Adaptive Stratified Active Statistical Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

Active Statistical Inference (ASI) leverages machine learning predictions to guide label acquisition and improve statistical inference under limited labeling budgets, but lacks the nuanced understanding of model reliability varying across the data manifold, leading to inefficient budget allocation. To address this, we propose Adaptive Stratified Active Statistical Inference (AdaStrat-ASI), a framework that replaces ASI's global sampling rule with stratum-specific local policies, designed via a model-guided 'scouting' phase. We provide theoretical results showing that AdaStrat-ASI achieves strictly lower asymptotic variance compared to ASI while preserving the inferential guarantees of ASI. We verify our theoretical findings through empirical results on real-world datasets demonstrate that AdaStrat-ASI yields tighter confidence intervals than existing baselines under the same labeling budget.

1 Introduction

High-quality labels data are often a bottleneck in data-driven research. Consequently, practitioners are increasingly adopting machine learning predictions as inexpensive proxies for missing labels, in applications ranging from remote sensing [28, 58, 51], proteomics [29], and electoral systems [59], enabling large-scale measurement possible in settings where exhaustive human annotation would be prohibitively expensive. Yet prediction is not inference: even highly accurate predictions do not, by themselves, provide statistical guarantees required for decision-making. Confidence intervals and hypothesis tests still require access to ground truth labels.

Active Statistical Inference (ASI) [65] addresses this gap by using predictions not as a substitute for labels but as a guide for acquiring them. It samples labels according to a single *global* uncertainty rule and then corrects the estimator using an inverse propensity weighting. This yields statistically valid inference for a broad class of M-estimation targets under tight budgets. The limitation is that ASI asks a global question of a local object. Its sampling rule normalizes uncertainty over the entire data pool, so all points compete for the same lableing budget on a single global scale. This is natural when predictive reliability is roughly homogeneous. In practice, this is seldom the case – a model may be accurate on one subpopulation, biased in another, calibrated in one region and confidently wrong in another [4, 17]. LLM annotators [34, 64, 49, 56] exhibit the same pattern: they may perform well on well-specified or easily verifable tasks [35, 12], yet become unreliable on open-ended, weakly grounded or demographically sensitive inputs [57, 26].

Can ASI be made more efficient by exploiting region-specific reliability differential without sacrificing validity?

We answer this question with AdaStrat-ASI, an adaptive stratified framework for active statistical inference. The key idea is to make reliability local: given a partition of data into strata – defined by domain knowledge, observed covariate groupings, or model-induced structure – AdaStrat-ASI replaces ASI's single global sampling rule with stratum-specific local policies. While lables are still

acquired actively within each stratum, the total budget allocation is no longer spread according to one global uncertainty rule. Instead, AdaStrat-ASI allocates more labels to strata where the predictor is locally less useful for inference, and fewer labels to strat where the predictor is locally stable. In an oracle design, these label budgets would be assigned to strata based according to their contribution to asymptotic variance of the final estimator. But this requires access to true labels to calculate. AdaStrat-ASI resolves this cold-start problem through a label-free scouting stage. Before querying any labels, it uses the black-box predictor on the unlabelled data to estimate a proxy difficulty score for each stratum, measuring the local variability of proxy loss gradients which serve as a model-based signal for how much the stratum may affect the final estimator. This results in a Neyman-style allocation (Section 4.1). After this allocation step, a local ASI sampling rule is applied within each stratum. Thus, our method separates two decisions which are conflated in global ASI: *which points* to label within a region, and *how much budget* that region should receive. We summarize our contributions as follows:

- **Method.** We propose Adaptive Stratified Active Statistical Inference (AdaStrat-ASI), a stratified extension of ASI that replaces one global sampling policty with stratum-specific local policies under a fixed labeling budget to improve statistical efficiency.
- **Validity.** We prove that our method is statistically valid: under standard regularity conditions, it is asymptotically normal and supports valid confidence intervals (Theorem 5.3).
- **Efficiency.** We characterize an oracle optimal stratum allocation (Proposition 5.4), relate stratum difficulty to predictor error (Lemma 5.5), and provide a sufficient condition under which AdaStrat-ASI provably attains a strictly smaller asymptotic parameter covariance than ASI (Theorem 5.6), clarifying when stratification yields a principled efficiency gain.
- **Empirical validation.** Our results empirically validate our theoretical claims, demonstrating AdaStrat-ASI's narrower confidence intervals under identical labeling budgets compared against standard baselines on synthetic and real-world datasets. The source code for this project will be made publicly available upon acceptance of the manuscript.

Technical Overview. While stratification is standard, our focus is *ML-integrated inference*. Any inference framework that treats the predictor as a globally homogeneous oracle is theoretically misaligned with reality [55, 5]. AdaStrat-ASI resolves this misalignment by treating the latent structure of model performance in strata as the central theme in the sampling design by proposing a novel estimator (Eq. 3). We provide the first theoretical analysis within the ASI framework that explicitly links model error to inference efficiency—showing how it enters the asymptotic covariance (Lemma C.3 in Appendix) and drives the Neyman-style oracle label allocation across strata (Lemma 5.5), offering a nuanced understanding of *why* and *where* ML-guided inference succeeds or fails. Further, since optimal label allocation depends on unknown labels, we introduce a novel, *label-free scouting* mechanism that uses cheaply available proxy predictions $f(X)$ to estimate per-stratum "difficulty" and allocate a fixed labeling budget *without peeking* at the ground truth, effectively solving the cold-start problem for stratified active learning. Taken together, stratification is the mathematically natural response to the heterogeneity inherent in black-box predictors and is a principled route to improved statistical efficiency (Theorem 5.6).

2 Related Works

ML-enabled Data Annotation and Inference. A burgeoning body of literature exists on inference from adaptively collected data, encompassing works such as Kato et al. [31], Zhang et al. [61], Cook et al. [14], Lin et al. [37]. Additionally, adaptive experimental design, including approaches like Hahn et al. [23], Chandak et al. [8], has garnered significant attention. Under the regime of black-box predictions being well calibrated, ASI aligns most closely with our pursuits. It positions itself uniquely by adaptively leveraging modern, black-box machine learning uncertainties for data collection while adhering to budget constraints for statistical inference. ASI shares similarities with prediction-powered inference (PPI) and other recent works on inference with ML predictions [2, 3, 44, 42]. Notably, ASI considers the same class of inferential targets studied in Angelopoulos et al. [2, 3], Zrnic and Candès [66]. Recently, Li et al. propose Robust Sampling in ASI to mitigate the risk of unreliable uncertainty estimates by interpolating between uniform and uncertainty-based sampling, ensuring performance no worse than uniform sampling under model misspecification. In parallel, LLMs have shown great potential in handling text-annotation tasks without prior task-specific training, sometimes even outperforming crowd workers [21, 63, 38, 10, 33]. Inspired by active learning's role in LLM Annotation [62, 41], Gligoric et al. extended ASI to leverage LLM-provided confidence to guide

targeted human annotation and produce statistically valid estimates and confidence intervals under the confidence-driven inference (CCI) framework.

Stratified Sampling and Adaptivity. Standard variance reduction techniques like Stratified Sampling can be viewed as a statistical 'divide-and-conquer' strategy: rather than estimating parameters over a complex, heterogeneous data manifold, one partitions the space into homogeneous strata and solves simpler local estimation problems [13]. Although related in goal, Fisch et al. combined the variance reduction benefits of Stratified Sampling for PPI. However AdaStrat-ASI is fundamentally different in approach and idea. AdaStrat-ASI, is an extension of ASI i.e. our goal is to strategically collect data to yield powerful inference, leveraging known structure in data for variance reduction. Further, the efficiency of method hinges on label allocation inspired by Neyman allocation, which dictates that the optimal sample distribution is proportional to true variance [45]. However, the internal variance for the task of statistical inference is an unknown a priori. Our work addresses this cold start problem by creating a approximation of the optimality criterion using a two-stage adaptive design, reminiscent of Explore-then-Commit (EtC) strategies in bandit literature [20] or stratified Monte Carlo methods [6](discussed in detail in Section 4).

Region-based Active Learning. Zhang and Chaudhuri demonstrated that in agnostic settings, absolute disagreement [24] is overly conservative. They introduced confidence-rated predictors to study disagreement-based learning in sub-regions. This idea was further extended to contextual bandits through the lens of "Sub-Regions of Disagreement" [43]: treating disagreement as a binary, global property masks opportunities for exploitation; instead, the input space should be decomposed into sub-regions where the optimal policy is locally identifiable, effectively isolating the "hard" regions of the space. Our approach also shares conceptual lineage with Region-Based Active Learning (RBAL) [15], which adapts the sampling strategy based on the difficulty of learning within specific sub-regions of the feature space. Similar to RBAL, AdaStrat-ASI acknowledges that model performance is not uniform; however, while RBAL focuses on optimizing a hypothesis h during training, our work focuses on the statistical inference of population parameters.

3 Preliminaries

Problem Setup. We observe unlabeled samples $X_1, \ldots, X_n \sim \mathbb{P}_X$, with corresponding labels Y_i initially unobserved, and aim to estimate a population parameter θ^* defined as the minimizer of an expected loss:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell_\theta(X, Y)],$$

for a convex loss function ℓ_θ. Here, $(X, Y) \sim \mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$ denotes a generic feature-label pair drawn from the underlying data distribution. This formulation captures a broad class of M-estimation problems, including mean estimation, quantiles, and regression coefficients.

Active Statistical Estimator. We assume access to an off-the-shelf predictive model $f(X)$ for labels $Y \in \mathbb{R}$ given covariates $X \in \mathcal{X}$. We design a *sampling rule* $\pi : \mathcal{X} \to [0, 1]$ and collect label Y_i with probability $\pi(X_i)$. Our goal is to perform inference under a strict labeling budget ($n_b \ll n$). To overcome this, we use $f(X)$ as a strategic guide to select the most informative points to label. We denote the number of empirically collected labels as n_{lab}. The intuition is to model $\pi(.)$ to have a high(low) value for instances where f is uncertain(certain). At a technical level, the rule is derived from a measure of model-output uncertainty $u(x)$ (computed from f). The definition of uncertainty in classification and regression is taken from Section 4 of Zrnić and Candès [65]. Let $\xi_i \sim \text{Bernoulli}(\pi(X_i))$ indicate whether point i is labeled, so that $n_{\text{lab}} = \sum_{i=1}^{n} \xi_i$. To stay under a budget n_b, we set $\pi(x) = \frac{u(x)}{\mathbb{E}[u(X)]} \cdot \frac{n_b}{n}$, which yields $\mathbb{E}[n_{\text{lab}}] = \mathbb{E}[\pi(X)] \cdot n \leq n_b$.

The *active estimator* i.e. Equation 1 from Zrnić and Candès [65] is defined as $\hat{\theta}^\pi = \arg\min_\theta L^\pi(\theta)$, where our *active empirical risk* is as follows,

$$L^\pi(\theta) := \frac{1}{n} \sum_{i=1}^{n} \left[\ell^f_{\theta,i} + \left(\ell_{\theta,i} - \ell^f_{\theta,i} \right) \frac{\xi_i}{\pi(X_i)} \right] \tag{1}$$

To simplify notation, let $L(\theta) = \mathbb{E}[\ell_\theta(X, Y)]$, $\ell_{\theta,i} = \ell_\theta(X_i, Y_i)$, and $\ell^f_{\theta,i} = \ell_\theta(X_i, f(X_i))$. It is evident that the estimator exemplifies augmented inverse propensity weighting (AIPW) estimation

[50]. Further, when the sampling rule is uniform, i.e., $\pi(x) = \frac{n_b}{n}$ for all x, the estimator imitates the prediction-powered estimator [2]. Classical baselines such as $\hat{\theta}^{\texttt{noML}} := \arg\min_\theta \frac{1}{n_b} \sum_{i=1}^{n_b} \ell_{\theta,i}$ utilize pure labels.

In many data-efficient learning and inference settings, a complementary goal is to ensure that labeled data are well-distributed across the underlying data manifold [7, 52, 16]. Motivated by this perspective, we quantify the geometric spread of the labeled set via the diameter of the induced manifold. For a labeled dataset $\mathcal{X}_{\text{label}} \subset \mathbb{R}^d$ we define diameter of manifold $\mathrm{Diam}(.)$ as,

$$\mathrm{Diam}(\mathcal{X}_{\text{lab}}) \;=\; \max_{\substack{x_i,\, x_j \in \mathcal{X}_{\text{lab}} \\ i \neq j}} \big\|x_i - x_j\big\| \tag{2}$$

The diameter captures the maximal geometric separation among labeled points and serves as a proxy for the spread of the labeled data manifold.

Stratification and Model Performance. In accordance with previous works [19], we assume that the input space X is partitioned in advance into K non-empty, mutually exclusive, and exhaustive strata $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_K)$, where K is a finite integer. Let $n_k = |\mathcal{A}_k|$ denote the size of stratum k, such that $\sum_{k=1}^{K} n_k = n$. Let w_k represent the probability mass of stratum k in the finite population, where $w_k = \mathbb{P}(X \in \mathcal{A}_k)$. Naturally, w_k becomes $\frac{n_k}{n}$. A major premise of our work is that the performance of f is not uniform across the input space [55, 25, 5]. To formalize this, let $\mathbb{E}[(f(X) - Y)^2]$ be the expected MSE(Mean Squared Error) measuring $f(X)$'s fidelity to Y. While natural in regression settings, there exist extensions on MSE in classification settings(e.g., Hamming distance for binary; 0–1 loss for nominal multiclass; MSE for ordinal labels). We characterize the model's reliability within each stratum k by an upper bound on its expected error, denoted by ϵ_k. This means, $\mathbb{E}[(f(X) - Y)^2 \mid X \in \mathcal{A}_k] \leq \epsilon_k$. We further define $\epsilon^* = \max_k \epsilon_k$ as the worst-case error across all strata. In our framework, a lower ϵ_k implies $f(X)$ is a stronger proxy for Y in stratum k.

4 Main Algorithm

AdaStrat-ASI is designed to approximate the *oracle* way of spending a fixed labeling budget across strata. If we knew, for each stratum $\mathcal{A}_k$, how much labeling within it would reduce the variance of the final M-estimator, then the optimal design would follow a Neyman-style principle: allocate more labels to strata that are (i) prevalent (larger w_k) and (ii) *hard* for the predictor in the sense of contributing more to the estimator's asymptotic covariance. Formally, the oracle allocation depends on the stratum-specific, label-dependent quantities (e.g. true error, and resulting covariance contribution), and is therefore unavailable at design time. Our key observation is that these oracle difficulty terms admit a reasonable *label-free proxy* that can be estimated using only the black-box predictor $f(\cdot)$. We define a stratum difficulty score $\hat{\sigma}_k^f$ using the proxy Hessian and gradient covariance computer from $f(\cdot)$ on unlabeled data, and then allocate the main labeling budget according to $n_{b_k} \propto w_k \cdot \sqrt{\hat{\sigma}_k^f}$ which mirrors the oracle Neyman allocation but requires no access to Y. Algorithm 1 implements this idea via a short *scouting* phase for estimating $\hat{\sigma}_k^f$ followed by *attacking* phase for sampling and labeling according to the resulting stratum budgets, and finally computing an ASI-style(Eq. 1) estimator. Additional details on AdaStrat-ASI are described as follows.

