Stratified Prediction-Powered Inference for Hybrid Language Model Evaluation

Adam Fisch[†,*] **Joshua Maynez**[†,*] **R. Alex Hofer**[†]

Bhuwan Dhingra[†] **Amir Globerson**[‡] **William W. Cohen**[†]

[†]Google DeepMind [‡]Google Research

{fisch,joshuahm,rofer,bdhingra,amirg,wcohen}@google.com

Abstract

Prediction-powered inference (PPI) is a method that improves statistical estimates based on limited human-labeled data. PPI achieves this by combining small amounts of human-labeled data with larger amounts of data labeled by a reasonably accurate—but potentially biased—automatic system, in a way that results in tighter confidence intervals for certain parameters of interest (e.g., the mean performance of a language model). In this paper, we propose a method called Stratified Prediction-Powered Inference (StratPPI), in which we show that the basic PPI estimates can be considerably improved by employing simple data stratification strategies. Without making any assumptions on the underlying automatic labeling system or data distribution, we derive an algorithm for computing provably valid confidence intervals for population parameters (such as averages) that is based on stratified sampling. In particular, we show both theoretically and empirically that, with appropriate choices of stratification and sample allocation, our approach can provide substantially tighter confidence intervals than unstratified approaches. Specifically, StratPPI is expected to improve in cases where the performance of the autorater varies across different conditional distributions of the target data.

1 Introduction

Evaluating machine learning models requires *evaluation* data. In particular, to iteratively improve on a method during development, or to reliably report an improvement, one needs to confidently and quantitatively assess the performance of the method. This is especially challenging for large language models (LLMs), where gathering high-quality annotations for generations can be difficult and time-consuming—and can ultimately become quite costly if gathering more than a few hundred examples.

One often-proposed approach to avoiding the evaluation bottleneck is to use a secondary LLM-based system to judge the output of the primary one. For instance, if the primary task is developing an LLM-based question-answering (QA) system, one can use a second LLM-based system that rates question/answer pairs as acceptable or not [6; 8; 17]. However, automated raters (i.e., *autoraters*) may be biased relative to the human raters they are intended to model, as others have noted [1; 9; 18; 23]. This can become substantially worse when models are tailored to hill climb on the autorater metrics, and eventually cease to become a good metric at all—a phenomenon commonly referred to as Goodhart's law, or reward hacking in the context of reinforcement learning from human feedback [13; 24].

We thus have two signals for assessing model performance. The first is human labels, which are typically accurate, but expensive to collect. As a result, usually only a small sample size is available, and

[*] Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

an estimate based on these samples alone will have high variance. The second is autorater predictions, which are easy to collect for large sample sizes, but may also be systematically biased. The above may suggest that one must make a choice between either (i) a high-variance, but unbiased, estimate from a small human sample, or (ii) a lower-variance, but biased, autorater-based estimate. However, it turns out there are also statistically valid ways of *combining* the auto-rater and human data for hybrid evaluations. Following [1; 3; 7] we call such methods *prediction-powered inference* (PPI) methods. At a high level, PPI-based methods operate by using a small sample of examples labeled by both humans and autoraters to estimate the bias of the autorater. This bias is then used as a rectifier for the autorater estimate. The resulting estimate can then be shown to provide improved (i.e., tighter) confidence intervals for properties of interest for the system being evaluated, such as its true mean accuracy.

A weakness of standard PPI, however, is that it does not take heterogeneity into account. For example, in our QA setting, an autorater may have one accuracy when predicting if a model answer is correct, and a different accuracy when predicting if it is incorrect. This is especially true in cases where a correct answer is easy to verify (e.g., it is also present in Wikipedia), but harder to refute (e.g., no relevant external search results can be retrieved to either support or contradict it). In these settings, it may make sense to apply a different PPI strategy within each subdomain, depending on the local quality of the autorater. Moreover, we claim that such heterogenous settings are to be quite expected in practical applications. Inspired by the rich prior literature on stratified sampling and survey design [12; 19; 21], we therefore propose a *stratified* approach to PPI, and show that it can be very advantageous when performance varies across subdomains, for either the autorater, or the model being evaluated.

On a technical level, it is not immediately clear how to apply stratification to PPI, as it involves two types of samples: one sample that is labeled by both humans and an autorater, and another (typically much larger) sample that is only labeled by the autorater. Extending the analysis of [3], in this work we show how confidence intervals based on the asymptotic normality of weighted M-estimators [32] can in fact be derived for the stratified PPI setting. The next challenge we address is how to determine the sample sizes used for stratification (i.e., the sample size of each stratum). Similar to recent work for active statistical inference [37], we further derive optimal rates that depend on certain moments of the underlying distribution that are generally unknown—and provide an approach for effectively approximating these. Finally, we provide extensive empirical evidence showing that our stratified approach (StratPPI) leads to considerably improved confidence intervals over both classical inference methods that use only human labels, and the baseline (unstratified) PPI approach.

2 Related Work

Prediction Powered Inference (PPI) was introduced in [1] as a method for obtaining tighter confidence intervals for parameters learned in supervised machine learning (e.g., coefficients in logistic regression) by also leveraging other model-based predictions on additional, unlabeled data. Related ideas were explored by [30], but with a focus on bootstrapping as a way for obtaining confidence intervals. PPI was then extended in several directions. For example [36] showed how the labeled data can be used for both estimating the parameters and the autorater model. PPI++ [3] showed how to obtain confidence intervals that are easy to compute efficiently, and introduced a parameter for weighting the predictions of the autorater such that the overall statistical efficiency can be improved. As noted in previous works on PPI, these approaches are closely related to other statistical methods for introducing control variates based on autoevaluators [9], as well as augmented inverse propensity weighting [27] (see discussion in [1; 3]). Like this paper, prior work has also focused on using PPI/PPI++ for evaluating machine learning systems with autoraters, including for ranking chatbots in the Chatbot Arena [7] and evaluating retrieval-augmented generation systems [28].

Most relevant to our setting, the recent work of [37] focuses on *active* sampling of examples to label during PPI, where the total number of queried labels is random, but less than the sampling budget n in expectation. Specifically, it proposes to label examples for which the autorater is less confident in its predictions, and corrects for the sampling bias with a variant of inverse propensity weighting. In contrast, our somewhat simplified approach takes inspiration from stratified sampling where a coarse stratification is defined in advance, and the total number of labeled samples is constant. Like [37], we derive an analogous optimal allocation strategy for a given budget, though we only apply it at the stratum level, and not for individual examples. Furthermore, while the variable allocation helps reduce variance in heterogenous settings, our stratified treatment also allows for a stratum-specific extension of [3]'s estimated tuning parameters that further improves performance in complementary ways.

