_{t=1}^T \mathbb{E}[\mathbf{1}\{t \leq \tau_i\}] = \mathbb{E}[\min(\tau_i, T)]. \end{aligned}$$

Hence,

$$\mathbb{E}[D_i] = 1.$$

We now prove unbiasedness using the IPW randomization identity. For any stage t ,

$$\mathbb{E} \left[\frac{\delta_{i,t}}{\pi_{i,t}} \mid \mathcal{F}_{t-1}, Y_i \right] = 1,$$

since $\pi_{i,t} = \mathbb{P}(\delta_{i,t} = 1 \mid \mathcal{F}_{t-1})$ and the sampling decision does not depend on the unobserved gold label Y_i .

Therefore,

$$\begin{aligned} \mathbb{E}[D_i Y_i] &= \frac{1}{T} \mathbb{E} \left[Y_i \left(\sum_{t \leq \min(\tau_i, T)} \frac{\delta_{i,t}}{\pi_{i,t}} + (T - \tau_i)_+ \right) \right] \\ &= \frac{1}{T} \mathbb{E}[Y_i (\min(\tau_i, T) + (T - \tau_i)_+)] \\ &= \mathbb{E}[Y_i] = \mu. \end{aligned}$$

Averaging over i yields

$$\mathbb{E}[\hat{\mu}_{\text{IPW,CAL}}] = \mu.$$

□

B.3.2. CONSISTENCY

Proof. From the martingale decomposition, write

$$M_{i,t} := \delta_{i,t} - \pi_{i,t}, \quad \mathbb{E}[M_{i,t} \mid \mathcal{F}_{t-1}] = 0.$$

Using the identity

$$\sum_{t \leq \min(\tau_i, T)} \frac{\delta_{i,t}}{\pi_{i,t}} = \min(\tau_i, T) + \sum_{t \leq \min(\tau_i, T)} \frac{M_{i,t}}{\pi_{i,t}},$$

and the fact that $\min(\tau_i, T) + (T - \tau_i)_+ = T$, we obtain

$$D_i = 1 + \frac{1}{T} \sum_{t \leq \min(\tau_i, T)} \frac{M_{i,t}}{\pi_{i,t}}.$$

Hence

$$D_i Y_i - \mu = \frac{Y_i}{T} \sum_{t \leq \min(\tau_i, T)} \frac{M_{i,t}}{\pi_{i,t}}.$$

Since instances are independent across i , we have

$$\text{Var}(\hat{\mu}_{\text{IPW,CAL}}) = \frac{1}{n} \text{Var}(D_i Y_i).$$

We now bound $\text{Var}(D_i Y_i)$. Using iterated expectation and the martingale property,

$$\begin{aligned} \text{Var}(D_i Y_i) &= \mathbb{E} \left[\left(\frac{Y_i}{T} \sum_{t \leq \min(\tau_i, T)} \frac{M_{i,t}}{\pi_{i,t}} \right)^2 \right] \\ &\leq \frac{Y_i^2}{T^2} \mathbb{E} \left[\sum_{t=1}^T \frac{M_{i,t}^2}{\pi_{i,t}^2} \mathbf{1}_{\{\tau_i \geq t\}} \right], \end{aligned}$$

where cross terms vanish because $\{M_{i,t}\}$ is a martingale difference sequence.

Since $\delta_{i,t} \in \{0, 1\}$,

$$\mathbb{E}[M_{i,t}^2 \mid \mathcal{F}_{t-1}] = \text{Var}(\delta_{i,t} \mid \mathcal{F}_{t-1}) \leq \pi_{i,t},$$

hence

$$\mathbb{E} \left[\frac{M_{i,t}^2}{\pi_{i,t}^2} \right] \leq \mathbb{E} \left[\frac{1}{\pi_{i,t}} \right].$$

Under the positivity condition $\pi_{i,t} \geq \underline{\pi} > 0$, we obtain

$$\text{Var}(D_i Y_i) \leq C$$

for some finite constant C independent of n .