4.1 Scouting Stage

In this phase, we leverage the availability of unlabeled data and the black-box model $f(x)$ to estimate the geometric properties of the loss landscape without querying any human labels. Let $\hat{\theta} := \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(X_i, f(X_i))$. First, we compute the proxy Hessian $\hat{H}_f$ using the entire unlabeled dataset at $\hat{\theta}$. This matrix captures the global steepness of the proxy loss surface, serving as a substitute for the unknown true Hessian. Then, for each stratum k, compute the covariance matrix $\hat{V}_k^f$ of the loss gradients, using $f(x)$, at $\hat{\theta}$. This measures how much the loss gradient vector fluctuates within that specific sub-population, solely based on the model's perspective. Finally, we combine the global curvature and local variance to compute the *difficulty* score $\hat{\sigma}_k^f = \mathrm{Tr}(\hat{H}_f^{-1} \hat{V}_k^f \hat{H}_f^{-1})$. This scalar quantifies how hard it is to minimize the loss in stratum k relative to the global objective. As a

result, we get a set of scores $\{\hat{\sigma}_1^f, \ldots, \hat{\sigma}_K^f\}$ that dictate the allocation of the labeling budget in the attacking stage.

4.2 Attacking Stage

Let n_{b_k} denote the number of labels allocated to stratum k, such that $\sum_{k=1}^K n_{b_k} = n_b$. We now distribute the labeling budget n_b to the areas identified as most critical using difficulty scores derived in the scouting stage: strata with higher weighted difficulty ($n_{b_k} \propto w_k \sqrt{\hat{\sigma}_k^f}$) receive more samples. In practice, n_{b_k} is realized by drawing $(n_{b_1}, \ldots, n_{b_K}) \sim \text{Multinomial}(n_b, \vec{z})$, where $z_k = \frac{w_k \sqrt{\hat{\sigma}_k^f}}{\sum_{j=1}^K w_j \sqrt{\hat{\sigma}_j^f}}$. This construction guarantees $\mathbb{E}[\sum_{k=1}^K n_{b_k}] = n_b$.

Once the budget is distributed, we locally apply ASI to each stratum. Specifically, we observe i.i.d. unlabeled pre-partitioned covariates $\mathcal{A}_{k1:n_k}$. We consider a budget–scaled family of sampling rules per stratum $\pi_{\eta_k}(x) = \eta_k\, u(x), \eta_k \in \mathcal{H} \subseteq \mathbb{R}^+$, and η_k is chosen to match the stratum label budget. We set $\hat{\eta}_k = \max\Big\{\eta_k \in \mathcal{H} : \eta_k \sum_{i=1}^{n_k} u(X_i) \leq n_{b_k}\Big\}$,. This gives rise to $\vec{\eta} := [\eta_1, \eta_2, \cdots \eta_k]$ and $\hat{\vec{\eta}} := [\hat{\eta}_1, \hat{\eta}_2, \cdots \hat{\eta}_k]$. The later deploys $\pi_{\hat{\vec{\eta}}}$, to ensure the expected number of collected labels does not exceed the overall budget. We denote $\hat{\theta}^{\vec{\eta}} \equiv \hat{\theta}^{\pi_{\vec{\eta}}}$. Optionally, to guarantee stability and some level of acceptance, i.e. to lower bound our acceptance probabilities, we rely on τ-mixing, i.e. essentially interpolation of our sampling rule and uniform sampling. Thus, $\pi_k^{(\tau)}(x) := \frac{n_{b_k}}{n_k}\left[(1-\tau)\,\frac{u(x)}{\hat{\mathbb{E}}_{X\sim\mathcal{A}_k}[u(X)]} + \tau\right]$, where $\tau \in (0,1)$. We query the oracle for labels Y_i based on the randomized decisions $\xi_{i,k} \sim \text{Bern}(\pi_k^{(\tau)}(X_i))$. It is worth noting, in the extreme situation of $[\hat{\sigma}_1^f, \ldots, \hat{\sigma}_K^f] = \vec{0}$ computed from Section 4.1, we use zero human labels for all strata($n_{b_k} = 0$ for all k) and purely rely on model output for computing the risk.

4.3 Final Estimation

Thus, our *stratified active estimator* is defined as $\hat{\theta}^{\pi_{\vec{\eta}}} = \arg\min_\theta L^{\pi_{\vec{\eta}}}(\theta)$, where a simple stratum-wise rearrangement of Eq. (1) gives us our unbiased *stratified active empirical risk*

$$L^{\pi_{\vec{\eta}}}(\theta) = \frac{1}{n}\sum_{k=1}^K \left[\sum_{i\in\mathcal{A}_k}\left[\ell_{\theta,i}^f + \frac{\xi_{i,k}}{\pi_k^\tau(X_i)}(\ell_{\theta,i} - \ell_{\theta,i}^f)\right]\right] \tag{3}$$

A short calculation show us that our method stays under budget; $\mathbb{E}[n_{\text{lab}}] = \mathbb{E}\Big[\sum_{k=1}^K \mathbb{E}[\sum_{i\in\mathcal{A}_k}\xi_{i,k} \mid n_{b_k}, \{X_i\}]\Big] = \mathbb{E}\Big[\sum_{k=1}^K \sum_{i\in\mathcal{A}_k}\pi_k(X)\Big] = \mathbb{E}\Big[\sum_{k=1}^K n_{b_k}\Big] \leq n_b$. On comparing Eq. (3) to Eq. (1), the only change comes from using a distinct sampling rule for every k, which in-turn depends on the label allocation from the previous scouting stage.

We provide some insights into data stratification for AdaStrat-ASI in the Appendix in Section E.4.

5 Theoretical Analysis

Like ASI, our setup also requires standard, mild smoothness assumptions on the loss ℓ_θ, articulated as follows:

Assumption 5.1. The loss ℓ is smooth if:

1. $\ell_\theta(x,y)$ is differentiable at θ^* for all (x,y);
2. ℓ_θ is locally Lipschitz around θ^*: there is a neighborhood of θ^* such that $\ell_\theta(x,y)$ is $C(x,y)$-Lipschitz and $\ell_\theta(x, f(x))$ is $C(x)$-Lipschitz in θ, where $\mathbb{E}[C(X,Y)^2] < \infty, \mathbb{E}[C(X)^2] < \infty$;
3. $L(\theta) = \mathbb{E}[\ell_\theta(X,Y)]$ and $L^f(\theta) = \mathbb{E}[\ell_\theta(X, f(X))]$ are double-differentiable, and $H_{\theta^*} = \nabla^2\mathbb{E}[\ell_{\theta^*}(X,Y)] = \nabla^2 L(\theta^*) \succ 0$;

Assumption 5.2. Additionally, we assume the gradient of the loss function $\nabla\ell_\theta(x,y)$ is stratumwise L_y-Lipschitz continuous with respect to the second argument y i.e. for $x \in A_k$, $\|\nabla\ell_\theta(x,y) - \nabla\ell_\theta(x,y')\| \leq L_y(x)\|y - y'\|$ where $\mathbb{E}[L_y(X)^2|\mathcal{A}_k] < \infty$.

In this section, we begin by proving that our AdaStrat-ASI produces valid confidence intervals and our estimator asymptotically normal. We then shed some light on the design of the Scouting Stage (Section 4.1) of Algorithm 1 in the Appendix B. We provide justification for relying on model predictions for label allocation. Finally, we show under mild conditions, our estimator is guaranteed to be more statistically efficient than the standard ASI estimator, resulting in a smaller asymptotic covariance.

In ASI it is established that the empirically chosen budget parameter $\hat{\eta}$ converges to the true optimal parameter η^* using Claim A.1 in Zrnić and Candès [65]. In AdaStrat-ASI, we run independent budget optimizations within each stratum k. We show simultaneous convergence happens for all strata(i.e $\mathbb{P}(\hat{\eta}_k \neq \eta_k^*) \to 0 \quad \forall k$) in Claim C.1 in Appendix.

Theorem 5.3. *Assume the loss is smooth(Ass. 5.1). Assume that for each stratum k, there exists an optimal budget parameter $\eta_k^* \in \mathcal{H}$ such that $\mathbb{P}(\hat{\eta}_k \neq \eta_k^*) \to 0$. Then, if $\hat{\theta}^{\vec{\eta}^*} \xrightarrow{p} \theta^*$, we have:*

$$\sqrt{n}(\hat{\theta}^{\hat{\vec{\eta}}} - \theta^*) \xrightarrow{d} N(0, \Sigma_{\vec{\eta}^*}), \textit{where}$$

$$\Sigma_{\vec{\eta}^*} = H_{\theta^*}^{-1}\left[\sum_{k=1}^{K} w_k \, \mathbb{V}\mathrm{ar}\left(\nabla \ell_{\theta^*}(X, f(X)) + \frac{\xi^{\eta_k^*}}{\pi_{\eta_k^*}(X)}\left(\nabla \ell_{\theta^*}(X,Y) - \nabla \ell_{\theta^*}(X, f(X))\right) \,\middle|\, X \in \mathcal{A}_k\right)\right] H_{\theta^*}^{-1},$$

and $\xi^{\eta_k^*} \sim \mathrm{Bern}(\pi_{\eta_k^*}(X))$. Consequently, for any consistent plug-in estimate $\hat{\Sigma} \xrightarrow{p} \Sigma_{\vec{\eta}^*}$, the confidence interval $\mathcal{C}_\alpha = (\hat{\theta}_j^{\hat{\vec{\eta}}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\Sigma}_{jj}}{n}})$ is a valid $(1-\alpha)$-asymptotic confidence interval for θ_j^*:

$$\lim_{n\to\infty} \mathbb{P}(\theta_j^* \in \mathcal{C}_\alpha) = 1 - \alpha.$$

A key challenge lies in the allocation of n_b labels across K strata. Let R_k^* represent the asymptotic weighted variance of the loss residuals within stratum k; $R_k^* := \mathrm{Tr}\left(H_{\theta^*}^{-1} W_k^* H_{\theta^*}^{-1}\right)$. W_k^* is the Weighted Covariance of the Residual Gradients in stratum k : $W_k^* = \mathbb{E}[\frac{1}{s_k(X)}(\nabla \ell_{\theta^*}(X,Y) - \nabla \ell_{\theta^*}(X, f(X)))(.)^\top | \mathcal{A}_k]$. Here, $s_k(X) = \frac{u(X)}{\mathbb{E}_{X\sim\mathcal{A}_k}[u(X)]}$ is the relative uncertainty-sampling score.

Proposition 5.4. *Given a fixed budget n_b and fixed stratum characteristics $\{w_k, R_k^*\}_{k=1}^K$, such that $w_k > 0$, $R_k^* > 0$, the allocation distribution vector $\{n_{b_1}, \ldots, n_{b_K}\}$ that minimizes AdaStrat-ASI's asymptotic variance $\Sigma_{\vec{\eta}^*}$ (Theorem 5.3) subject to $\sum_{k=1}^K n_{b_k} = n_b$, $n_{b_k} > 0$, $\forall k$ is unique and satisfies: $n_{b_k}^* = n_b \cdot p_k^*$, where $p_k^* = \frac{w_k\sqrt{R_k^*}}{\sum_{j=1}^K w_j \sqrt{R_j^*}}$*

When deeply inspected, on abstracting away H_{θ^*} and $s_k(X)$ from R_k^*, we notice it is affected by W_k^*. W_k^* further at a basal level decomposes as, $\mathrm{Cov}(\nabla \ell_{\theta^*}(X,Y) \mid \mathcal{A}_k)$ + $\mathrm{Cov}(\nabla \ell_{\theta^*}(X, f(X)) \mid \mathcal{A}_k)$ - $2\mathrm{Cov}(\nabla \ell_{\theta^*}(X,Y), \nabla \ell_{\theta^*}(X, f(X)) \mid \mathcal{A}_k)$. We notice R_k^* is small when the model perfectly tracks the variations in Y. Conversely, surviving variance is high when the model and true gradients' are misaligned (third term ≈ 0). The aforementioned allocation strategy prioritizes stratum where such *cancellations* fail. Intuitively, if $Y \approx \hat{Y}$ in stratum k, $\sqrt{R_k^*}$ decreases, implying spending less labeling budget there and relying more on the predictor. While this characterizes the optimal allocation for AdaStrat-ASI's estimator's variance reduction, the quantity R_k^* depends on both Y and $\hat{Y}$ and cannot be estimated without labeled data, making it unsuitable for zero-label scouting. This limitation is not exclusive to AdaStrat-ASI, but applies to any ML-enabled IPW estimator.

Lemma 5.5. *Assuming the loss function is smooth(Ass. 5.1), loss gradient is Lipschitz in y (Ass. 5.2) for any stratum k, and sampling policy enforces a strict lower bound(τ-mixing) on all selection probabilities. Then the square root of asymptotic weighted variance of the loss residuals $\sqrt{R_k^*}$ is bounded by the stratum-specific absolute model error $\sqrt{\epsilon_k}$. i.e.,*

$$\sqrt{R_k^*} = \mathcal{O}\left(\sqrt{\epsilon_k}\right)$$

Crucially, the inaccessibility of R_k^* does not undermine the validity of the algorithm. While R_k^* is a latent oracle quantity requiring access to true labels, Lemma 5.5 shows that it is upper bounded by the model error ϵ_k. This observation does not render the value of R_k^* irrelevant; rather, it shows that its magnitude is controlled by a measurable notion of model fidelity. In regimes where ϵ_k is small—whether assessed through validation data or prior domain knowledge an oracle allocation

would assign negligible labeling budget to that stratum. In fact, in the limit of a perfect local model ($\epsilon_k = 0$), the optimal allocation requires zero human labels for the stratum. Importantly, this collapse is not a smooth limit but a consequence of the discrete nature of budget allocation: once empirical difficulty estimates vanish across all strata ($\epsilon^* = 0$), no human labels are required.

Justification for using f during Scouting Stage. To obtain a practical and analyzable allocation rule, we rely on cheap, model predictions as discussed in Section 4.1. In a stratified regime, $\hat{\sigma}_k^f$ effectively measures the stability of the model's decision-making within a partition. $\hat{\sigma}_k^f$ is inherently designed to be small in stable, homogeneous regions where f behaves consistently. In such stable regions, a reasonably calibrated model is also likely to have consistently low residual variance. Similarly in a heterogenous strata, high proxy gradient variance indicates that the stratum lies along complex model decision boundaries leading to unstable predictions. These are precisely the regions where the underlying task is difficult, implying that $f(x)$ is prone to errors and R_k^* is large. We notice, $\hat{\sigma}_k^f$ and R_k^* are affected similarly under different conditions. While σ_k^f is not a labeled estimate of oracle difficulty R_k^*, it serves as a ranking signal for reallocating a fixed budget toward strata where the predictor's behavior is less stable. Similar gradient based variance heuristics have been used to capture fit difficulty in coreset selection [40] and impact of data on learning [1, 30, 54, 53], aligning well in practice.