In terms of stratification specifically, concurrent work [14] also proposed Bayesian variants of several PPI methods, including stratified approaches. In contrast, the methods here do not require introducing priors, do not require running expensive Bayesian inference, and give more conventional, frequentist, guarantees of performance. The credible intervals produced by Bayesian methods are related to, but different from, the confidence intervals produced here, and by other prior PPI approaches [1; 3; 7].

Finally, there is a conceptual similarity between the PPI rectifiers and post-hoc calibration of a classifier. There is a rich literature on calibrating classifiers (e.g., [22; 25]), but there is a clear difference in goals between calibration and PPI: the former is aimed at modifying a learned autorater to make better probabilistic predictions, and the latter is aimed at obtaining better confidence intervals by making use of an existing autorater. That said, clearly one approach to improving PPI is to use a better-calibrated classifier, perhaps by taking some of the labeled data and using it for re-calibration. This is similar in motivation to the cross-PPI approach of [36]. We leave such approaches as future work, but we do experimentally explore PPI approaches on both well-calibrated autoraters and poorly calibrated ones.

3 Preliminaries

We begin by briefly reviewing PPI [1], and specifically the efficient PPI++ variant of [3]. Here, and in the rest of the paper, upper-case letters (X) denote random variables; lower-case letters (x) denote constants, and script letters ($\mathcal{X}$) denote sets, unless specified. All proofs are deferred to Appendix A.

Following [1; 3; 7], we assume an empirical sample $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of n i.i.d. examples drawn from some unknown, but fixed, distribution $\mathbb{P}$, where $X_i \in \mathcal{X}$ is an input, and $Y_i \in \mathcal{Y}$ is the target output. We also assume a larger sample $\tilde{S}_N = \{(\tilde{X}_1, f(\tilde{X}_1)), \ldots, (\tilde{X}_N, f(\tilde{X}_N))\}$ where $N \gg n$, for which the target outputs are not available, but we have access to an *autorater function* $f \colon \mathcal{X} \to \mathcal{Y}'$ which provides an approximation of Y given the observed input (we use $\mathcal{Y}'$ to denote the output space of the autorater as it is possible that $\mathcal{Y} \neq \mathcal{Y}'$, e.g., if $f(X) = \mathbb{E}[Y \mid X]$).

As an example, assume X is a *(question, model answer)* tuple where the answer is produced by an LLM-based QA system we wish to evaluate, Y is a $0/1$ "gold" human rating of correctness of the answer, i.e., $\mathbf{1}\{\textit{human rates model answer as correct}\}$, and $f(X)$, the autorater, imperfectly predicts this human rating Y. One value of interest is then the average accuracy of the QA system, $\mathbb{E}[Y]$. More generally, given any convex loss function $\ell_\theta(x, y)$ satisfying certain regularity conditions (see Definition A.1), we are interested in estimating the d-dimensional parameter θ^*,

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}[\ell_\theta(X, Y)]. \tag{1}$$

In the statistics literature, this is broadly referred to as M-estimation (e.g., see [29]). Here, we focus on *mean* estimation, for which the loss $\ell_\theta(x, y) = \frac{1}{2}\|y - \theta\|^2$ has the optimum $\theta^* = \mathbb{E}[Y]$. However, our method generalizes to other typical losses such as the squared loss for linear regression, where the parameters θ^* are the regression coefficients, or, similarly, the log loss for other generalized linear models. Of course, the true θ^* is generally not known, because we cannot exactly calculate the expectation in (1). The goal of PPI, which we share in this work, is to use both the labeled and unlabeled data S_n and $\tilde{S}_N$ to obtain an asymptotically valid confidence interval $\mathcal{C}_{\alpha,j}$ for θ^*_j that for any $j \in [d]$ satisfies[1]

$$\lim_{n,N \to \infty} \mathbb{P}(\theta^*_j \in \mathcal{C}_{\alpha,j}) \geq 1 - \alpha. \tag{2}$$

The simplest confidence interval can be obtained using standard techniques by using only the labeled sample S_n, and ignoring the autorater data $\tilde{S}_n$. However, PPI employs a clever trick which allows for also using the unlabeled sample $\tilde{S}_N$ to get tighter confidence intervals, as described next.

3.1 A rectified prediction-powered loss

The key idea of PPI-based methods is to use autorater predictions to derive a low variance, but unbiased, estimate of the objective in (1). Consider the following "rectified" prediction-powered loss:

$$L^{\mathrm{PP}}(\theta) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \ell_\theta(\tilde{X}_i, f(\tilde{X}_i))}_{\text{autorater data loss}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, Y_i) - \ell_\theta(X_i, f(X_i))}_{\text{loss rectifier}}. \tag{3}$$

[1] In d-dimensions, we obtain a confidence interval for each coordinate. See [3] for other possible choices.

The rectifier term on the right hand side of (3) removes the bias of the autorater data so that $L^{\mathrm{PP}}(\theta)$ satisfies $\mathbb{E}[L^{\mathrm{PP}}(\theta)] = \mathbb{E}\left[\ell_\theta(X, Y)\right]$. However, when $f(X)$ is correlated with Y, the loss $L^{\mathrm{PP}}(\theta)$ will have lower variance. Unfortunately, $f(X)$ may not be a good predictor of Y in all cases. In fact, it is even possible for L^{PP} to be higher variance than the classical estimate (e.g., if $f(X)$ is anti-correlated with Y). Thus, to adapt to the quality of $f(X)$, PPI++ also introduces a tuning parameter $\lambda \in \mathbb{R}$:[2]

$$L_\lambda^{\mathrm{PP}}(\theta) = \frac{\lambda}{N}\sum_{i=1}^{N} \ell_\theta(\tilde{X}_i, f(\tilde{X}_i)) + \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, Y_i) - \lambda \ell_\theta(X_i, f(X_i)). \tag{4}$$

Clearly, L_λ^{PP} still remains unbiased for any value of λ. For $\lambda = 0$, the framework reduces to classical inference. In most cases, however, a proper choice of $\lambda \neq 0$ will result in lower variance. PPI++ also suggests how to automatically choose a data-dependent $\hat{\lambda}$ that converges to an optimal value.