Therefore,

$$\text{Var}(\hat{\mu}_{\text{IPW,CAL}}) \leq \frac{C}{n}.$$

Applying Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\mu}_{\text{IPW,CAL}} - \mu| > \varepsilon) \leq \frac{C}{n\varepsilon^2} \rightarrow 0.$$

Hence,

$$\hat{\mu}_{\text{IPW,CAL}} \xrightarrow{P} \mu.$$

□

B.3.3. ASYMPTOTIC NORMALITY

Proof. Recall the decomposition

$$D_i = 1 + \frac{1}{T} \sum_{t \leq \min(\tau_i, T)} \frac{M_{i,t}}{\pi_{i,t}}, \quad M_{i,t} = \delta_{i,t} - \pi_{i,t},$$

with $\mathbb{E}[M_{i,t} \mid \mathcal{F}_{t-1}] = 0$. Hence

$$D_i Y_i - \mu = \frac{Y_i}{T} \sum_{t \leq \min(\tau_i, T)} \frac{M_{i,t}}{\pi_{i,t}}.$$

For each fixed i , the sequence

$$\left\{ Y_i \frac{M_{i,t}}{\pi_{i,t}} \mathbf{1}_{\{\tau_i \geq t\}} \right\}_{t=1}^T$$

forms a martingale difference sequence with respect to $\{\mathcal{F}_t\}$. Therefore $D_i Y_i$ can be written as a finite sum of martingale increments.

Since instances are independent across i , the random variables

$$Z_i := D_i Y_i$$

are i.i.d. with $\mathbb{E}[Z_i] = \mu$ and finite variance (shown in the consistency proof). Hence, by the CLT,

$$\sqrt{n}(\hat{\mu}_{\text{IPW,CAL}} - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (Z_i - \mu) \Rightarrow \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \text{Var}(Z_i)$.

It remains to show consistency of the variance estimator

$$\hat{\mathbb{V}}_{\text{IPW,CAL}} = \frac{1}{n} \sum_{i=1}^n (Z_i - \hat{\mu}_{\text{IPW,CAL}})^2.$$

Since $\{Z_i\}$ are i.i.d. with finite second moment, the law of large numbers implies

$$\hat{\mathbb{V}}_{\text{IPW,CAL}} \xrightarrow{P} \sigma^2.$$

Finally, Slutsky's theorem yields

$$\frac{\sqrt{n}(\hat{\mu}_{\text{IPW,CAL}} - \mu)}{\sqrt{\hat{\mathbb{V}}_{\text{IPW,CAL}}}} \Rightarrow \mathcal{N}(0, 1).$$

□

B.4. Proof of Proposition 5.7

B.4.1. UNBIASEDNESS

Proof. We first prove unbiasedness. Recall

$$\psi_i = D_i Y_i + \frac{1}{T} \sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) p_{t-1}(X_i).$$

From the proof of Proposition 5.5, we already have

$$\mathbb{E}[D_i Y_i] = \mu.$$

Hence it suffices to show that the augmentation term has mean zero. For each $t \leq T$, since $\pi_{i,t} = \mathbb{P}(\delta_{i,t} = 1 \mid \mathcal{F}_{t-1})$ and $p_{t-1}(X_i)$ is $\mathcal{F}_{t-1}$ -measurable,

$$\begin{aligned} \mathbb{E} \left[\left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) p_{t-1}(X_i) \mid \mathcal{F}_{t-1}, \tau_i \geq t \right] &= p_{t-1}(X_i) \left(1 - \frac{\mathbb{E}[\delta_{i,t} \mid \mathcal{F}_{t-1}]}{\pi_{i,t}} \right) \\ &= 0. \end{aligned}$$

Taking expectation over the stopping time τ_i yields

$$\mathbb{E} \left[\frac{1}{T} \sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) p_{t-1}(X_i) \right] = 0.$$

Therefore,

$$\mathbb{E}[\psi_i] = \mu, \quad \mathbb{E}[\hat{\mu}_{\text{AIPW,CAL}}] = \mu.$$

□

B.4.2. CONSISTENCY AND ASYMPTOTIC NORMALITY

Proof. We prove consistency and asymptotic normality by comparing ψ_i with an idealized version that uses the probability limits of the prediction models.

Step 1: Decompose the estimator. Write

$$\psi_i = D_i Y_i + \frac{1}{T} \sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) p_{t-1}(X_i),$$

and define the deterministic limit version

$$\psi_i^* = D_i Y_i + \frac{1}{T} \sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) p^*(X_i),$$

where $p^*(X)$ is the probability limit implied by Assumption 5.6.

Then

$$\hat{\mu}_{\text{AIPW,CAL}} - \frac{1}{n} \sum_{i=1}^n \psi_i^* = \frac{1}{nT} \sum_{i=1}^n \sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) (p_{t-1}(X_i) - p^*(X_i)).$$

Denote this difference by R_n .