Further, we empirically validate this proxy on the California Housing Dataset (introduced later in Section E.2.1) by comparing the oracle allocation $\mathbf{p}^*$ (computed with labels) to our proxy allocation $\hat{\mathbf{p}}$, finding close agreement (Mean Absolute Error (MAE) of 0.0050, Figure 1). Moreover, for budget parity with competing baselines, we reserve the entire human labeling budget for the estimation stage; thus scouting must be zero-label, making cheap model predictions the natural source of information before labeling. To formalize AdaStrat-ASI's efficiency gains, we view ASI's asymptotic variance, through the lens of stratification. This is purely an analytical decomposition, as ASI does not algorithmically rely on stratified sampling. Recall, from Eq. (1) and Eq. (3), ASI and AdaStrat-ASI, differ in the use of sampling rule. The former relies on a global policy for the entire data manifold, however the latter relies on k local policies, one for each stratum. We also know, AdaStrat-ASI enforces fixed local allocation $n_{b_k}^{\text{AdaStrat-ASI}}$ (In previous sections, we refer to this simply as n_{b_k}).

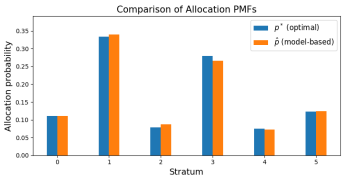


Figure 1: Allocation Comparison

But in ASI, $\pi(x) = \frac{n_b}{n} \cdot \frac{u(x)}{\mathbb{E}[u[X]]} \propto u(x)$. This means, high uncertainty points pull budget toward them strongly and low uncertainty points have a weak pull. Since $\pi(x) \propto u(x)$, this becomes: $\mathbb{E}[n_{b_k}^{\text{ASI}}] \propto \sum_{x \in \mathcal{A}_k} u(x)$, where $n_{b_k}^{\text{ASI}}$ denotes the random number of samples that fall into stratum k after ASI is ran. From this, we can clearly infer: in ASI, the implicit local budget allocation depends on the total uncertainty volume in that stratum. Let $\gamma_k = w_k^2 R_k^*$ denote the weighted oracle difficulty of stratum k.

Theorem 5.6. *Assume the conditions for asymptotic normality of the active M-estimator hold (Theorem 5.3). Let Σ_{ASI} and $\Sigma_{AdaStrat\text{-}ASI}$[1] be the asymptotic covariance matrices for the ASI and AdaStrat-ASI estimators respectively. Then Variance reduction is guaranteed if weighted sum of label reciprocals for ASI is more than AdaStrat-ASI i.e. if,*

$$\sum_{k=1}^{K} \frac{\gamma_k}{\mathbb{E}[n_{b_k}^{ASI}]} > \sum_{k=1}^{K} \frac{\gamma_k}{n_{b_k}^{AdaStrat\text{-}ASI}} \tag{4}$$

then,

$$\text{Tr}(\Sigma_{ASI}) > \text{Tr}(\Sigma_{AdaStrat\text{-}ASI})$$

Theorem 5.6 intuitively shows us, AdaStrat-ASI guarantees variance reduction by *stealing* samples from easy strata and giving them to hard strata. We further empirically study the difference in labelling distribution in Appendix E.9. One notes, if γ_k is constant for all k, then Eq. (4), is never true; the weighted distributions become the same. This aligns with AdaStrat-ASI's mechanism to exploit the heterogeneity across strata to improve efficiency.

[1]We use these notations instead of standard Σ_* and $\Sigma_{\vec{\eta}^*}$, for ease of readability and direct comparison.

The gap in Eq. 4 is large when the model's uncertainty (which drives ASI's allocation $\mathbb{E}[n_{b_k}^{\text{ASI}}]$) is misaligned with the model's true error, which drives AdaStrat-ASI's allocation. For example, if a model is "confidently wrong" in a specific stratum, ASI will starve it of samples, inflating the variance. On the other hand, AdaStrat-ASI will detect high pseudo-gradient fluctuations and will assign an informed budget $n_{b_k}^{\text{AdaStrat-ASI}}$, ensuring robust estimation.

It is also worth noting Eq. 4 is always true if n_{b_k} is the global minimizer of the true objective function as shown in Proposition 5.4. Thus the inequality holds if and only if the proxy Allocation is closer to the Oracle Allocation (see Figure 1). But more importantly, we wish to emphasize the mathematical safety of our scouting heuristic : because AdaStrat-ASI strictly decouples the Scouting stage from the Attacking stage, a failure of the proxy model is entirely benign to the statistical validity of the estimator (Theorem 5.3). That is, even if the scouting phase generates a highly suboptimal allocation vector, the resulting confidence intervals remain statistically valid. The heuristic only governs efficiency (variance reduction; Theorem 5.6), never validity. We provide more details in Appendix D.

6 Experiments

Motivational Synthetic Example. To isolate the effect of stratification on variance reduction, we construct a highly controlled 1D logistic regression setting. We draw $x \sim \mathcal{U}_{[-6,6]}$ and generate labels from a ground-truth model $y \mid x \sim \text{Bernoulli}(\sigma(\theta_1 \cdot x + \theta_0))$, where $(\theta_1, \theta_0) = (1, 0)$. A logistic regression model is fit on 5000 such samples, yielding an acquisition model $\hat{f}(x) = \sigma(\hat{\theta}_1 \cdot x + \hat{\theta}_0)$. We consider the task of estimating the population mean $\mathbb{E}[Y]$ under a deliberately structured test distribution consisting of two Gaussian clusters: $\mathcal{C}_1 : x \sim \mathcal{N}(8, 0.1^2), \quad y = 1, \qquad \mathcal{C}_2 : x \sim \mathcal{N}(-8, 0.1^2), \quad y = 0,$ each containing 5000 points. By construction, the true mean equals 0.5. Both clusters are symmetrically located far from the decision boundary at $x = 0$, and therefore lie in regions of near-deterministic predictions under $\hat{f}$ (Figure 3).

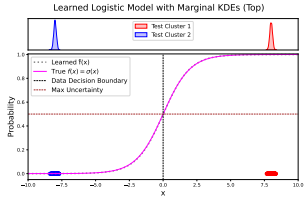


Figure 2: Toy Setup

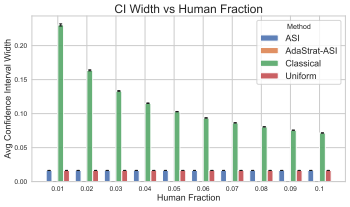


Figure 3: CI width

We notice both clusters lie in regions of low model uncertainty and are symmetrically positioned with respect to the decision boundary. Consequently, model uncertainty based sampling provides no additional signal for prioritization, causing ASI to effectively reduce to uniform sampling. For stratification, we rely on with the data-generating structure. As a result, the empirical variance estimates within each stratum collapse to zero across trials (Figure 4). Although contrived, this example highlights the mechanism by which stratification improves efficiency: when strata align with regions of high black-box model certainty and low outcome variability, stratified ASI can dramatically reduce both empirical variance(and in extreme cases labeling cost). In realistic settings, such perfect alignment is unlikely; nevertheless, this toy example serves to isolate and illustrate the maximal benefit achievable through stratification.

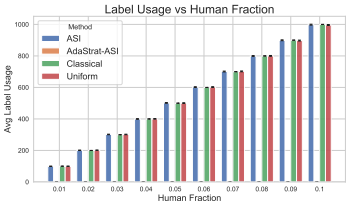


Figure 4: Label Usage

Real world Datasets. We provide experimental details in the Appendix (Section E.1).

AlphaFold Dataset. Inspired by Angelopoulos et al. [2] and Zrnić and Candès [65], we incorporate the AlphaFold dataset($n = 10,802$) sourced from the `ppi_py` package[2] as an application to proteomics. Rather than estimating the odds ratio of phosphorylation as in Angelopoulos et al. [2], Zrnić and Candès [65], we shift the target of inference to the *mean disorder rate* i.e. $\theta^* = \mathbb{E}[Y] = \mathbb{P}$(residue is disordered), to estimate the fraction of PTM-modified residues that are truly intrinsically disordered.

From Figure 5, we see AdaStrat-ASI offering the largest reduction in CI width(approximately 30%(ASI), 32%(Uniform), and 64%(Classical) reduction). AdaStrat-ASI's samples also yield

[2]Github Repository: `https://github.com/aangelopoulos/ppi_py/tree/main`

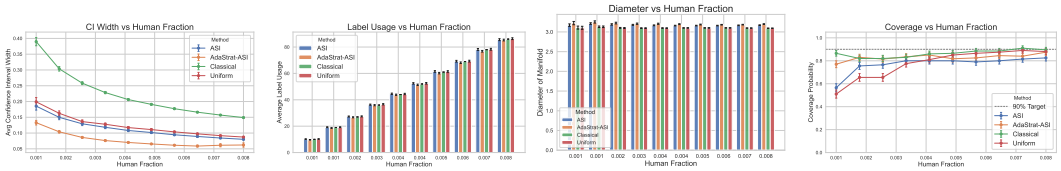


Figure 5: AlphaFold Dataset

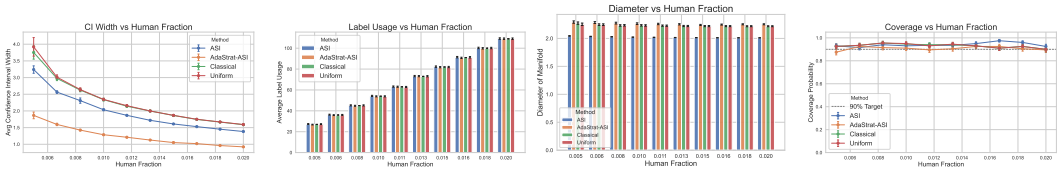


Figure 6: Politeness Dataset

the largest diameter. Interestingly, other ML-powered methods struggle with coverage at tight budgets($n_b < 0.004$). By contrast, AdaStrat-ASI decomposes the uncertainty into stratified sub-regions. This local estimation allows us to more accurately capture the residual variance within each stratum, effectively shielding the CI from becoming biased.

Politeness Dataset. We use the dataset[3] used by Gligoric et al. to gauge politeness of online text requests posted on Stack Exchange and Wikipedia ($n = 5,480$). In this estimation task, θ^* corresponds to the logistic regression coefficient β_{hedge} measuring the impact of a linguistic feature such as hedging on the perceived politeness, $\text{logit}(P(H_{\text{polite}} = 1|X_{\text{hedge}})) = \beta_{\text{hedge}}X_{\text{hedge}} + \beta_0$, where $X_{\text{hedge}} = 1$ indicates the presence of the hedge marker and $H_{\text{polite}} = 1$ indicates annotation as polite.

From Figure 6, all methods achieve coverage close to the nominal 90% level across budgets, indicating that inferential validity is preserved throughout. Among the ML-enabled approaches, AdaStrat-ASI most consistently maintains the best fidelity. Label usage is nearly identical across all methods. In particular, AdaStrat-ASI yields a large reduction in confidence interval width relative to other baselines across all human fractions, with the most pronounced gains at smaller budgets; up to 54% and 45.67% in case of Classical and ASI respectively. AdaStrat-ASI exhibits a modest but consistent increase in diameter indicating broader capture of the data manifold.

These results collectively highlight that, even when uncertainty alone is informative enough to maintain coverage, stratification provides an additional efficiency gain by promoting efficiency (reduced CI). Also, stratifying data first and then applying ASI within each stratum, catapults the budget sampling, causing broader labeled data manifold spread.

7 Conclusion

Limitations and Discussion. A potential pitfall of AdaStrat-ASI, inherited from ASI, is the reliance on well-calibrated estimates of model uncertainty. A poorly-calibrated model may also lead to suboptimal allocation during scouting. On datasets with strata having lack of inter-stratum heterogeneity, AdaStrat-ASI may not be able to reduce CI width(Theorem 5.6). Finally, like ASI, our theoretical guarantees rest on the assumption that data is i.i.d.

We propose Adaptive Stratified Active Statistical Inference (AdaStrat-ASI), a novel inference framework that directs labeling effort toward regions with the greatest impact on the statistical efficiency. By leveraging model uncertainty locally, AdaStrat-ASI achieves more powerful inferences with better spread of labeled data. We provide theoretical guarantees for its asymptotic validity and demonstrate its practical effectiveness across datasets. We believe this work establishes a new principle for efficient data collection, showing that latent structure in data of the statistical problem is as crucial as the uncertainty of the predictive model.

[3]Github Repository: `https://github.com/kristinagligoric/confidence-driven-inference/tree/main/datasets`

References

[1] C. Agarwal, D. D'souza, and S. Hooker. Estimating example difficulty using variance of gradients, 2022. URL https://arxiv.org/abs/2008.11600.

[2] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference, 2023. URL https://arxiv.org/abs/2301.09633.

[3] A. N. Angelopoulos, J. C. Duchi, and T. Zrnic. Ppi++: Efficient prediction-powered inference. 2024. URL https://arxiv.org/abs/2311.01453.

[4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 2025-12-13.

[5] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

[6] A. Carpentier, R. Munos, and A. Antos. Adaptive strategy for stratified monte carlo sampling. *Journal of Machine Learning Research*, 16(68):2231–2271, 2015. URL http://jmlr.org/papers/v16/carpentier15a.html.

[7] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 1995. URL https://projecteuclid.org/journals/statistical-science/volume-10/issue-3/Bayesian-Experimental-Design-A-Review/10.1214/ss/1177009939.full.

[8] Y. Chandak, S. Shankar, V. Syrgkanis, and E. Brunskill. Adaptive instrument design for indirect experiments, 2023. URL https://arxiv.org/abs/2312.02438.

[9] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM, Aug. 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

[10] C.-H. Chiang and H.-y. Lee. Can large language models be an alternative to human evaluations? In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870/.

[11] J. A. Clayton and F. S. Collins. Policy: Nih to balance sex in cell and animal studies. *Nature*, 509(7500):282–283, 2014. URL https://www.nature.com/articles/509282a.

[12] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[13] W. Cochran. *Sampling Techniques*. Wiley Series in Probability and Statistics. Wiley, 1977. ISBN 9780471162407. URL https://books.google.com/books?id=8Y4QAQAAIAAJ.

[14] L. N. Cook, A. Mishler, A. Ramdas, and T. Zrnić. Semiparametric efficient inference from adaptively collected data. In *Proceedings of the Conference on Learning Theory and Representations*, 2024. URL https://openreview.net/forum?id=mfSgIZpwi0.

[15] C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang. Region-based active learning. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2801–2809. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/cortes19a.html.

[16] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 208–215, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390183. URL https://doi.org/10.1145/1390156.1390183.

[17] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):aao5580, 2018. doi: 10.1126/sciadv.aao5580. URL https://www.science.org/doi/10.1126/sciadv.aao5580.

[18] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, Aug. 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

[19] A. Fisch, J. Maynez, R. A. Hofer, B. Dhingra, A. Globerson, and W. W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 111489–111514. Curran Associates, Inc., 2024. doi: 10.52202/079017-3541. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c9fcd02e6445c7dfbad6986abee53d0d-Paper-Conference.pdf.

[20] A. Garivier, E. Kaufmann, and T. Lattimore. On explore-then-commit strategies, 2016. URL https://arxiv.org/abs/1605.08988.