3.2 A prediction-powered confidence interval

We next describe the PPI++ method for deriving confidence intervals based on the rectified loss. Let

$$\hat{\theta}_{\hat{\lambda}}^{\mathrm{PP}} = \operatorname*{argmin}_{\theta \in \Theta} L_{\hat{\lambda}}^{\mathrm{PP}}(\theta). \tag{5}$$

be the prediction-powered estimate. Standard analysis for M-estimators can then be extended to show that the scaled difference of the estimate $\hat{\theta}_{\hat{\lambda}}^{\mathrm{PP}}$ and the true θ^* is asymptotically normally distributed. PPI++ then leverages this result to compute a confidence interval for θ^* that is asymptotically valid.

Theorem 1 (PPI++, [3]). *Assume that $\hat{\lambda} \xrightarrow{p} \lambda$ and $\frac{n}{N} \to r \geq 0$. Let $H_{\theta^*} := \mathbb{E}[\nabla^2 \ell_{\theta^*}]$, and*

$$V_{f,\theta^*}^{\lambda} := \lambda^2 \mathrm{Cov}(\nabla \ell_{\theta^*}(\tilde{X}, f(\tilde{X}))), \quad V_{\Delta,\theta^*}^{\lambda} := \mathrm{Cov}(\nabla \ell_{\theta^*}(X, Y) - \lambda \nabla \ell_{\theta^*}(X, f(X))), \tag{6}$$

where $\lambda \in \mathbb{R}$ is a hyper-parameter. Then under the regularity conditions of Definition A.1, we have that $\sqrt{n}(\hat{\theta}_{\hat{\lambda}}^{\mathrm{PP}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^\lambda)$, where $\Sigma^\lambda := H_{\theta^*}^{-1}(r \cdot V_{f,\theta^*}^{\lambda} + V_{\Delta,\theta^*}^{\lambda})H_{\theta^*}^{-1}$.

Corollary 1 (PPI++ CI, [3]). *Let $\widehat{\Sigma}^{\hat{\lambda}}$ be the plug-in estimate for Σ^λ using S_n and $\tilde{S}_N$. Define*

$$\mathcal{C}_{\alpha,j}^{\mathrm{PP}} := \left(\hat{\theta}_{\hat{\lambda},j}^{\mathrm{PP}} \pm z_{1-\frac{\alpha}{2}} \sqrt{n^{-1}\widehat{\Sigma}_{jj}^{\hat{\lambda}}}\right), \tag{7}$$

where z_β denotes the β-quantile of the standard normal distribution. Then for any $j \in [d]$ it holds that $\lim_{n,N\to\infty} \mathbb{P}\left(\theta_j^* \in \mathcal{C}_{\alpha,j}^{\mathrm{PP}}\right) \geq 1 - \alpha$.

As mentioned above, [3] further showed that for an appropriate choice of λ (that can be analytically derived and estimated via a $\hat{\lambda}$), the trace of the covariance matrix, $\mathrm{Tr}(\Sigma^\lambda)$, is at most that of the covariance matrix derived without using autorater data, implying that the PPI++-based confidence set $\mathcal{C}_\alpha^{\mathrm{PP}}$ can always be at least as tight as that of the classical M-estimator (and often much tighter in practice).

4 Stratified prediction-powered inference

We now present our approach for improving PPI++ estimates via *stratification*. In particular, we show how optimizing a rectified loss computed via stratified sampling can lead to a consistent, but lower variance, estimate of the optimal parameter θ^*, and correspondingly tighter confidence intervals. Consider the QA example from the previous section. For most autoraters, it is reasonable to assume that the strength of their performance can vary, depending on the type of input being presented. For instance, an autorater might be accurate at predicting whether an answer to an unambiguous question (e.g., *"What is the capital of France?"*) is correct, but relatively poor at inferring if an answer to an open-ended question (e.g., *"What is the best way to cook a hamburger?"*) is acceptable or not. Splitting the problem space into different domains allows us to derive a more specialized form of the prediction powered loss that can better adapt to this autorater heterogeneity via stratified sampling.

Formally, assume that the input space $\mathcal{X}$ 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. For each stratum,

[2] As noted in [3], for some losses ℓ_θ we must have $\lambda \in [0, 1]$ to guarantee convexity of L_λ^{PP}. The main focus of this paper, however, is on mean estimation with $\ell_\theta(x, y) = \frac{1}{2}\|y - \theta\|^2$, which is convex for any $\lambda \in \mathbb{R}$.

we further assume that we can estimate $w_k = \mathbb{P}(X \in \mathcal{A}_k)$ arbitrarily well using large amounts of unlabeled data or prior knowledge; we treat them as known constants here for simplicity. To collect the labeled and unlabeled datasets S_n and $\tilde{S}_N$, we then follow a standard stratified sampling procedure in which we draw two i.i.d. sets of samples of fixed size n_k and N_k, respectively, from each stratum k, where $\sum_{k=1}^K n_k = n$ and $\sum_{k=1}^K N_k = N$. The relative sizes n_k/n and N_k/N are free parameters; they can simply be the natural rates, $n_k/n \approx N_k/N \approx w_k$, or systematically decided (see §4.3). Note that this is a fundamentally different sampling model from standard `PPI`: here examples are i.i.d. within each strata, and independent (but not necessarily identically distributed) across each strata.

Next, we define the stratified loss via the weighted sum, $L_\lambda^{\mathrm{SPP}}(\theta) = \sum_{k=1}^K w_k L_{k,\lambda_k}^{\mathrm{PP}}(\theta)$, where w_k is the stratum weight, $\lambda = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k$ are now *stratum-specific* tuning parameters, and $L_k^{\mathrm{PP}}(\theta)$ is the conditional prediction-powered loss computed within each stratum, i.e.,

$$L_{k,\lambda_k}^{\mathrm{PP}}(\theta) = \frac{\lambda_k}{N_k} \sum_{i=1}^{N_k} \ell_\theta(\tilde{X}_{ik}, f(\tilde{X}_{ik})) + \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_\theta(X_{ik}, Y_{ik}) - \lambda_k \ell_\theta(X_{ik}, f(X_{ik})). \tag{8}$$

As before, each $L_{k,\lambda_k}^{\mathrm{PP}}$ is an unbiased estimate of the conditional loss. Like `PPI++`, we also allow for data-dependent parameters $\hat{\lambda}_k$ that we show how to automatically tune for best performance in §4.2.

4.1 A stratified prediction-powered confidence interval

We now present our main result, which is a confidence interval for θ^* based on the stratified loss. More precisely, the result states that, as in `PPI++`, the minimizer of the stratified loss,

$$\hat{\theta}_{\hat{\lambda}}^{\mathrm{SPP}} = \operatorname*{argmin}_\theta L_{\hat{\lambda}}^{\mathrm{SPP}}(\theta), \tag{9}$$

has an asymptotically normal distribution with mean θ^*. See Algorithm 1 for pseudocode.