Step 2: Show $R_n = o_p(n^{-1/2})$. For each fixed i and t , the term

$$\left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) (p_{t-1}(X_i) - p^*(X_i)) \mathbf{1}\{\tau_i \geq t\}$$

is a martingale difference with respect to $\{\mathcal{F}_t\}$, since $p_{t-1}(X_i)$ is $\mathcal{F}_{t-1}$ -measurable and $\mathbb{E}[\delta_{i,t} \mid \mathcal{F}_{t-1}] = \pi_{i,t}$.

Using independence across i and the martingale property (cross terms vanish),

$$\begin{aligned} \text{Var}(R_n) &= \text{Var} \left(\frac{1}{nT} \sum_{i=1}^n \sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) (p_{t-1}(X_i) - p^*(X_i)) \right) \\ &\leq \frac{1}{n^2 T^2} \sum_{i=1}^n \text{Var} \left(\sum_{t \leq \min\{\tau_i, T\}} \left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right) (p_{t-1}(X_i) - p^*(X_i)) \right) \\ &= \frac{1}{n^2 T^2} \sum_{i=1}^n \sum_{t=1}^T \mathbb{E} \left[\left(1 - \frac{\delta_{i,t}}{\pi_{i,t}} \right)^2 (p_{t-1}(X_i) - p^*(X_i))^2 \mathbf{1}\{\tau_i \geq t\} \right] \\ &\leq \frac{C}{n} \mathbb{E} \left[(p_{t-1}(X) - p^*(X))^2 \right], \end{aligned}$$

for some constant C depending only on T and the positivity bound on $\pi_{i,t}$.

By Assumption 5.6, $p_{t-1}(X) \xrightarrow{p} p^*(X)$ and the second moment is bounded, hence

$$\mathbb{E}[(p_{t-1}(X) - p^*(X))^2] \rightarrow 0.$$

Therefore,

$$\text{Var}(R_n) = o\left(\frac{1}{n}\right), \quad R_n = o_p(n^{-1/2}).$$

Step 3: Analyze the idealized estimator. For ψ_i^* , the prediction term is now deterministic given X_i . As in the proof of Proposition 5.5, ψ_i^* is a finite sum of martingale differences and therefore has finite variance. Moreover,

$$\mathbb{E}[\psi_i^*] = \mu.$$

Since instances are independent across i , the CLT gives

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \psi_i^* - \mu \right) \Rightarrow \mathcal{N}(0, \sigma^2), \quad \sigma^2 = \text{Var}(\psi_i^*).$$

Step 4: Transfer the result back to $\hat{\mu}_{\text{AIPW,CAL}}$. Because $R_n = o_p(n^{-1/2})$, Slutsky's theorem yields

$$\sqrt{n}(\hat{\mu}_{\text{AIPW,CAL}} - \mu) \Rightarrow \mathcal{N}(0, \sigma^2).$$

Consistency follows immediately from the law of large numbers applied to ψ_i^* and the fact that $R_n \rightarrow 0$ in probability.

Finally, the sample variance $\hat{V}_{\text{AIPW,CAL}}$ is consistent for σ^2 by the law of large numbers, and another application of Slutsky's theorem gives

$$\frac{\sqrt{n}(\hat{\mu}_{\text{AIPW,CAL}} - \mu)}{\sqrt{\hat{V}_{\text{AIPW,CAL}}}} \Rightarrow \mathcal{N}(0, 1).$$

□

C. Experiment detail

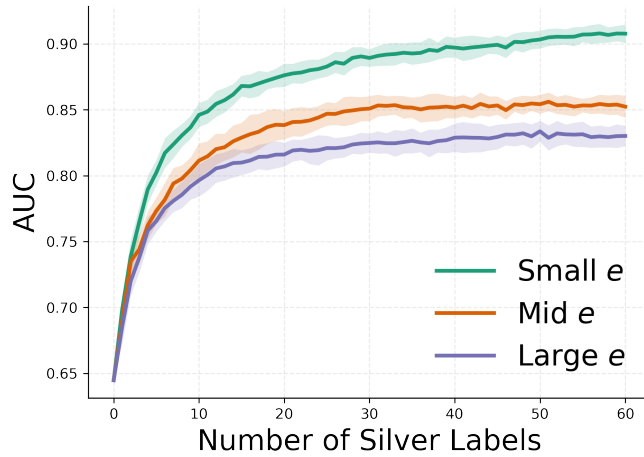


Figure 6. Caption