[21] F. Gilardi, M. Alizadeh, and M. Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. URL https://www.pnas.org/doi/10.1073/pnas.2305016120.

[22] K. Gligoric, T. Zrnic, C. Lee, E. Candes, and D. Jurafsky. Can unconfident LLM annotations be used for confident conclusions? In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.179. URL https://aclanthology.org/2025.naacl-long.179/.

[23] J. Hahn, K. Hirano, and D. Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011. doi: 10.1198/jbes.2009.08161. URL https://doi.org/10.1198/jbes.2009.08161.

[24] S. Hanneke. Theory of disagreement-based active learning. *Found. Trends Mach. Learn.*, 7(2–3):131–309, June 2014. ISSN 1935-8237. doi: 10.1561/2200000037. URL https://doi.org/10.1561/2200000037.

[25] D. Hovy and A. Søgaard. Tagging performance correlates with author age. In C. Zong and M. Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2079. URL https://aclanthology.org/P15-2079/.

[26] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, Jan. 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[27] G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015. URL https://doi.org/10.1017/CBO9781139025751.

[28] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. URL https://www.science.org/doi/abs/10.1126/science.aaf7894.

[29] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. URL https://doi.org/10.1038/s41586-021-03819-2.

[30] A. Katharopoulos and F. Fleuret. Not all samples are created equal: Deep learning with importance sampling. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2525–2534. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/katharopoulos18a.html.

[31] M. Kato, T. Ishihara, J. Honda, and Y. Narita. Efficient adaptive experimental design for average treatment effect estimation, 2025. URL https://arxiv.org/abs/2002.05308.

[32] M. Kearns and A. Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019. URL https://dl.acm.org/doi/10.5555/3379082.

[33] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2024. URL https://arxiv.org/abs/2310.08491.

[34] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

[35] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

[36] P. Li, T. Zrnic, and E. Candès. Robust sampling for active statistical inference, 2025. URL https://arxiv.org/abs/2511.08991.

[37] L. Lin, K. Khamaru, and M. J. Wainwright. Semi-parametric inference based on adaptively collected data, 2025. URL https://arxiv.org/abs/2303.02534.

[38] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

[39] S. Lohr. *Sampling: Design and Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2021. ISBN 9781000478235. URL https://books.google.com/books?id=DahGEAAAQBAJ.

[40] S. H. A. Mahmood, M. Yin, and R. Khanna. On the support vector effect in dnns: Rethinking data selection and attribution. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 1020–1031, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709295. URL https://doi.org/10.1145/3690624.3709295.

[41] K. Margatina, G. Vernikos, L. Barrault, and N. Aletras. Active learning by acquiring contrastive examples. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.51. URL https://aclanthology.org/2021.emnlp-main.51/.

[42] J. Miao, X. Miao, Y. Wu, J. Zhao, and Q. Lu. Assumption-lean and data-adaptive post-prediction inference, 2024. URL https://arxiv.org/abs/2311.14220.

[43] P. R. Mohanty. Beyond Disagreement-based Learning for Contextual Bandits. 7 2023. doi: 10.25394/PGS.23739957.v1. URL https://hammer.purdue.edu/articles/thesis/Beyond_Disagreement-based_Learning_for_Contextual_Bandits/23739957.

[44] K. Motwani and D. Witten. Revisiting inference after prediction, 2024. URL https://arxiv.org/abs/2306.13746.

[45] J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934. ISSN 09528385. URL http://www.jstor.org/stable/2342192.

[46] C. Ober, D. A. Loisel, and Y. Gilad. Sex-specific genetic architecture of human disease. *Nature Reviews Genetics*, 9(12):911–922, 2008. URL https://www.nature.com/articles/nrg2415.

[47] OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[48] R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3): 291–297, 1997. URL https://doi.org/10.1016/S0167-7152(96)00140-X.

[49] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

[50] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994. doi: 10.1080/01621459.1994.10476818. URL https://doi.org/10.1080/01621459.1994.10476818.

[51] E. Rolf, J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, 2021. URL https://doi.org/10.1038/s41467-021-24638-z.

[52] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

[53] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, page 1289–1296, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

[54] M. Toneva, A. Sordoni, R. T. des Combes, A. Trischler, Y. Bengio, and G. J. Gordon. An empirical study of example forgetting during deep neural network learning, 2019. URL https://arxiv.org/abs/1812.05159.

[55] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, page 1521–1528, USA, 2011. IEEE Computer Society. ISBN 9781457703942. doi: 10.1109/CVPR.2011.5995347. URL https://doi.org/10.1109/CVPR.2011.5995347.

[56] J. T.-Z. Wei, F. Zufall, and R. Jia. Operationalizing content moderation "accuracy" in the digital services act, 2024. URL https://arxiv.org/abs/2305.09601.

[57] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

[58] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. URL https://dl.acm.org/doi/10.5555/3016387.3016457.

[59] M. Zdun. Machine politics: How america casts and counts its votes. *Reuters*, 2022. URL https://www.reuters.com/world/us/us-midterm-elections-how-america-casts-counts-its-votes-2022-11-02/.

[60] C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning, 2014. URL https://arxiv.org/abs/1407.2657.

[61] K. W. Zhang, L. Janson, and S. Murphy. Statistical inference with m-estimators on adaptively collected data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=TJOQw_vMlAj.

[62] R. Zhang, Y. Li, Y. Ma, M. Zhou, and L. Zou. LLMaAA: Making large language models as active annotators. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=B6Gdg7u04y.

[63] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

[64] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, Mar. 2024. doi: 10.1162/coli_a_00502. URL https://aclanthology.org/2024.cl-1.8/.

[65] T. Zrnić and E. J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html. PMLR v235.

[66] T. Zrnic and E. J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi: 10.1073/pnas.2322083121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2322083121.

Appendix

A Impact Statement

This work advances principled statistical inference under limited labeling budgets by improving use of labeled data and variance efficiency through stratified active sampling. Beyond efficiency gains, the stratified framework enables explicit control over how labeling budgets are allocated across predefined subgroups, allowing practitioners to enforce balance or representational constraints when required. This capability is particularly relevant in settings such as survey research, social science, and AI-assisted data annotation, where subgroup considerations may be mandated for reliability or policy reasons. In the bigger picture, when applied responsibly, such efficiency gains may reduce costs, respondent burden, and reliance on large-scale labeling efforts. As a methodological contribution, the societal impact is indirect and depends on responsible downstream use.

B Algorithm

Algorithm 1 Adaptive Stratified Active Statistical Inference

Input: Unlabeled data partitions $\mathcal{A}_k = \{X_{ik}\}_{i=1}^{n_k}$ for $k \in [K]$; total labeling budget n_b; predictive model f; loss $\ell_\theta(x, y)$; proxy parameter using f $\hat{\theta}$; error level α; strata weight $w_k = \frac{n_k}{n}$.

— Phase 1: Scouting —

1: Compute the global proxy Hessian $\hat{H}_f := \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ell_{\hat{\theta}}(X_i, f(X_i))$

2:

for each stratum $k \in [K]$ **do**

2.1: Compute the proxy Covariance Matrix $\hat{V}_k^f := \mathrm{Cov}\big(\nabla \ell_{\hat{\theta}}(X, f(X)) | X \in \mathcal{A}_k\big)$

2.2 Assemble $\hat{\Sigma}_k := \hat{H}_f^{-1} \hat{V}_k^f \hat{H}_f^{-1}$

2.3: Compute $\hat{\sigma}_k^f := \mathrm{Tr}\Big(\hat{\Sigma}_k\Big)$

end for

— Phase 2: Attacking —

3: For each k, allocate the main budget using *proxy budget allocation weight* based on the scouting results i.e. $(n_{b_1}, \ldots, n_{b_K}) \sim \mathrm{Multinomial}(n_b, \vec{z})$, where $z_k = \frac{w_k \sqrt{\hat{\sigma}_k^f}}{\sum_{j=1}^{K} w_j \sqrt{\hat{\sigma}_j^f}}$

4:

for each stratum $k \in [K]$ **do**

4.1: Compute sampling probabilities for all $x \in \mathcal{A}_k$: $\pi_k(x) = u(x) \cdot \hat{\eta}_k$, $\hat{\eta}_k = \frac{n_{b_k}}{n_k} \cdot \frac{1}{\hat{\mathbb{E}}_{X \sim \mathcal{A}_k}[u(X)]}$

4.2 Optionally: Select $\tau \in (0, 1)$ and choose sampling rule $\pi_k^{(\tau)}(x) = (1-\tau) \cdot \pi_k(x) + \tau \cdot \pi_k^{\mathrm{unif}}$, where $\pi_k^{\mathrm{unif}} = \frac{n_{b_k}}{n_k}$.

4.3 Sample labeling decisions: $\xi_{i,k} \sim \mathrm{Bern}(\pi_k^{(\tau)}(X_i)), \quad i \in [n_k]$

4.4 Collect labels $\{Y_i : \xi_{i,k} = 1\}$

end for

— Phase 3: Final Estimation —

5: Compute stratified active estimator $(\hat{\theta}^{\pi_{\vec{\eta}}})$ (minimizer of Eq. (3))

C Proofs

Claim C.1 (Consistency of $\hat{\vec{\eta}}$ for AdaStrat-ASI)**.** *Suppose that for each stratum $k \in [K]$, the local budget fraction converges: $n_{b_k}/n_k \xrightarrow{p} q_k \in (0, 1)$. If $\mathcal{H}$ is discrete and for every k there is no $\eta \in \mathcal{H}$ such that $\eta \mathbb{E}[u(X) \mid X \in \mathcal{A}_k] = q_k$ exactly, then there exists a unique optimal vector $\vec{\eta}^* = (\eta_1^*, \ldots, \eta_K^*) \in \mathcal{H}^K$ such that:*

$$\mathbb{P}(\hat{\vec{\eta}} \neq \vec{\eta}^*) \to 0$$

Proof. Proof of Claim C.1. For any fixed stratum k, the estimated parameter $\hat{\eta}_k$ is defined as:

$$\hat{\eta}_k = \max\left\{\eta \in \mathcal{H} : \eta \frac{1}{n_k} \sum_{i \in \mathcal{A}_k} u(X_i) \leq \frac{n_{b_k}}{n_k}\right\}.$$

This condition is equivalent to:

$$\eta \leq \frac{n_{b_k}/n_k}{\frac{1}{n_k}\sum_{i \in \mathcal{A}_k} u(X_i)}.$$

In the numerator (by assumption), $\frac{n_{b_k}}{n_k} \xrightarrow{p} q_k$. In the denominator, by Law of Large Numbers (within stratum k), the empirical average converges in probability to the population expectation: $\frac{1}{n_k}\sum_{i \in \mathcal{A}_k} u(X_i) \xrightarrow{p} \mathbb{E}[u(X) \mid X \in \mathcal{A}_k]$.

Applying the Continuous Mapping Theorem,

$$\frac{n_{b_k}/n_k}{\frac{1}{n_k}\sum_{i \in \mathcal{A}_k} u(X_i)} \xrightarrow{p} \frac{q_k}{\mathbb{E}[u(X) \mid X \in \mathcal{A}_k]}.$$

Let the gap to the nearest incorrect parameter be:

$$\varepsilon_k = \min_{\eta \in \mathcal{H}} \left|\eta - \frac{q_k}{\mathbb{E}[u(X) \mid \mathcal{A}_k]}\right|.$$

By the assumption that no η exactly equals the target ratio, we have $\varepsilon_k > 0$. Define the true optimal parameter for stratum k as:

$$\eta_k^* = \max\left\{\eta \in \mathcal{H} : \eta \mathbb{E}[u(X) \mid \mathcal{A}_k] \leq q_k\right\}.$$

On the event E_k that the empirical ratio $\frac{n_{b_k}/n_k}{\hat{\mathbb{E}}_k[u]}$ is within distance ε_k of the population ratio, we must have $\hat{\eta}_k = \eta_k^*$. Convergence in probability implies $\mathbb{P}(E_k^c) \to 0$ as $n_k \to \infty$. Since K is finite, we apply a union bound over all strata:

$$\mathbb{P}(\hat{\vec{\eta}} \neq \vec{\eta}^*) = \mathbb{P}\left(\bigcup_{k=1}^{K} \{\hat{\eta}_k \neq \eta_k^*\}\right) \leq \sum_{k=1}^{K} \mathbb{P}(\hat{\eta}_k \neq \eta_k^*) \to 0.$$

Thus, the estimated parameter vector converges exactly to the optimal vector with probability approaching 1. □

Proof for Theorem 5.3. We assume the smoothness conditions of Assumption 5.1 hold. Let $n_k = \sum_{i=1}^{n} \mathbb{I}(X_i \in A_k)$ be the random number of unlabeled samples falling into stratum k. Because the total pool of n samples is drawn i.i.d. from the marginal distribution P_X, $n_k \sim \text{Binomial}(n, w_k)$. By the Weak Law of Large Numbers, $n_k/n \xrightarrow{p} w_k$.

The AdaStrat-ASI estimator $\hat{\theta}^{\hat{\vec{\eta}}}$ is defined as the minimizer of the post-stratified empirical risk (Equation 3):

$$L_n^{\hat{\vec{\eta}}}(\theta) := \sum_{k=1}^{K} w_k \left(\frac{1}{n_k} \sum_{i \in A_k} L_{\theta,i}^{\hat{\eta}_k}\right)$$

where $L_{\theta,i}^{\eta_k} = \ell_\theta(X_i, f(X_i)) + (\ell_\theta(X_i, Y_i) - \ell_\theta(X_i, f(X_i)))\frac{\xi_i}{\pi_{\eta_k}(X_i)}$.

Let $\vec{\eta} = (\eta_1, \ldots, \eta_K) \in \mathcal{H}^K$ parameterize the piecewise sampling policy $\pi_{\vec{\eta}}(X_i) = \sum_{k=1}^{K} \mathbb{I}(X_i \in A_k)\pi_{\eta_k}(X_i)$. For a given $\vec{\eta}$, we define the single-sample surrogate loss as $L_{\theta,i}^{\vec{\eta}} = \ell_\theta(X_i, f(X_i)) + (\ell_\theta(X_i, Y_i) - \ell_\theta(X_i, f(X_i)))\frac{\xi_i^{\vec{\eta}}}{\pi_{\vec{\eta}}(X_i)}$. We define the stratified empirical expectation and empirical process operators over the full sample of size n:

$$\mathbb{E}_n[g(L_\theta^{\vec{\eta}})] := \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in A_k} g(L_{\theta,i}^{\vec{\eta}}) \quad ; \quad \mathbb{G}_n[g(L_\theta^{\vec{\eta}})] := \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i \in A_k}\left(g(L_{\theta,i}^{\vec{\eta}}) - \mathbb{E}[g(L_{\theta,i}^{\vec{\eta}}) \mid A_k]\right)$$

Notice that $\mathbb{E}_n[L_\theta^{\hat{\vec{\eta}}}]$ exactly matches our stratified empirical risk $L^{\pi_{\vec{\eta}}}(\theta)$ from Eq. (3).