Theorem 2. *Assume that $\hat{\lambda}_k \xrightarrow{p} \lambda_k$, $\frac{n}{N} \to r$ for any $r \geq 0$, $\frac{n_k}{n} \to \rho_k$ for any $\rho_k > 0$, and $\frac{N_k}{N} \to \tilde{\rho}_k$ for any $\tilde{\rho}_k > 0$. Let $H_{k,\theta^*} := \mathbb{E}[\nabla^2 \ell_{\theta^*}(X, Y) \mid X \in \mathcal{A}_k]$, and*

$$V_{k,f,\theta^*}^{\lambda_k} := \lambda_k^2 \mathrm{Cov}(\nabla \ell_{\theta^*}(\tilde{X}, f(\tilde{X})) \mid \tilde{X} \in \mathcal{A}_k), \tag{10}$$

$$V_{k,\Delta,\theta^*}^{\lambda_k} := \mathrm{Cov}(\nabla \ell_{\theta^*}(X, Y) - \lambda_k \nabla \ell_{\theta^*}(X, f(X)) \mid X \in \mathcal{A}_k), \tag{11}$$

where $\lambda_k \in \mathbb{R}$ is a hyper-parameter. Then, under the regularity conditions of Definition A.1,

$$\sqrt{n}(\hat{\theta}_{\hat{\lambda}}^{\mathrm{SPP}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_w^\lambda), \tag{12}$$

where $\Sigma_w^\lambda := A_w^{-1} B_w^\lambda A_w^{-1}$ and:

$$A_w := \sum_{k=1}^K w_k H_{k,\theta^*} \qquad B_w^\lambda := \sum_{k=1}^K w_k^2 \left(\frac{r}{\tilde{\rho}_k} \cdot V_{k,f,\theta^*}^{\lambda_k} + \frac{1}{\rho_k} \cdot V_{k,\Delta,\theta^*}^{\lambda_k} \right). \tag{13}$$

To obtain the result, we combine unstratified `PPI++` with the asymptotic properties of weighted M-estimators [32]. The resulting confidence interval for θ^* is then derived analogously to Corollary 1.

Corollary 2. *Let $\widehat{\Sigma}_w^{\hat{\lambda}}$ be the plug-in estimate for Σ_w^λ using S_n and $\tilde{S}_N$. Define*

$$\mathcal{C}_{\alpha,j}^{\mathrm{SPP}} := \left(\hat{\theta}_{\hat{\lambda},j}^{\mathrm{SPP}} \pm z_{1-\frac{\alpha}{2}} \sqrt{n^{-1} \widehat{\Sigma}_{w,jj}^{\hat{\lambda}}} \right), \tag{14}$$

where z_β denotes the β-quantile of the standard normal distribution. Then for any $j \in [d]$,

$$\lim_{n,N \to \infty} \mathbb{P}\left(\theta_j^* \in \mathcal{C}_{\alpha,j}^{\mathrm{SPP}}\right) \geq 1 - \alpha. \tag{15}$$

The form of the stratified prediction-powered confidence interval is similar to that of `PPI++`, except that it is based off of the weighted stratum-conditional covariance matrices. The effect of this change, however, is significant. In fact, in the case of mean estimation, we show that the asymptotic variance of `StratPPI` is at most that of `PPI++` (even without any additional tuning of λ_k and ρ_k).

Algorithm 1 Stratified prediction-powered inference for general M-estimators (StratPPI)

Definitions: f is the autorater. Inputs $\{(\tilde{X}_{ik}, f(\tilde{X}_{ik}))\}_{i=1}^{N_k}$ include the autorater predictions on sampled unlabeled data for each partition $\mathcal{A}_k$, $k \in [K]$. Inputs $\{(X_{ik}, Y_{ik}, f(X_{ik}))\}_{i=1}^{n_k}$ include the autorater predictions on sampled labeled data for each partition $\mathcal{A}_k$, $k \in [K]$. w_k is the partition weight for $\mathcal{A}_k$, $k \in [K]$. α is the confidence interval coverage error tolerance.

```
# Pick weighting parameters, see §4.2.
Select $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$
# Solve for the minimizer of the stratified, prediction-powered empirical loss.
$\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}} = \mathrm{argmin}_\theta\, L^{\mathrm{SPP}}_{\hat{\lambda}}(\theta)$
# Estimate the Hessian of the true expected loss at $\theta^*$.
$\hat{A}_w = \sum_{k=1}^K \frac{w_k}{n_k} \sum_{i=1}^{n_k} \nabla^2 \ell_{\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}}(X_{ik}, Y_{ik})$
# Build the estimated covariance matrix from each stratified component of the loss.
$\widehat{\Sigma}^{\hat{\lambda}}_w = \mathbf{0}_{d \times d}$
for $k = 1, 2, \ldots, K$ do
    $\widehat{V}_f = \hat{\lambda}_k^2 \widehat{\mathrm{Cov}}_{N_k}\left(\nabla \ell_{\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}}(\tilde{X}_{ik}, f(\tilde{X}_{ik}))\right)$
    $\widehat{V}_\Delta = \widehat{\mathrm{Cov}}_{n_k}\left(\nabla \ell_{\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}}(X_{ik}, Y_{ik}) + \hat{\lambda}_k \nabla \ell_{\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}}(X_{ik}, f(X_{ik}))\right)$
    $\widehat{\Sigma}^{\hat{\lambda}}_w = \widehat{\Sigma}^{\hat{\lambda}}_w + w_k^2 \hat{A}_w^{-1} \left(\frac{\widehat{V}_f}{N_k} + \frac{\widehat{V}_\Delta}{n_k}\right) \hat{A}_w^{-1}$
# Return coordinate-wise confidence intervals for $\theta^*$.
$\mathcal{C}^{\mathrm{SPP}}_\alpha = \left\{\hat{\theta}^{\mathrm{SPP}}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\Sigma}^{\hat{\lambda}}_{w,jj}} \,:\, j \in [d]\right\}$
```

Proposition 1. *Let $\lambda_k \in \mathbb{R}$ be any constant for all strata, and fix ρ_k and $\tilde{\rho}_k$ to their natural rates w_k. Then for $\ell_\theta(x, y) = \frac{1}{2}\|y - \theta\|^2$ and any stratification $(\mathcal{A}_1, \ldots, \mathcal{A}_K)$, we have $\mathrm{Tr}(\Sigma^\lambda_w) \leq \mathrm{Tr}(\Sigma^\lambda)$, where Σ^λ is the asymptotic covariance matrix of PPI++. Furthermore, we have that equality holds if and only if both $\mathbb{E}[Y \mid X \in \mathcal{A}_k]$ and $\mathbb{E}[f(X) \mid X \in \mathcal{A}_k]$ are the same for all strata.*