From C.1, we show uniform consistency of the vector-valued Split ($\hat{\vec{\eta}}$). Because $\vec{\eta}^*$ is fixed, we can apply the standard empirical process bounds to the global risk. By the differentiability and local Lipschitzness of the loss, for any $h_n = O_P(1)$ we have stochastic equicontinuity:

$$\mathbb{G}_n \left[\sqrt{n}(L^{\vec{\eta}^*}_{\theta^*+h_n/\sqrt{n}} - L^{\vec{\eta}^*}_{\theta^*}) - h_n^\top \nabla L^{\vec{\eta}^*}_{\theta^*} \right] \xrightarrow{p} 0$$

By definition, this is equivalent to:

$$n\mathbb{E}_n[L^{\vec{\eta}^*}_{\theta^*+h_n/\sqrt{n}} - L^{\vec{\eta}^*}_{\theta^*}] = n(L(\theta^* + h_n/\sqrt{n}) - L(\theta^*)) + h_n^\top \mathbb{G}_n[\nabla L^{\vec{\eta}^*}_{\theta^*}] + o_P(1)$$

A second-order Taylor expansion around the population minimizer θ^* yields:

$$n\mathbb{E}_n[L^{\vec{\eta}^*}_{\theta^*+h_n/\sqrt{n}} - L^{\vec{\eta}^*}_{\theta^*}] = \frac{1}{2} h_n^\top H_{\theta^*} h_n + h_n^\top \mathbb{G}_n[\nabla L^{\vec{\eta}^*}_{\theta^*}] + o_P(1)$$

Crucially, because $\mathbb{P}(\hat{\vec{\eta}} \neq \vec{\eta}^*) \to 0$, we can substitute the empirical estimate $\hat{\vec{\eta}}$ without altering the asymptotic behavior:

$$n\mathbb{E}_n[L^{\hat{\vec{\eta}}}_{\theta^*+h_n/\sqrt{n}} - L^{\hat{\vec{\eta}}}_{\theta^*}] = \frac{1}{2} h_n^\top H_{\theta^*} h_n + h_n^\top \mathbb{G}_n[\nabla L^{\vec{\eta}^*}_{\theta^*}] + o_P(1)$$

We evaluate the previous display at $h_n = \hat{h}_n := \sqrt{n}(\hat{\theta}^{\hat{\vec{\eta}}} - \theta^*)$ (which is $O_P(1)$ by consistency) and at $h_n = \tilde{h}_n := -H_{\theta^*}^{-1} \mathbb{G}_n[\nabla L^{\vec{\eta}^*}_{\theta^*}]$. Following the algebraic manipulation from ASI:

$$\frac{1}{2} \left(\sqrt{n}(\hat{\theta}^{\hat{\vec{\eta}}} - \theta^*) - \tilde{h}_n \right)^\top H_{\theta^*} \left(\sqrt{n}(\hat{\theta}^{\hat{\vec{\eta}}} - \theta^*) - \tilde{h}_n \right) + o_P(1) \leq 0$$

Since $H_{\theta^*} \succ 0$, it must be that $\sqrt{n}(\hat{\theta}^{\hat{\vec{\eta}}} - \theta^*) - \tilde{h}_n \xrightarrow{p} 0$.

We now analyze the gradient vector $\sqrt{n}\nabla L_n^{\vec{\eta}^*}(\theta^*)$, accounting for the random n_k:

$$\sqrt{n}\nabla L_n^{\vec{\eta}^*}(\theta^*) = \sum_{k=1}^K w_k \sqrt{n} \left(\frac{1}{n_k} \sum_{i \in A_k} \nabla L^{\eta_k^*}_{\theta^*,i} \right)$$

Let $\mu_k = \mathbb{E}[\nabla L^{\eta_k^*}_{\theta^*,i} \mid A_k]$. We rewrite the scaled gradient by multiplying and dividing by $\sqrt{n_k}$:

$$\sqrt{n}\nabla L_n^{\vec{\eta}^*}(\theta^*) = \sum_{k=1}^K w_k \sqrt{\frac{n}{n_k}} \left(\frac{1}{\sqrt{n_k}} \sum_{i \in A_k} (\nabla L^{\eta_k^*}_{\theta^*,i} - \mu_k) \right)$$

On expanding,

$$\sqrt{n}\nabla L_n^{\vec{\eta}^*}(\theta^*) = \sum_{k=1}^K w_k \sqrt{\frac{n}{n_k}} \left(\frac{1}{\sqrt{n_k}} \sum_{i \in A_k} (\nabla L^{\eta_k^*}_{\theta^*,i} - \mu_k) \right) + \sqrt{n} \sum_{k=1}^K w_k \mu_k$$

Because the estimator deterministically multiplies by the true population weights w_k, this sum exactly perfectly matches the population first-order condition $\sum_{k=1}^K w_k \mu_k = 0$. Therefore, the entire drift term vanishes algebraically before any variance is taken:

$$\sqrt{n} \sum_{k=1}^K w_k \mu_k = \sqrt{n}(0) = 0$$

Because this deterministic drift is forced to zero for every finite sample n, it cannot contribute to the asymptotic variance. The variance of the remaining sum depends strictly on the centered within-stratum gradients.

Now, by the standard Central Limit Theorem within each stratum,

$$Z_k := \frac{1}{\sqrt{n_k}} \sum_{i \in A_k} (\nabla L^{\eta_k^*}_{\theta^*,i} - \mu_k) \xrightarrow{d} \mathcal{N}(0, V_k)$$

, where $V_k = \text{Var}(\nabla L^{\eta_k^*}_{\theta^*,i} \mid X \in A_k)$.

By Slutsky's theorem, since $\sqrt{n/n_k} \xrightarrow{p} 1/\sqrt{w_k}$, we have $w_k\sqrt{n/n_k}Z_k \xrightarrow{d} \sqrt{w_k}\mathcal{N}(0, V_k) \equiv \mathcal{N}(0, w_k V_k)$.

Because the samples are independent across strata, the sum converges to a normal distribution with the sum of the covariances:

$$\sqrt{n}\nabla L_n^{\vec{\eta}^*}(\theta^*) \xrightarrow{d} \mathcal{N}\left(0, \sum_{k=1}^K w_k V_k\right)$$

Applying this to the Taylor expansion yields $\sqrt{n}(\hat{\theta}^{\hat{\vec{\eta}}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\vec{\eta}^*})$, where the covariance is exactly the weighted sum of the within-stratum variances:

$$\Sigma_{\vec{\eta}^*} = H_{\theta^*}^{-1}\left(\sum_{k=1}^K w_k \text{Var}\left(\nabla \ell_{\theta^*}(X, f(X)) + \frac{\xi^{\eta_k^*}}{\pi_{\eta_k^*}(X)}\Delta \mid A_k\right)\right) H_{\theta^*}^{-1}$$

where $\Delta = \nabla \ell_{\theta^*}(X, Y) - \nabla \ell_{\theta^*}(X, f(X))$. The final statement follows by Slutsky's theorem. □

Lemma C.2. *Assume standard M-estimation regularity from Theorem 5.3. In Algorithm 1 the AdaStrat-ASI estimator's asymptotic variance's trace admits the decomposition*

$$\text{Tr}\big(\Sigma_{\vec{\eta}^*}\big) = C + n\sum_{k=1}^K \frac{w_k^2}{n_{b_k}} R_k^*, \tag{5}$$

where C does not depend on $(n_{b_1}, \ldots, n_{b_K})$, and

$$R_k^* := \text{Tr}\big(H_{\theta^*}^{-1} W_k^* H_{\theta^*}^{-1}\big), \qquad W_k^* := \mathbb{E}\left[\frac{1}{s_k(X)}\Delta\Delta^\top \;\middle|\; X \in \mathcal{A}_k\right]. \tag{6}$$

, where,

$$\Delta := \nabla \ell_{\theta^*}(X, Y) - \nabla \ell_{\theta^*}(X, f(X)).$$

Proof. We directly start from result of Theorem 5.3

$$\Sigma_{\vec{\eta}^*} = H_{\theta^*}^{-1}\left(\sum_{k=1}^K w_k V_k\right) H_{\theta^*}^{-1}$$

We now decompose V_k. By the law of total variance conditioned on $X \in A_k$, and using $\xi_i \sim \text{Bernoulli}(\pi_{\eta_k^*}(X_i))$, we separate the variance of the true gradient from the variance induced by active sampling:

$$V_k = \text{Var}(\nabla \ell_{\theta^*}(X, Y) \mid A_k) + \mathbb{E}\left[\left(\frac{1}{\pi_{\eta_k^*}(X)} - 1\right)\Delta\Delta^\top \;\middle|\; A_k\right]$$

By our active sampling design, the policy inside stratum k is $\pi_{\eta_k^*}(X) = \frac{n_{b,k}}{n_k} s_k(X)$. Substituting this yields:

$$V_k = \text{Var}(\nabla \ell_{\theta^*}(X, Y) \mid A_k) - \mathbb{E}[\Delta\Delta^\top \mid A_k] + \frac{n_k}{n_{b,k}}\mathbb{E}\left[\frac{1}{s_k(X)}\Delta\Delta^\top \;\middle|\; A_k\right]$$

Asymptotically, $n_k/n \to w_k$, so the scaling factor becomes $\frac{n_k}{n_{b,k}} \to \frac{nw_k}{n_{b,k}}$. The limiting within-stratum variance becomes:

$$V_k = \tilde{C}_k + \frac{nw_k}{n_{b,k}} W_k^*$$

where $\tilde{C}_k = \text{Var}(\nabla \ell_{\theta^*}(X, Y) \mid A_k) - \mathbb{E}[\Delta\Delta^\top \mid A_k]$ is the irreducible variance independent of the active labeling budget. Finally, we substitute this back into the global covariance matrix and apply the trace operator $\text{Tr}(\cdot)$:

$$\text{Tr}(\Sigma_{\vec{\eta}^*}) = \sum_{k=1}^{K} w_k \text{Tr}(H_{\theta^*}^{-1} \tilde{C}_k H_{\theta^*}^{-1}) + n \sum_{k=1}^{K} \frac{w_k^2}{n_{b,k}} \text{Tr}(H_{\theta^*}^{-1} W_k^* H_{\theta^*}^{-1})$$

Let $C = \sum_{k=1}^{K} w_k \text{Tr}(H_{\theta^*}^{-1} \tilde{C}_k H_{\theta^*}^{-1})$. Substituting the definition $R_k^* := \text{Tr}(H_{\theta^*}^{-1} W_k^* H_{\theta^*}^{-1})$, we obtain the final decomposition:

$$\text{Tr}(\Sigma_{\vec{\eta}^*}) = C + n \sum_{k=1}^{K} \frac{w_k^2}{n_{b,k}} R_k^*$$

This completes the proof.

□

Proof for Proposition 5.4.

Proof. Under Lemma C.2, we formulate the constrained optimization problem using the method of Lagrange multipliers. The Lagrangian $\mathcal{L}$ is defined with respect to the decision variables n_{b_k}: $\mathcal{L}(n_{b_1}, \ldots, n_{b_K}, \lambda) = n \sum_{k=1}^{K} \frac{w_k^2 R_k^*}{n_{b_k}} + \lambda \left(\sum_{k=1}^{K} n_{b_k} - n_b \right)$ where λ is the Lagrange multiplier enforcing the budget constraint. To find the critical points, we take the partial derivative with respect to each n_{b_k} and set it to zero: $\frac{\partial \mathcal{L}}{\partial n_{b_k}} = -\frac{n w_k^2 R_k^*}{n_{b_k}^2} + \lambda = 0$ Rearranging the terms to isolate the decision variable n_{b_k}: $n_{b_k}^2 = \frac{n w_k^2 R_k^*}{\lambda} \implies n_{b_k} = \sqrt{n} \frac{w_k \sqrt{R_k^*}}{\sqrt{\lambda}}$ This relation holds for all k. We determine the constant λ by summing n_{b_k} over all strata and applying the budget constraint: $\sum_{k=1}^{K} n_{b_k} = \sum_{k=1}^{K} \sqrt{n} \frac{w_k \sqrt{R_k^*}}{\sqrt{\lambda}} = \frac{\sqrt{n}}{\sqrt{\lambda}} \sum_{k=1}^{K} w_k \sqrt{R_k^*} = n_b$. Solving for the scaling factor $\frac{\sqrt{n}}{\sqrt{\lambda}}$: $\frac{\sqrt{n}}{\sqrt{\lambda}} = \frac{n_b}{\sum_{j=1}^{K} w_j \sqrt{R_j^*}}$. Finally, substituting this back into the expression for n_{b_k} yields the optimal allocation rule: $n_{b_k}^* = w_k \sqrt{R_k^*} \left(\frac{n_b}{\sum_{j=1}^{K} w_j \sqrt{R_j^*}} \right)$

One further notices, the objective is strictly convex in n_{b_k} over $n_{b_k} > 0$ because each term has a second derivative($\frac{2 n w_k R_k^*}{n_{b_k}^3} > 0$). So any critical point is the unique global minimizer.

□

Proof. *Proof* for Lemma 5.5.

By definition, $\epsilon_k := \mathbb{E}\big[(Y - f(X))^2 \mid X \in \mathcal{A}_k\big]$.

Recall our notations,

$H_{\theta^*} := \mathbb{E}[\nabla^2 \ell_{\theta^*}(X, Y)]$, $\Delta := \nabla \ell_{\theta^*}(X, Y) - \nabla \ell_{\theta^*}(X, f(X))$, $W_k^* := \mathbb{E}\Big[\frac{1}{s_k(X)} \Delta\Delta^\top \,\Big|\, X \in \mathcal{A}_k \Big]$, $R_k^* := \text{Tr}\big(H_{\theta^*}^{-1} W_k^* H_{\theta^*}^{-1}\big)$

We know, $\pi_k(x) = \frac{n_{b_k}}{n_k} s_k(x)$. On τ-mixing,

$$\begin{aligned} \pi_k^\tau(x) &= \frac{n_{b_k}}{n_k} s_k(x)(1 - \tau) + \frac{n_{b_k}}{n_k} \tau \\ &= \frac{n_{b_k}}{n_k} (s_k(x)(1 - \tau) + \tau) \\ &= \frac{n_{b_k}}{n_k} (s_k^\tau(x)) \end{aligned}$$

This implies $\forall x$,

$$s_k^\tau(x) > \tau \tag{7}$$

Also, from Assumption 5.2, gradient is Lipschitz in the target y: there exists a $L_y(X)$ with $\mathbb{E}[L_y(X)^2 \mid \mathcal{A}_k] \leq$
$L_{y,k}^2 < \infty$ such that for all y, y',

$$\|\nabla \ell_{\theta^*}(X, y) - \nabla \ell_{\theta^*}(X, y')\| \leq L_y(X)\,|y - y'|$$

Thus, by Lipschitz-in-y, $\|\Delta\| \leq L_y(X)\,|Y - f(X)|$.