More generally, although our results will hold for arbitrary stratifications, it is best if they are heterogeneous, and chosen such that the individual stratum-conditional variances are minimized. For example, if an autorater systematically *over-estimates* the model's performance on one subdomain, but systematically *under-estimates* the model's performance on another, then splitting these subdomains into different strata can result in a lower variance L^{SPP}_λ. Similarly, if an autorater has much higher noise on some subdomains than others, it can be beneficial to stratify on those subdomains—and then either lower λ_k, allocate a higher proportion of samples n_k/n and N_k/N, or both for the noisier stratas. In §5 we empirically demonstrate that the stratified estimator can indeed lead to considerably tighter confidence intervals in practice, especially with additional tuning of λ_k and ρ_k, as discussed next.

4.2 Optimal weighting of the autorater predictions

In [3] it was shown that the optimal value of λ for PPI++ (i.e., the one minimizing the variances of the estimator and the corresponding confidence interval) could be found in closed form. We now present a simple extension of this result to the stratified case. For notational convenience, we use the shorthand $\nabla \ell_{k,\theta} := \nabla \ell_\theta(X, Y) \mid X \in \mathcal{A}_k$ and $\nabla \ell^f_{k,\theta} := \nabla \ell_\theta(X, f(X)) \mid X \in \mathcal{A}_k$.

Proposition 2. *Assume that $\tilde{\rho}_k$ and ρ_k are fixed. Then the tuning parameters $(\lambda^*_1, \ldots, \lambda^*_k)$, where*

$$\lambda^*_k = \frac{\mathrm{Tr}\left(A_w^{-1}(\mathrm{Cov}(\nabla \ell_{k,\theta^*}, \nabla \ell^f_{k,\theta^*}) + \mathrm{Cov}(\nabla \ell^f_{k,\theta^*}, \nabla \ell_{k,\theta^*}))A_w^{-1}\right)}{2\left(1 + \frac{n_k}{N_k}\right) \mathrm{Tr}\left(A_w^{-1} \mathrm{Cov}(\nabla \ell^f_{k,\theta^*}) A_w^{-1}\right)}, \tag{16}$$

minimize the cumulative asymptotic variance, $\mathrm{Tr}(\Sigma^\lambda_w)$.

Note that $\mathrm{Tr}(\Sigma^\lambda_w)$ is proportional to the total size of the confidence interval $\mathcal{C}^{\mathrm{SPP}}_\alpha$ in (14). Furthermore, as in [3], we can use plug-in estimates for the terms in (16) to compute a $\hat{\lambda}_k$, where $\hat{\lambda}_k \xrightarrow{p} \lambda^*_k$. From (16), we can see that λ^*_k is closely related to the correlation coefficient of the (curvature-scaled) gradients for minimizing the loss on an autorater label versus a true label. Intuitively, the more correlated these terms are, the more we can rely on the autorater labels for finding the true minimizer of $\mathbb{E}[\ell_\theta(X, Y)]$. For 1-$d$ mean estimation in particular, we can see that λ_k^* takes on a simple form:

Example 1 (λ_k^* for mean estimation)**.** *Consider the* 1-d *mean loss:* $\ell_\theta(x, y) = \frac{1}{2}(y - \theta)^2$. *Then:*

$$\lambda_k^* = \frac{\mathrm{Cov}(Y, f(X))}{(1 + \frac{n_k}{N_k})\mathrm{Var}(f(X))} \approx \frac{\mathrm{Cov}(Y, f(X))}{\mathrm{Var}(f(X))} \quad \textit{for large } N_k, \tag{17}$$

which is equivalent to the optimal linear regression coefficient, $\min_{\lambda_k} \mathbb{E}[|Y - \lambda_k f(X)|^2 \mid X \in \mathcal{A}_k]$.

4.3 Optimal allocation of the sampling budget

Our stratification approach has an additional hyperparameter ρ that can also be tuned to reduce variance. Recall that ρ_k determines the ratio between the labeled data size n_k for stratum k and the overall data size n (i.e., $\sum_k n_k = n$). $\tilde{\rho}_k$ is similarly defined for the unlabeled data. Any strictly positive values of ρ_k and $\tilde{\rho}_k$ are valid to be used, though not all values will improve performance. It turns out that the optimal ρ_k values can be exactly calculated, as the following proposition shows.

Proposition 3. *Assume that* λ_k *is fixed. Then the sampling rates* $(\rho_1^*, \dots, \rho_k^*)$ *and* $(\tilde{\rho}_1^*, \dots, \tilde{\rho}_k^*)$*, where*

$$\rho_k^* = \frac{w_k \sqrt{\mathrm{Tr}(A_w^{-1} V_{k,\Delta,\theta^*}^{\lambda_k} A_w^{-1})}}{\sum_{k'=1}^K w_{k'} \sqrt{\mathrm{Tr}(A_w^{-1} V_{k',\Delta,\theta^*}^{\lambda_{k'}} A_w^{-1})}} \quad \textit{and} \quad \tilde{\rho}_k^* = \frac{w_k \sqrt{\mathrm{Tr}(A_w^{-1} V_{k,f,\theta^*}^{\lambda_k} A_w^{-1})}}{\sum_{k'=1}^K w_{k'} \sqrt{\mathrm{Tr}(A_w^{-1} V_{k',f,\theta^*}^{\lambda_{k'}} A_w^{-1})}} \tag{18}$$

minimize the cumulative asymptotic variance, $\mathrm{Tr}(\Sigma_w^\lambda)$.

Although the solution for ρ_k^* is informative, it is not necessarily practical, as it depends on knowing $A_w^{-1} V_{k,\Delta,\theta^*}^{\lambda_k} A_w^{-1}$; this in turn depends on θ^* and $\mathbb{P}(X, Y)$, which are both unknown.[3] In the special case of mean estimation, however, it turns out there is no dependence on θ^*. To address the remaining dependence on $\mathbb{P}(X, Y)$, we propose to use autorater confidence scores, assuming they are available. Specifically, assume $\mathcal{Y}$ is discrete, and let $c(y \mid x)$ be the confidence of the autorater in label y given input x, where $c(y \mid x)$ approximates $\mathbb{P}(Y = y \mid X = x)$. This will result in the estimate for $\mathrm{Tr}(A_w^{-1} V_{k,\Delta,\theta^*}^{\lambda_k} A_w^{-1})$ below, which can then be plugged into the expression for ρ_k^* in Proposition 3.