Squaring and taking stratumwise expectation,

$$\begin{aligned}\mathbb{E}[\|\Delta\|^2 \mid \mathcal{A}_k] &\leq \mathbb{E}[L_y(X)^2 (Y - f(X))^2 \mid \mathcal{A}_k] \\ &\leq L_{y,k}^2\, \mathbb{E}[(Y - f(X))^2 \mid \mathcal{A}_k] \\ &= L_{y,k}^2\, \epsilon_k \end{aligned} \tag{8}$$

Now,

$$\begin{aligned}\|W_k^*\|_{\text{op}} &\leq \mathbb{E}\left[\frac{1}{s_k^\tau(X)} \|\Delta\Delta^\top\|_{\text{op}} \mid \mathcal{A}_k\right] \\ &= \mathbb{E}\left[\frac{1}{s_k^\tau(X)} \|\Delta\|^2 \mid \mathcal{A}_k\right] \\ &\leq \frac{1}{\tau}\mathbb{E}[\|\Delta\|^2 \mid \mathcal{A}_k] \\ &\leq \frac{L_{y,k}^2}{\tau}\epsilon_k.\end{aligned}$$

Third line and fourth line are true because of Eq. 7 and Eq. 8 respectively.

Finally, assuming $\|H_{\theta^*}^{-1}\|_F < \infty$,

$$R_k^* = \text{Tr}(H_{\theta^*}^{-1} W_k^* H_{\theta^*}^{-1}) \leq \|H_{\theta^*}^{-1}\|_F^2\, \|W_k^*\|_{\text{op}} \leq \frac{\|H_{\theta^*}^{-1}\|_F^2}{\tau}\, L_{y,k}^2\, \epsilon_k$$

Since R_k^* and ϵ_k are non negative, for all k

$$\sqrt{R_k^*} = \mathcal{O}(\sqrt{\epsilon_k})$$

□

Proof. Proof for Theorem 5.6.

The idea of strata exists explicitly and implicitly in AdaStrat-ASI and ASI respectively; AdaStrat-ASI uses pre-stratification((3)), whereas we analyze ASI under a post-stratification((1)) lens under the same strata composition.

Let $n_k^{\text{ASI}} := \sum_{i=1}^n \mathbf{1}\{X_i \in A_k\}$ be the random number of samples that fall into stratum k, such that $\sum n_k^{\text{ASI}} = n$. Let $\hat{w}_k = \frac{n_k^{\text{ASI}}}{n}$ be the empirical weight. Because the total pool is drawn i.i.d., the vector of counts follows a Multinomial distribution: $(n_1^{\text{ASI}}, \ldots, n_K^{\text{ASI}}) \sim \text{Multinomial}(n, (w_1, \ldots, w_K))$. By the properties of the Multinomial distribution, the empirical weights $\hat{w}_k$ fluctuate around the true weights w_k. By the Central Limit Theorem, the scaled fluctuations converge to a Normal distribution:

$$\sqrt{n}(\hat{w} - w) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\text{multinomial}})$$

where the covariance matrix of the weights has entries $\text{Cov}(\hat{w}_j, \hat{w}_k) = w_j(1 - w_j)$ if $j = k$, and $-w_j w_k$ if $j \neq k$.

Let $\mu_k = \mathbb{E}[\nabla L_{\theta^*,i}^{\eta_k^*} \mid A_k]$ be the true mean gradient in stratum k. The population first-order optimality condition guarantees that exactly $\sum_{k=1}^K w_k \mu_k = 0$.

Standard ASI(Equation 1) computes the unweighted global average. We scale it by $\sqrt{n}$ to find its asymptotic distribution:

$$\sqrt{n}\nabla L_n^{\text{ASI}}(\theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla L_i$$

We can group this sum into the K strata, and center the inner sums around μ_k:

$$\sqrt{n}\nabla L_n^{\text{ASI}}(\theta^*) = \sum_{k=1}^{K} \frac{1}{\sqrt{n}} \sum_{i \in A_k} (\nabla L_i - \mu_k) + \sum_{k=1}^{K} \frac{1}{\sqrt{n}} \sum_{i \in A_k} \mu_k$$

On closely inspecting the second term, the sum $\sum_{i \in A_k} \mu_k$ is just $n_k^{\text{ASI}} \mu_k$.

$$\text{Second Term} = \sum_{k=1}^{K} \frac{n_k^{\text{ASI}}}{\sqrt{n}} \mu_k = \sqrt{n} \sum_{k=1}^{K} \left(\frac{n_k^{\text{ASI}}}{n} \right) \mu_k = \sqrt{n} \sum_{k=1}^{K} \hat{w}_k \mu_k$$

Because $\sum_{k=1}^{K} w_k \mu_k = 0$, we can subtract it from this term without changing its value:

$$\text{Second Term} = \sqrt{n} \sum_{k=1}^{K} (\hat{w}_k - w_k) \mu_k$$

This term represents the error caused by the random sample sizes (n_k). Because $\sqrt{n}(\hat{w} - w)$ converges to the Multinomial covariance matrix, the variance of this linear combination is:

$$\text{Var}\left(\sum_{k=1}^{K} \sqrt{n}(\hat{w}_k - w_k) \mu_k \right) = \sum_{k=1}^{K} w_k \mu_k \mu_k^\top - \left(\sum_{k=1}^{K} w_k \mu_k \right) \left(\sum_{k=1}^{K} w_k \mu_k \right)^\top$$

Because the second bracket is strictly 0 (the first-order condition), the variance of the second term is exactly:

$$\sum_{k=1}^{K} w_k \mu_k \mu_k^\top \equiv \Delta_{\text{Bias}} > 0$$

Thus, ASI's empirical gradient contains the term $\sqrt{n} \sum \hat{w}_k \mu_k$. Because $\hat{w}_k$ is a random variable, this term fluctuates, injecting exactly Δ_{Bias} into the final asymptotic covariance matrix. By forcing the estimator to use w_k instead of the empirical $\hat{w}_k$, AdaStrat-ASI algebraically destroys the multinomial fluctuation term before the variance is taken.

Now, performing a symmetrical k-stratum decomposition analysis on ASI's asymptotic variance as seen in Proof of Lemma C.2,

$$\begin{aligned} \text{Tr}(\Sigma_{\text{ASI}}) &= \Delta_{\text{Bias}} + T(\sum_{k=1}^{K} w_k (C_k + n_k \mathbb{E}\left[\frac{1}{n_{b_k}^{\text{ASI}}} \right] \mathbb{E}\left[\frac{\Delta \Delta^\top}{s_k(X)} \;\middle|\; Z = k \right])) \\ &= \Delta_{\text{Bias}} + C + n \sum_{k=1}^{K} w_k^2 R_k^* \mathbb{E}\left[\frac{1}{n_{b_k}^{\text{ASI}}} \right] \end{aligned} \tag{9}$$

Subtracting Eq. (5) from Eq. (9),

$$\begin{aligned} \text{Gap} &= \text{Tr}(\Sigma_{\text{ASI}}) - \text{Tr}(\Sigma_{\text{AdaStrat-ASI}}) \\ &= \underbrace{\Delta_{\text{Bias}}}_{\text{Stability Gap(Term 1)}} + n \cdot \underbrace{\sum_{k=1}^{K} w_k^2 R_k^* \left(\mathbb{E}\left[\frac{1}{n_{b_k}^{\text{ASI}}} \right] - \frac{1}{n_{b_k}^{\text{AdaStrat-ASI}}} \right)}_{\text{Efficiency Gap(Term 2)}} \end{aligned}$$

The gap shown above reveals that AdaStrat-ASI reduces asymptotic variance through two distinct mechanisms: *stability* (via stratification) and *efficiency* (via allocation). We analyze the conditions under which the gap $\text{Tr}(\Sigma_{\text{ASI}}) - \text{Tr}(\Sigma_{\text{AdaStrat-ASI}})$ is strictly positive.

1. Term 1: The first term compares reduction of between-stratum variance. Since Δ_{Bias} is the trace of a quadratic form involving a covariance matrix, it is strictly non-negative.

2. Term 2: The second term compares the allocation strategies. By using Jensen's inequality, on convex function $f(x) = 1/x$, we have:

$$\sum_{k=1}^{K} \gamma_k \mathbb{E}\left[\frac{1}{n_{b_k}^{\text{ASI}}}\right] \geq \underbrace{\sum_{k=1}^{K} \frac{\gamma_k}{\mathbb{E}[n_{b_k}^{\text{ASI}}]}}_{\text{(Fixed Average)}}$$

The above inequality implies that removing the randomness of the sample counts (fixing $n_{b_k}^{\text{ASI}}$ to its mean) reduces variance.

But it is given that,

$$\sum_{k=1}^{K} \frac{\gamma_k}{\mathbb{E}[n_{b_k}^{\text{ASI}}]} \geq \sum_{k=1}^{K} \frac{\gamma_k}{n_{b_k}^{\text{AdaStrat-ASI}}}$$

This implies,

$$\text{Tr}(\Sigma_{\text{ASI}}) > \text{Tr}(\Sigma_{\text{AdaStrat-ASI}})$$

Hence proved. □

Lemma C.3. *Assuming non-degenerate difficulty scores $0 < \hat{\sigma}_k^f < \infty \quad \forall k$. In Algorithm 1, as the model becomes perfect ($\epsilon^* \to 0$), the asymptotic variance of AdaStrat-ASI converges to the irreducible variance C at a rate of $\mathcal{O}(\epsilon^*)$.*

Proof.

$$\begin{aligned}
\text{Tr}(\Sigma_{\vec{\eta}^*}) &= C + n \sum_{k=1}^{K} \frac{w_k^2}{n_{b_k}} R_k^* \\
&= C + n \sum_{k=1}^{K} \frac{w_k^2}{n_b \left(\frac{w_k \sqrt{\hat{\sigma}_k^f}}{C'}\right)} R_k^* \qquad \left(C' := \sum_{j=1}^{K} w_j \sqrt{\hat{\sigma}_j^f}\right) \\
&= C + \frac{nC'}{n_b} \sum_{k=1}^{K} w_k \frac{R_k^*}{\sqrt{\hat{\sigma}_k^f}} \\
&= C + \mathcal{O}(\sum_{k=1}^{K} w_k \frac{\epsilon_k}{\sqrt{\hat{\sigma}_k^f}}). \\
&= C + \mathcal{O}(\epsilon^*)
\end{aligned}$$

The first line follows from Lemma C.2. The second line is true asymptotically.[4] The fourth line follows from Lemma 5.5.

This also confirms, the variance is bounded by the weighted average of the Noise-to-Signal ratio. □

D Validity of Inference

Our AdaStrat-ASI estimator (Eq.(3)) depends on the attacking-phase sampling probabilities $\pi_k(x)$ only through their realized values in the denominator. When we write:

$$\pi_k(x) = \frac{n_{b_k}}{n_k} \cdot \frac{u(x)}{\mathbb{E}[u(X) \mid A_k]}$$

the budget fraction n_{b_k}/n_k enters as a deterministic scaling constant in the asymptotic limit (by assumption: $n_{b_k}/n_k \to q_k \in (0, 1)$; see Claim C.1). Our proof for Theorem 5.3 does not depend

[4] Since the budget allocation follows a multinomial distribution, $n_{b_k}/n_b \to z_k$ almost surely in expectation. Thus, for the asymptotic variance analysis, we treat the allocation as deterministic: $n_{b_k} = n_b z_k$.

whether this q_k comes from Oracle Neyman allocation ($\propto \sqrt{R_k^*}$), informed empirical assistance(i.e. Scouting Stage), random allocation (z_k arbitrary) or failed scouting (any convergent sequence); the structure of the asymptotic variance formula remains identical. Fundamentally we care about CLT holding within each fixed strata: $n_k \to \infty$(consequence of $n \to \infty$) and that the sampling probabilities $\pi_{\eta_k^*}(X_i)$ are bounded away from zero. We only care about the ratio $\sqrt{n/n_k} \to 1/\sqrt{w_k}$(by Slutsky's Theorem) based on the stratum prevalence, not the budget allocation mechanism. Overall, the numerical value of q_k changes, affecting efficiency, not validity of our method.

E Experiments

E.1 Training Details

For AlphaFold Dataset, each observation corresponds to a post-translationally modified protein residue, where the binary response $Y_i \in \{0, 1\}$ indicates whether residue i belongs to an experimentally validated intrinsically disordered region (IDR). Crucially, AlphaFold2's pLDDT-based disorder predictions $\hat{Y}_i \in [0, 1]$ are provided directly by `ppi_py`, requiring no additional proxy model training. However for the other real-world datasets, we follow a consistent sample-splitting protocol as in Zrnić and Candès [65]. Specifically, 50% of the data is used to learn models that guide annotation allocation—including the predictive model in ASI(XGBoost [9]) and the auxiliary error model used to calibrate LLM(GPT-4o [47]) confidence while statistical inference is performed exclusively on the remaining half. To clarify, for the latter, the LLM itself is not trained on the data, the error model that maps verbalized confidence scores to squared annotation error is learned on the first half and evaluated only on the held-out half(CCI; Section 3.2 from Gligoric et al. [22]). This separation preserves the inferential validity guarantees of ASI. For simplicity sake, we report confidence-intervals(CIs) and coverage for the first parameter with $\alpha = 0.1$. We repeat our analysis for 500 trials. All experiments were run on a single RTX A6000.

E.2 Additional Real Data Experiment

E.2.1 California Housing Dataset

For this setup, we use the California Housing dataset ($n = 20640$) [48] with the goal to predict the median house value for California districts(expressed in hundreds of thousands of dollars), using demography, location, and general information, while minimizing over squared loss.

From Figure 7, all methods achieve coverage close to the nominal 90% level, confirming inferential validity. Notably, among ML-enabled approaches, AdaStrat-ASI consistently maintains best fidelity to target coverage. Here, label usage is nearly identical across methods for a given human fraction, ruling out differences in annotation cost as the primary driver of performance gaps.

Despite this, AdaStrat-ASI consistently achieves the smallest confidence interval width across all budgets. Relative to ASI, this corresponds to an approximate 10%˘14% reduction in CI width across most human fractions, with larger gains at smaller budgets. Moreover, AdaStrat-ASI selects labeled points with the largest diameter (from Eq. (2)), i.e., the widest spread across the data manifold.

E.3 Runtime Overhead

AdaStrat-ASI adds a small computational overhead primarily due to the the zero-label scouting phase compared to ASI. For instance on the Housing Dataset, ASI takes 7.336 ± 0.369 ms and AdaStrat-ASI takes 8.565 ± 0.189 ms.

E.4 Guidelines on stratifying data

The effectiveness of stratified inference critically depends on how strata are defined. In this section we shed light on practical ways to identify strata in the inference data. As is standard in stratified sampling, strata with very small sample sizes are merged with neighboring or similar strata to avoid unstable variance estimates and degenerate allocations [39, 13].