Example 2 (ρ_k^* for mean estimation)**.** *Consider the* 1-d *mean loss:* $\ell_\theta(x, y) = \frac{1}{2}(y - \theta)^2$. *Then:*

$$\mathrm{Tr}(A_w^{-1} V_{k,\Delta,\theta^*}^{\lambda_k} A_w^{-1}) = \mathrm{Var}(Y - \lambda_k f(X) \mid X \in \mathcal{A}_k). \tag{19}$$

(19) *can then be estimated using the observed, but unlabeled, samples scored by the autorater for each stratum* k, $\tilde{X}_{ik} = \tilde{x}_{ik}$, $i = 1, \dots, N_k$, *as* $\mathrm{Var}(Y - \lambda_k f(X) \mid X \in \mathcal{A}_k) \approx \hat{\sigma}_k^2$, *where*

$$\hat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} \sum_{y \in \mathcal{Y}} c(y \mid \tilde{x}_{ik}) \left(y - \lambda_k f(\tilde{x}_{ik}) - \hat{\mu}_k\right)^2 \quad \textit{and} \tag{20}$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \sum_{y \in \mathcal{Y}} c(y \mid \tilde{x}_{ik}) \left(y - \lambda_k f(\tilde{x}_{ik})\right). \tag{21}$$

For d-dimensional data, the result is similar, but with a sum of d variances (one for each dimension). We also provide a simplified expression for $\hat{\sigma}_k^2$ in Appendix B when $f(\tilde{x}) := \sum_{y \in \mathcal{Y}} c(y \mid \tilde{x}) \cdot y$. Importantly, as we are free to use any $\rho_k > 0$, using this estimate still preserves the asymptotic coverage guarantees in (14), regardless of if the confidence estimate $c(y \mid x)$ is calibrated or not. Empirically, we show that using this heuristic can indeed lead to substantial improvements.

5 Experimental results

We compare our stratified estimator, `StratPPI`, to two baselines: (i) the classical estimate, which uses only the labeled data, S_n; and (ii) `PPI++`, which uses both S_n and $\tilde{S}_n$. All of our experiments focus on

[3] Similarly, the optimal solution to $\tilde{\rho}_k$ also depends on the unknown θ^* in general, though this term is less important to optimize if we assume N to be large. In practice, we always keep $\tilde{\rho}_k$ fixed to the natural rate, w_k.

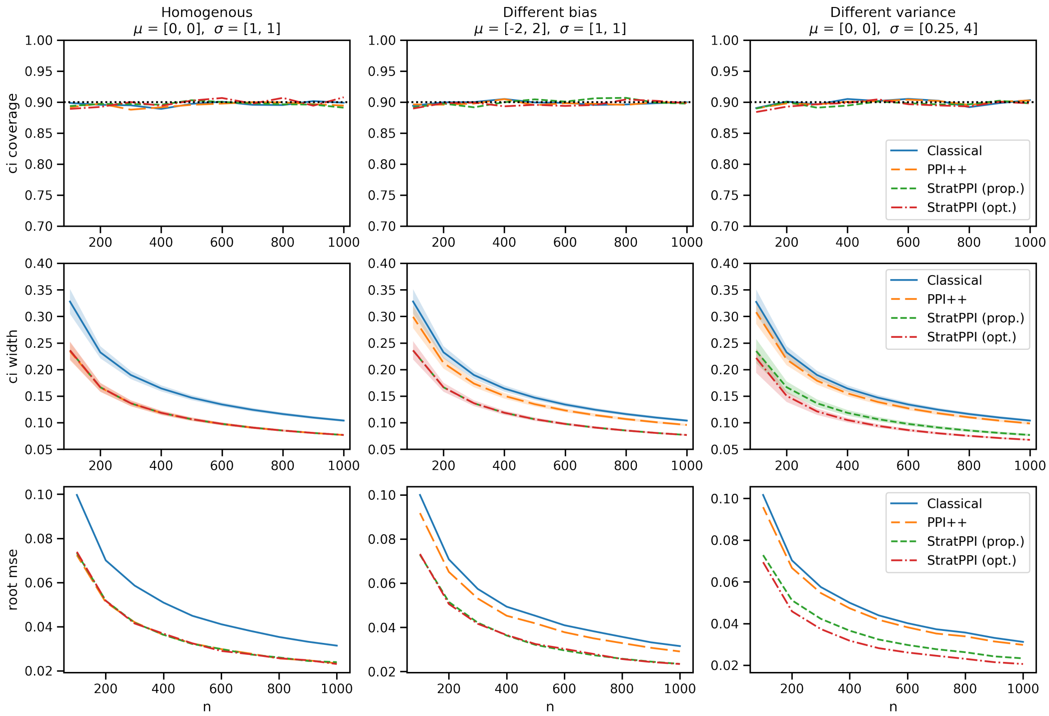

Figure 1: Mean estimation simulation study with $K = 2$ and $\alpha = 0.1$. The top row plots coverage (i.e., the fraction of the cases where the CI contained the true parameter value θ^*). The middle row plots the mean CI width (↓ is better). Shaded areas plot the $16/84$ quantiles across 5k trials. The bottom row plots the RMSE of $\hat{\theta}^{\text{SPP}}$ computed across the $5k$ trials, which shares the same trend with the mean CI width, as the estimator is unbiased. The left column shows a setting where strata are homogeneous, and `StratPPI` provides the no benefits over standard `PPI++` (but is not worse). The middle and right columns show heterogeneous settings where the autorater has either a different bias (μ) or variance (σ) per stratum, in which case `StratPPI` helps substantially. As strata variances are known, we only report proportional and optimal sample allocation results for `StratPPI`.

1-d mean estimation. We explore three different allocation strategies for `StratPPI`: the first is to set $\rho_k = w_k$ to be data proportional (`StratPPI Prop.`), the second is to set ρ_k optimally via the oracle $\rho_k = \rho_k^*$ (`StratPPI Opt.`), and the third is to use the approximation, $\rho_k \propto w_k \hat{\sigma}_k$, in Example 2 for $\lambda_k = 1$ when confidence scores are available (`StratPPI Heur.`). We use λ-tuning for both `PPI++` and `StratPPI`, as outlined in §4.2. Additional experimental results are given in Appendix C.

5.1 Simulation studies

We start with a simple synthetic experiment that is an analogue of §7.7.1 in [3]. Our goal is to estimate the mean outcome $\mathbb{E}[Y]$, where $Y \sim \mathcal{N}(0, 1)$. We assume that the input space $\mathcal{X}$ is partitioned into $K = 2$ strata, $(\mathcal{A}_1, \mathcal{A}_2)$, of equal mass $\mathbb{P}(X \in \mathcal{A}_1) = \mathbb{P}(X \in \mathcal{A}_2) = 0.5$. We then assume that predictions are formed as $f(X_{ik}) = Y_{ik} + \mu_k + \sigma_k \epsilon_{ik}$, where $\epsilon_{ik} \sim \mathcal{N}(0, 1)$. In other words, the predictions do not depend on the covariates X_{ik}, other than to reflect a stratum-specific noise σ_k and bias μ_k. We test three different scenarios: (i) where the two strata are homogeneous with $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$; (ii) where the two strata have different prediction biases, $\mu_1 \neq \mu_2$; and (iii) where the two strata have different prediction noise levels, $\sigma_1 \neq \sigma_2$. For each experiment, we sample $N = 10{,}000$ total predictions $f(\tilde{X})$ using $\tilde{\rho}_1 = \tilde{\rho}_2 = 0.5$, i.e., proportional to masses of the two hypothetical, equal-weight strata. We then vary the total number n of labeled examples Y, where the allocation is chosen according to ρ (which differs depending on if we are using `StratPPI Prop.` or `StratPPI Opt.`). We show results in Figure 1 for the mean confidence interval (CI) size and coverage (i.e., the fraction of the cases where the CI contained the true parameter value θ^*) of each method, averaged over $1k$ trials. As the plots in Figure 1 illustrate, when the underlying strata are homogeneous (left column), `StratPPI` behaves similar to `PPI++`. However, when the strata are heterogeneous (middle and right columns), `StratPPI` outperforms both baselines significantly—whereas `PPI++` becomes barely

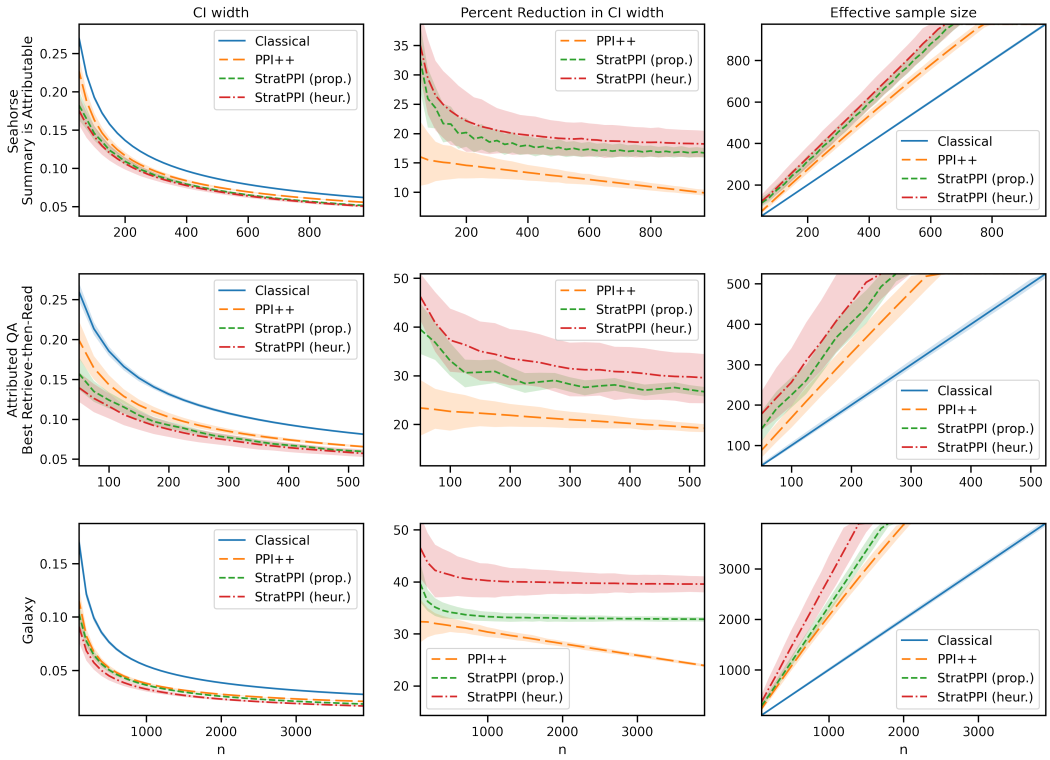

Figure 2: Mean estimation on real data with $K = 10$ and $\alpha = 0.05$. The x axis plots the number of human-labeled examples n; the y axis plots CI width, percent reduction in CI width against the classical estimate , and the effective sample size (the amount of human labels necessary to match the same confidence interval via classical inference). Shaded areas plot the $16/84$ quantiles across $1k$ trials. All `StratPPI` methods improve over classical inference and `PPI++`.

more powerful than the classical estimator. Additionally, we see that when the variance differs per stratum (right column), optimal allocation of the sampling budget indeed provides additional benefit.

5.2 Real data studies

We now demonstrate how our method performs on real datasets, where the underlying structure of the autorater is unknown. To partition $\mathcal{X}$, we choose to focus on stratifications that are based on the autorater's predictions, $f(X)$, based on the intuition that autorater performance can often differ across the type of predictions that it makes. Concretely, if the output space $\mathcal{Y}'$ of f is discrete, then we define $\mathcal{A} = \mathcal{Y}'$; otherwise we define $\mathcal{A}$ based on the equal-mass quantiles of $\mathcal{Y}'$ (which can be estimated by sampling a large set of unlabeled X and applying the autorater). We set $K = 10$. For all experiments, we plot performance as a function of n, where n is the number of human ratings our system is allowed to observe from each dataset. The remainder of the dataset (including any data points that are unlabeled, or labeled but with the labels removed) is used for the autorater sample $\tilde{S}_N$.

Seahorse. The `Seahorse` dataset [11] focuses on multilingual summarization. The authors considered generative models that output summaries for a document, and collected labels for many systems that cover serveral dimensions of summary quality. We focus on one quality dimension—whether the summary is fully attributable to the source document—and on one summarization system—a fine-tuned 13B parameter mT5 model [33]. The autorater models for each dimension are also mT5-XXL finetuned models, which output probability scores. The data contains 2727 examples for these two tasks, all of which have both human ratings as well as autorater scores.