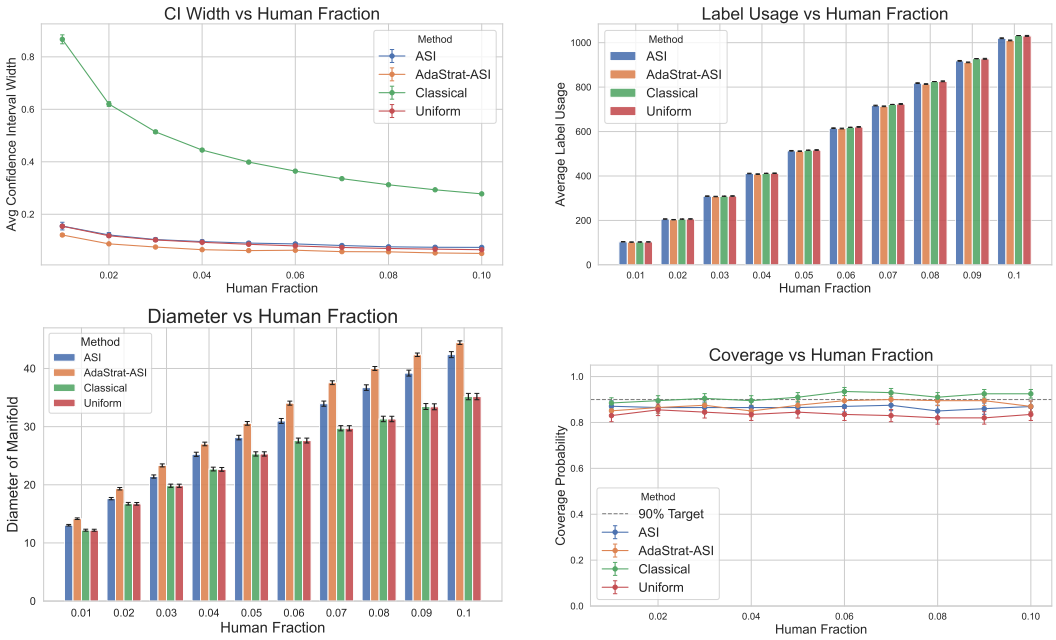


Figure 7: Housing Dataset

E.4.1 Domain Knowledge

In many real-world applications the choice of stratifying variable is not uniquely prescribed. Instead, it often hinges on practitioner intuition and domain expertise. Practitioners may select stratification variables based on prior empirical findings, exploratory analyses, or contextual understanding of the data-generating process. Such judgment-driven stratification leverages known scientific, biological, or social structure to partition the data into subpopulations that are expected to exhibit relatively homogeneous response behavior. For instance, in biomedical and clinical studies, stratifying by sex, age group, tissue type, or disease subtype is often essential, as these factors are known to induce systematic differences in outcomes and variability [11, 46]. Similarly, in social science and policy analysis, stratification by demographic or geographic categories (e.g., race, gender, region) is routinely employed to account for heterogeneity in behavior and responses [27]. In our framework, these domain-driven strata provide a natural and effective starting point.

For stratification in the California Dataset (Section E.2.1), we use the categorical variable `ocean_proximity`. Figure 8 shows the comparison between normalized trivial weight-based allocation and weighted model-based allocation across strata for the California Housing dataset.

First, our proxy distribution is markedly non-uniform, indicating that the black-box model identifies systematic differences across strata and is non-trivially informative. Such structure validates the use of proxy-based scouting in AdaStrat-ASI: the allocation reflects meaningful variation in local difficulty, which is subsequently exploited during the attacking stage to reallocate a fixed labeling budget more effectively. Second, our proxy allocation differs from population-proportional sampling by a total variation distance of 0.093, indicating a considerable redistribution of sampling mass across four strata.

We take this opportunity to underline the importance of measuring the difficulty scores ($\hat{\sigma}_k^f$) in the Scouting Stage. Instead of allowing the scouting phase to determine allocation weights i.e. Figure 1, we use an allocation vector purely based on strata size i.e. Figure 8(Left). We use the former allocation method in Section E.2.1. Comparing results in Figure 9 to the results presented in Section E.2.1, there is not much change in width, label-usage, and diameter. However, with purely weight-based allocation, coverage is severely impacted.

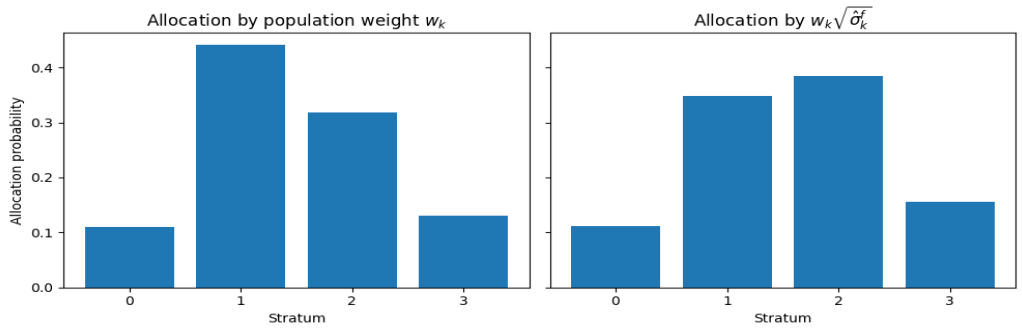


Figure 8: Influence of difficulty scores in Label Allocation

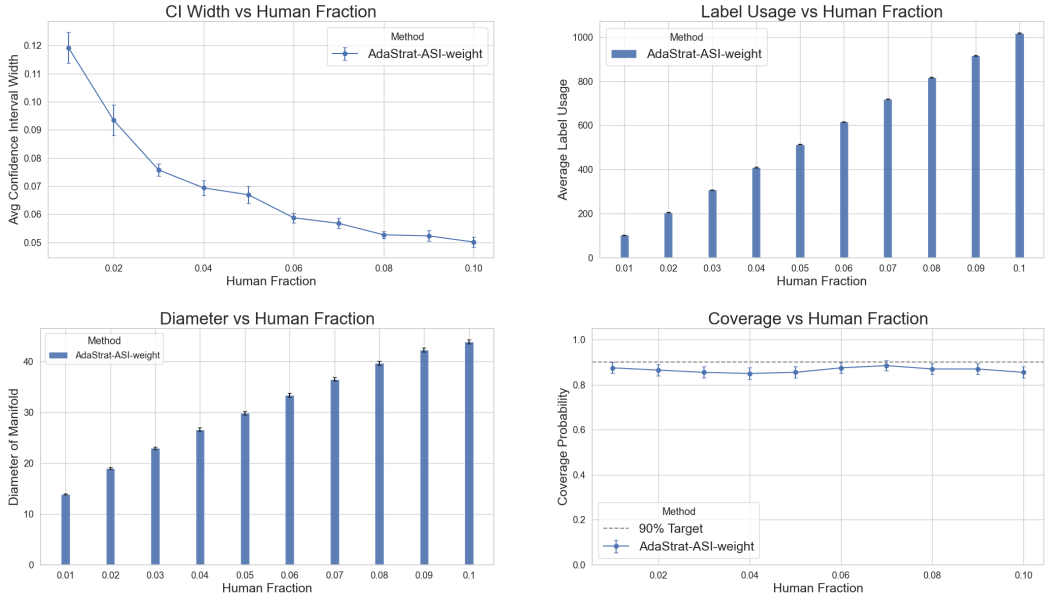


Figure 9: Impact of pure weight-based label allocation in Housing Dataset

E.4.2 Black Box Model based Stratification

In some applications, domain knowledge is unavailable or difficult to exploit, rendering intuition-driven stratification infeasible. This commonly arises when practitioners are provided with anonymized or abstracted datasets, where features lack semantic interpretation due to privacy, proprietary constraints, or institutional policies [18, 32].

To address this scenario, we adopt a model-based stratification strategy that leverages cheap proxy information provided by a black-box predictive model f, which is already central to the ASI framework. Concretely, we stratify the unlabeled pool by jointly considering the model's prediction's $f(x)$ and uncertainty $u(x)$, yielding partitions that reflect both the model's belief and its surety. Our stratifcation approach is loosely inspired by Section 5.2 of Fisch et al. [19].

Intuitively, regions where the model is highly confident are treated as coarse strata because the model's predictions are expected to be reliable proxies for the true response, and additional labels are less informative. Conversely, regions of high uncertainty are partitioned more finely, benefiting from targeted labeling. This construction reinforces our ideology: if a strata is problematic (i.e., high uncertainty), chunk it into smaller fragments to minimize its contribution to global variance (w_k). A similar argument applies to constructing larger strata where the model's predictions are reliable.

In practice, this can be implemented via grid-based stratification over the joint space of $(f(x), u(x))$. Breakpoints along each axis may be chosen using simple heuristics such as quantiles, histogram

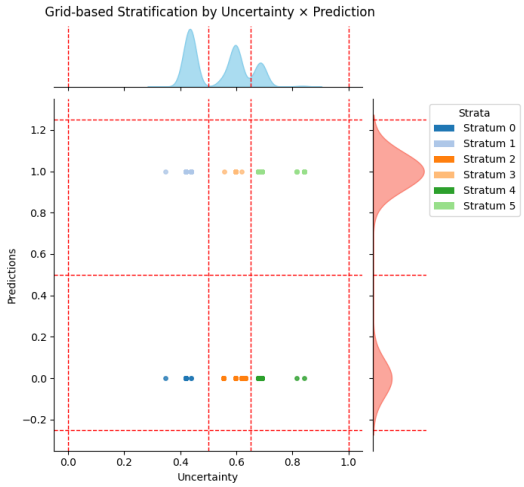


Figure 10: Stratification in Politeness Dataset

modes, or change-point detection, resulting in rectangular strata whose sizes shrink as uncertainty increases. Figure 10 illustrates this procedure on the Politeness dataset (Section 6).

E.5 Impact of Model Sensitivity on Label Allocation

To gauge impact on label allocation on $f(x)$ we extend our synthetic 1D example from Section 6. The setup remains exactly identical i.e. in Stratum 2, the predictor stays perfect: $f(x) = 0$, except in Stratum 1, we deliberately make the predictions noisy. $f(x) \sim 1 + \mathcal{N}(0, \sigma^2)$.

We run a small experiment for a fixed budget of 500 labels and $\sigma^2 = 1$. Because our scouting phase realizes Stratum 2 is perfectly solved ($f(x) = \hat{Y} = 0$), $\hat{\sigma}_2^f$ comes out to be 0, same as in the previous setting. Consequently, 0 budget labels were assigned to Stratum 2. However, in the case of Stratum 1, because of the artificial noise injection, $\hat{\sigma}_1^f \neq 0$ anymore(0.9708 to be precise). This causes Stratum 1 to be assigned all 500 labels.

This setup perfectly demonstrates the impact on model sensitivity on the scouting stage. Thus, the proxy variance $\hat{\sigma}_1^f$ correctly detected the instability, and AdaStrat-ASI dynamically shifted the entire labeling budget to Stratum 1.

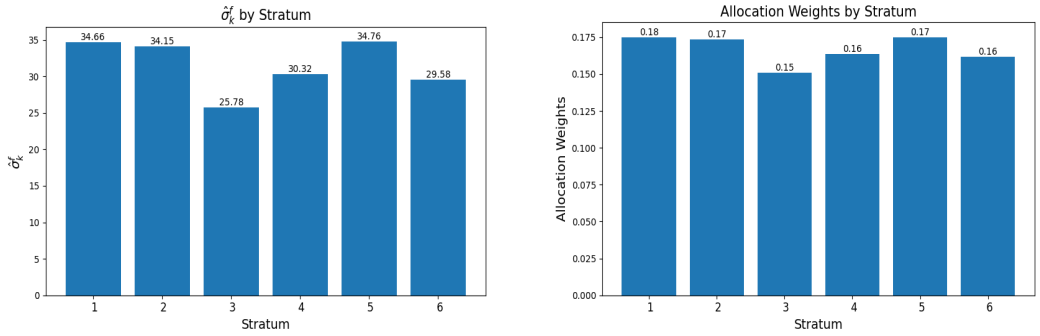
E.6 Random Stratification Ablation

As formalized in our variance reduction guarantee (Theorem 5.6), our method strictly requires the strata to capture latent heterogeneity. While we suggest ways to stratify data E.4, if the chosen strata are uninformative, AdaStrat-ASI's performance can be severely affected. the budget allocation ($\hat{\sigma}_k^f$) approaches uniformity.

To empirically demonstrate this, we run an ablation experiment where we construct $\mathcal{A}$ using a purely random split of the data into K equally sized strata(for simplicity), ignoring domain knowledge and model confidence entirely. We set $K = 6$ for direct comparison, on the Politeness Dataset(Section 6).

From Figure 11(Right), the proxy difficulty scores across these random strata are nearly identical. In this case, $n_{b_k} \propto w_k \approx \frac{1}{K}$. The allocation weights that serve input to the multinomial distribution follow an almost uniform distribution($\approx \frac{1}{6}$). Stratification was clearly not *useful* here.

From Figure 14 compared to the other baselines discussed in Section E.4 and illustrated in Figure 6, label usage and diameter for random allocation are quite similar. Compared to ASI, we still reduce CI-width. Interestingly, per our experimental design γ_k for all strata is roughly the same, which is why the efficiency gap is not exploitable. But, the reduction in CI widths comes from the



(a) Strata-wise difficulty

(b) Allocation PMF

Figure 11: Impact of Random Stratification on Budget Allocation

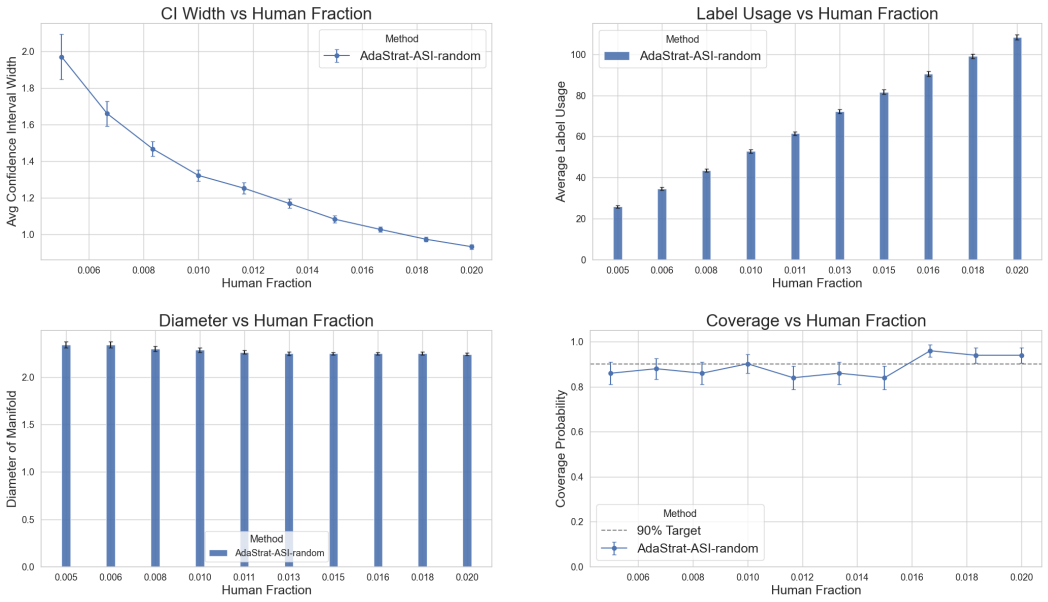


Figure 12: Impact of Random Stratification on Politeness Dataset

post-stratification structure i.e. Δ_{Bias}. Thus, random stratification kills the Efficiency Gap (budget is flat), but preserves the Stability Gap (structural post-stratification).

Compared to Informed Stratification, the width is almost identical, except for a slight difference in the stretch range of $0.01 - 0.013$, where informed Stratification(ours) provides a minor improvement. However, random stratification falls short in terms of coverage.

Poor Coverage. Based on our random stratification setup, each Strata 1-6 contains a random mix of easy points and hard points. Now, when the Attacking Stage begins, inside every stratum, it samples points using active sampling policy $\pi_k(x) \propto u(x)$. Because starta are a random mix of everything, it contains some points with massive uncertainty $u(x)$ and many points with tiny uncertainty. The algorithm will heavily over-sample the high-uncertainty points and largely ignore the low-uncertainty points inside that bucket.

Since Equation (3) follows AIPW, we divide by the sampling probability: $\frac{1}{\pi_k(x)}$. Since the budget allocation is even, the algorithm in expectation is designed to exhaust their respective respective strata budgets. This design makes it highly likely that low-uncertainty points (where $\pi_k(x)$ is very small) are randomly drawn. Naturally, those point gets a massive weight in the final average. Because strata are a random mix of everything, the normalization factor $\mathbb{E}[u(X) \mid A_k]$ is essentially the global

average uncertainty. The probabilities $\pi_k(x)$ stretch from very near 1 to very near 0. This creates extreme propensity weights.

When an estimator has extreme propensity weights, under limited labeling budget, the asymptotic Central Limit Theorem takes a very long time to kick in. The empirical variance estimate used to construct the CIs becomes highly unstable and downward-biased. Hence, the algorithm calculates tight CIs, but because the weights are so erratic, the actual estimates bounce around wildly outside that CI.

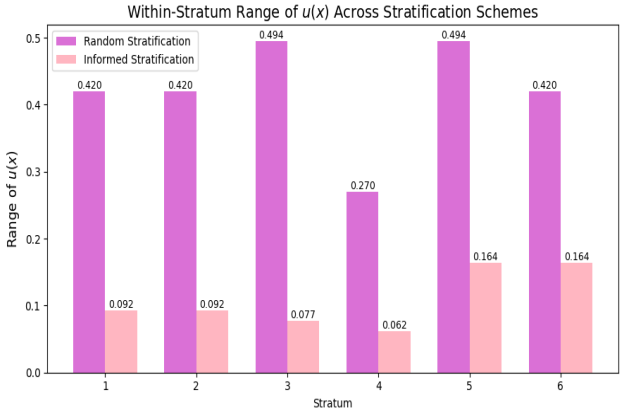


Figure 13: Range of Uncertainty across Strata

Figure 13 compares the within-stratum range of $u(x)$ under random and informed stratification. Since the two stratification procedures induce different partitions of the data, the stratum indices are not directly comparable across methods; the plot is intended only to contrast overall within-stratum homogeneity. The difference in performance goes on to show that informed stratification, in fact, facilitates the emergence on CLT within each stratum.

E.7 Random Allocation Ablation

We retain the informed stratification setup from Section 6. However, instead of allowing the scouting phase to determine allocation weights(Figure 1), we manually inject a purely randomly generated allocation vector:

$$\vec{z} = [0.1734, 0.0821, 0.2567, 0.1349, 0.0913, 0.2616]$$

These weights are generated independently of the data (summing to 1.0 via normalization) and represent a “failed" or “uninformed" scouting phase that assigns budget capriciously. This tests whether the method is robust to misspecified allocation—i.e., when budget is concentrated in arbitrary strata rather than those with proxy variance.

Compared to results illustrated in Figure 6, random budget allocation performs similarly in terms of label usage and diameter against other baselines. Specifically, compared to ASI, random budget allocated outperforms in terms of CI Width. This can be attributed to our pre-stratification mechanism((3)), enforcing $\Delta_{\text{Bias}} > 0$. This is almost identical to what we see in the case of random strata allocation (Section E.6). However, compared to informed stratification, CI-width and coverage have noticeably different results.

We notice wider confidence intervals across all budgets, reflecting the suboptimal allocation violating the Neyman condition of Theorem 5.6. This is followed by severe undercoverage ($< 80\%$) at low budgets ($n_b < 0.012$), driven by budget starvation in high-variance strata preventing within-stratum CLT convergence. Here stabilization occurs at $87\% - 89\%$ coverage once all strata accumulate sufficient samples ($n_b \geq 0.012$) allowing CLT to activate. So, coverage eventually gets close, but finite-sample corrections from the arbitrary allocation still permanently degrades it. In fact, the coverage would asymptotically approach 90% as $n_b \to \infty$, but the convergence is governed by the most starved stratum, requiring budgets beyond our experimental range to fully eliminate the calibration gap.

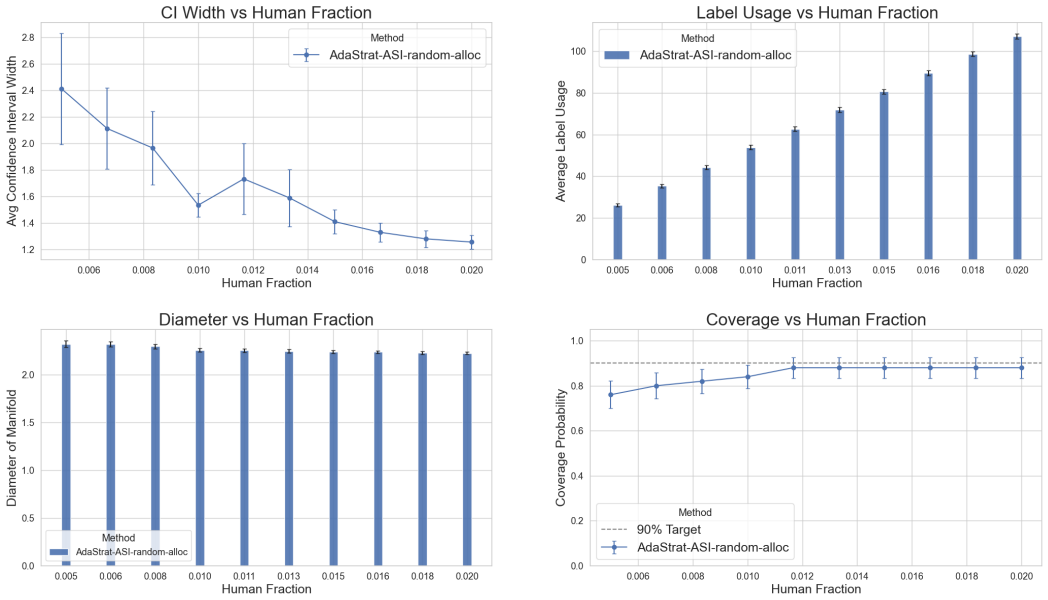


Figure 14: Impact of Random Label Allocation on Politeness Dataset

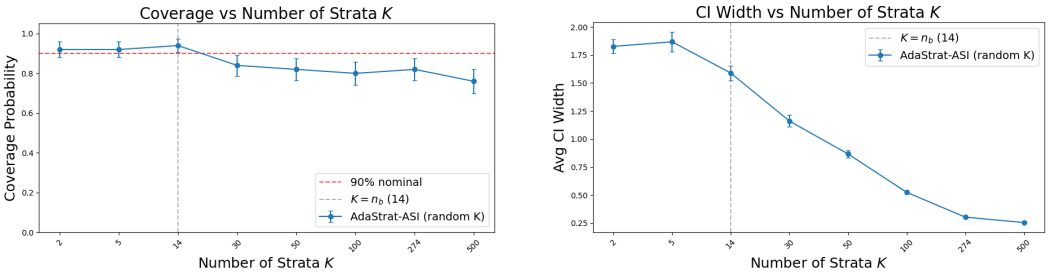


Figure 15: Impact of Rampant Stratification on Politeness Dataset

This confirms that stratification *and* allocation are complementary necessities. Even with strata that perfectly capture model heterogeneity, arbitrary budget allocation destroys the variance reduction guarantee. The scouting phase is not merely a convenience but a critical component for valid, efficient inference. So while practitioners *can* choose to skip the Scouting Stage to save on runtime(Section E.3), the performance becomes severely suboptimal.

E.8 Effect of K on AdaStrat-ASI

In this section we perform an ablation on K i.e. our strata count. We borrow the random stratification setup from Appendix E.6 , where we create K strata via random partitioning of the pool, varying $K \in \{2, 5, 14, 30, 50, 100, 274, 500\}$ on the Politeness dataset at a fixed labelling budget of 0.05% (14 labelled points in total). For each value of K we report empirical coverage and CI width. As shown in Figure 15, CI width and Coverage decrease as K increases. Specifically, coverage remains near the nominal level $1 - \alpha = 0.90$ for small to moderate $K \approx 14$, however, where the *expected* number of labelled points per stratum, $\frac{n_b}{K}$ drops below one, coverage degrades sharply and CI width becomes unreliable. This behaviour is expected: the within-stratum sandwich variance estimator relies on a consistent estimate of V_k, which requires sufficiently many labelled observations per stratum for the central limit theorem to hold locally. Hence

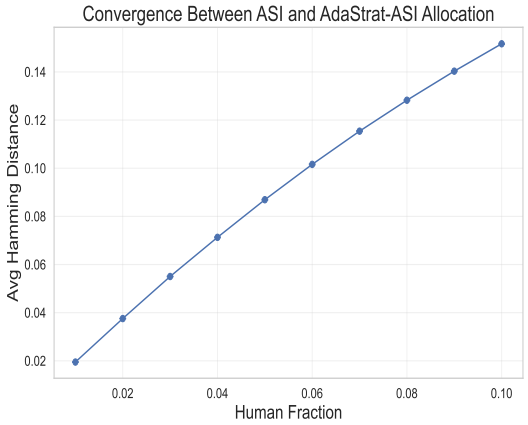


Figure 16: Difference in Label Distributions of ASI and AdaStrat-ASI

E.9 Does stratification(scouting stage) make any difference?

AdaStrat-ASI is not designed to reduce labeling cost or to conserve budget. By construction, it operates under a fixed labeling budget n_b and, in expectation, expends the same number of labels as standard ASI. This follows directly from the multinomial budget allocation in the attacking stage and is consistently observed empirically(Section 6). The objective of AdaStrat-ASI is instead to alter *how* a fixed labeling budget is allocated across the data space, rather than *how much* labeling is performed.

To gauge this impact, for the California Housing Dataset(Section E.2.1) for a fixed total labeling budget n_b, we define two random binary vectors over the inference pool of size n: $\boldsymbol{\xi}^{\text{ASI}} \in \{0,1\}^n$, and $\boldsymbol{\xi}^{\text{AdaStrat}} \in \{0,1\}^n$ where, $\sum_i \xi_i^{\text{ASI}} = \sum_i \xi_i^{\text{AdaStrat}} = n_b$, but the locations of the ones differ due to stratified reallocation.

We then compute

$$\bar{H}(\boldsymbol{\xi}^{\text{ASI}}, \boldsymbol{\xi}^{\text{AdaStrat}}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\xi_i^{\text{ASI}} \neq \xi_i^{\text{AdaStrat}}\}$$

and report its average over trials as shown in Figure 16.

- Low Budget: Both methods select very few points. Most entries in both vectors are zero. Hence average Hamming distance is close to zero.
- High Budget: The Hamming distance increases because stratification actively reshapes the allocation pattern under the same total budget.

Though not shown, at large budgets $n_b \to n$, both vectors will converge to the all-ones vector. Any allocation rule becomes identical $\bar{H} \to 0$. So the curve must eventually come back down.

In theory, the true shape is unimodal (low $\to$ high $\to$ zero). This analysis highlights that the primary impact of stratification lies in redistributing labels across strata when labels are scarce, which is precisely the regime where allocation decisions matter most due to practical constraints.

Thus, the variance reduction (Theorem 5.6) is not just theoretical; it stems from a concrete shift in which specific data points are labeled. The gap represents the *correction* AdaStrat-ASI applies to preventing over-spending on popular but easy strata.

We formalize this by modeling the unnormalized Hamming distance $H(\beta)$ as a quadratic:

$$H(\beta) = A\beta^2 + B\beta + C$$

where $\beta = \frac{n_b}{n}$.

We know, $H(0) = 0$. This implies, $C = 0$. We also know, $H(1) = 0$. This implies $A = -B$.

Recall, $H(\boldsymbol{\xi}^{\text{ASI}}, \boldsymbol{\xi}^{\text{AdaStrat}}) = H(\boldsymbol{\xi}^{\text{ASI}}, \boldsymbol{\xi}^{\text{AdaStrat}}) = |\boldsymbol{\xi}^{\text{ASI}} \cup \boldsymbol{\xi}^{\text{AdaStrat}}| - |\boldsymbol{\xi}^{\text{ASI}} \cap \boldsymbol{\xi}^{\text{AdaStrat}}| = 2n_b - 2|\boldsymbol{\xi}^{\text{ASI}} \cap \boldsymbol{\xi}^{\text{AdaStrat}}|$

For ease of notation, let $I = |\boldsymbol{\xi}^{\text{ASI}} \cap \boldsymbol{\xi}^{\text{AdaStrat}}|$.

Let ρ be the correlation between the selection policies.

Case 1: Perfect Correlation ($\rho = 1$). If the policies are identical, they pick the exact same top-n_b items. Thus, $I = n_b$.

Case 2: Random/Uncorrelated ($\rho = 0$). If one policy is random relative to the other, the intersection is just the probability of picking the same item twice: $P(\text{pick}) \times P(\text{pick}) = \beta \times \beta$. Thus, $I = n \cdot \beta^2 = n_b \cdot \beta$.

We model the actual intersection as a weighted average of Perfect Overlap ($\rho = 1$) and Random Overlap, ($\rho = 0$) weighted by the correlation coefficient ρ:

$$I(\beta) = \rho \cdot (n_b) + (1 - \rho) \cdot (n_b \beta)$$

On replacing and simplifying we get, $H(\beta) = 2n(1 - \rho)(\beta - \beta^2)$

Naturally the normalized Hamming Distance becomes,

$$\bar{H}(\beta) = (-\beta^2) \cdot 2(1 - \rho) + (\beta) \cdot 2(1 - \rho)$$

Magnitude: $\bar{H}(\beta)$'s value is maximized for β at $\beta = 0.5$ i.e.

$$\bar{H}(\beta) = \frac{1 - \rho}{2}$$

Similarly, $\bar{H}(\beta)$'s value is maximized for ρ at $\rho = 0$ i.e.

$$\bar{H}(\beta) = 2\beta(1 - \beta)$$

These findings jointly indicate that $0 \leq \bar{H}(\beta) \leq 0.5 \; \forall \beta, \rho$.

Interpretation: The amplitude A is directly influenced by $(1 - \rho)$, which represents the 'decorrelation' between ASI and AdaStrat-ASI. Consequently, when ρ approaches 0, indicating a high degree of decorrelation, the Hamming Distance over low label budgets exhibits a more pronounced upward curvature, resulting in a higher Hamming Distance. In practical terms, stratifying and then sampling reduces the correlation (ρ) because we concentrate on sampling from k partitions within the n-dimensional data vector, rather than the entire data vector directly.

Conversely, when the global (ASI) and local (AdaStrat-ASI) sampling rules converge, decorrelation approaches 0. This flattening of the curve ($\bar{H}(\beta) \to 0$) signifies that there is no actual change in the labeling distributions, which aligns with intuitive reasoning.