AttributedQA. In attributed question answering [6], the goal of the QA system is to output both an answer, and a retrieved document that provides support for that answer. The system is only considered to be correct if the answer is both correct, and indeed supported by the linked document. We evaluate the highest-scoring "retrieve-and-read" QA system from this dataset,and define our autorater to be an 11B parameter T5 model [26] fine-tuned on a collection of natural language entailment tasks [15]. Like Seahorse, the model predicts a probability for whether or not the QA system gave an attributable answer. This dataset has 1000 human labels and 3000 autorater labels.

Galaxy. To demonstrate the generality of StratPPI beyond LLM-based settings, we also consider the Galaxy dataset [31], where the task is to estimate the fraction of spiral galaxies in the local universe. The autorater used here is a ResNet classifier applied to images from the Sloan Digital Sky Survey (SDSS) [34]. The classifier estimates the probability that the galaxy in question is spiral. We use $16{,}743$ observations from the dataset which contain both the human and autorater labels.

Methodology. We follow a procedure similar to [1; 3] to study CI estimates as a function of n. We report results on the following estimates: Classical inference, PPI++, StratPPI Prop. and StratPPI Heur. We do not report StratPPI Opt. since it is unknown for real data. For each value of n, we sample n of the cases with human labels at an allocation rate ρ_k (this rate is determined differently in StratPPI Prop. and StratPPI Heur.). As noted above, we construct the unlabeled dataset $\tilde{S}_N$ by joining the remaining labeled data (without utilizing the true labels) with any unlabeled data available for that dataset. We repeat this over 1000 trials, and for each trial we obtain the CI width. We report the mean of these widths in Figure 2, as well as the percent reduction in width over the classical inference baseline, $\frac{|\mathcal{C}_\alpha^{\text{classical}}|-|\mathcal{C}_\alpha^{\text{method}}|}{|\mathcal{C}_\alpha^{\text{classical}}|} \times 100\%$. We also report the "effective sample size", which we define as the number of samples required to obtain a CI of the same width as the method at sample size n when using the classical inference baseline instead.

Results. Figure 2 shows a large improvement for StratPPI methods over both PPI++ and classical inference for most datasets. Specifically, though all CIs decrease substantially in absolute size with n as expected, we see that improvements in the percent reduction in CI width over PPI++ can be as large as $0.10 \rightarrow 0.20$, $0.20 \rightarrow 0.30$, and $0.25 \rightarrow 0.35$ points in Seahorse, AttributedQA, and Galaxy, respectively. Furthermore, we can observe that many of the datasets exhibit heterogeneous characteristics, for which heuristic allocation helps considerably. In Galaxy, this accounts for a $+10$ percent reduction in CI width. In Seahorse and AttributedQA, the improvements are less strong but still clearly apparent. The practical implication of these results is that when limited by a human rating budget, StratPPI is able to produce an estimate of the mean with fewer human ratings via stratification and sampling allocation. For example, for the Seahorse dataset, we can see from the right column in Figure 2 that StratPPI Heur. with only 300 human ratings will be approximately as confident about the mean as the classical estimate that utilizes 600 human ratings, a factor of $2\times$.

6 Conclusion

As systems built on top of large language models continue to become more and more advanced, it becomes increasingly challenging to evaluate their performance using automatic tools. Manual labeling, on the other hand, is slow and expensive. Methods which therefore save on annotation cost are critical for reliably evaluating models, and knowing when they are improving—or degrading. Prediction-powered inference (PPI) is a promising class of such hybrid evaluation methods, since it can leveraged to provably produce statistically valid confidence intervals, while also effectively reducing the number of human labels needed to obtain intervals of certain width. Our results demonstrate that we can push PPI even further by introducing a method for performing even lower variance M-estimation by employing stratified sampling. In particular, we find that stratifications based on the predictions of the autoraters themselves proves to be an powerful stratification technique.

Limitations. While the confidence intervals produced by StratPPI (and PPI) enjoy coverage guarantees, these guarantees are asymptotic. When finite-sample performance is of particular importance, techniques that afford stronger guarantees might be preferred [2; 4; 5]. We also note some non-trivial aspects of the stratified sampling setup: (i) the number of strata has to be fixed; if K scales with n a more careful treatment is required; and (ii) the assumed observation model is different from traditional i.i.d. settings—we must be able to sample fixed sized samples from each partition; and (iii) performance may not be improved if the selected stratification is not statistically useful. Furthermore, in practice, the human data might have already been collected, in which case the studied stratified sampling setup does not directly apply. We leave the study of such post-stratified estimators to future work.

Broader impacts. This paper introduces a more powerful statistical method for evaluating LLMs, by merging human evaluations with autoraters in a way that is aware of subdomain differences. Our goal is to help power more reliable evaluations with lower annotation effort, both in terms of cost and time.

Acknowledgements. We thank Jacob Eisenstein, Jonathan Berant, Anastasios Angelopoulos, Taylan Cemgil, Arnoud Doucet, and Chris Dyer for helpful comments and discussions.

References

[1] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/abs/10.1126/science.adi6000.

[2] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint: 2208.02814*, 2023.

[3] Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint: 2311.01453*, 2023.

[4] Anastasios Nikolas Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint: 2110.01052*, 2021.

[5] Stephen Bates, Anastasios Nikolas Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution free, risk controlling prediction sets. *arXiv preprint: 2101.02703*, 2020.

[6] Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint: 2212.08037*, 2023.

[7] Pierre Boyeau, Anastasios N. Angelopoulos, Nir Yosef, Jitendra Malik, and Michael I. Jordan. Autoeval done right: Using synthetic data for model evaluation. *arXiv preprint: 2403.07008*, 2024.

[8] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.20. URL https://aclanthology.org/2022.emnlp-main.20.

[9] Arun Chaganty, Stephen Mussmann, and Percy Liang. The price of debiasing automatic metrics in natural language evalaution. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1060. URL https://aclanthology.org/P18-1060.

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

[11] Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.584. URL https://aclanthology.org/2023.emnlp-main.584.

[12] William G. Cochran. *Sampling Techniques, 3rd Edition.* John Wiley, 1977. ISBN 0-471-16240-X.

[13] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[14] R. Alex Hofer, Joshua Maynez, Bhuwan Dhingra, Adam Fisch, Amir Globerson, and William W. Cohen. Bayesian prediction-powered inference. *arXive preprint: 2405.06034*, 2024.

[15] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. *arXiv preprint: 2204.04991*, 2022.

[16] Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. *arXiv preprint: 2407.18370*, 2024.

[17] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.307. URL https://aclanthology.org/2023.acl-long.307.

[18] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. URL https://aclanthology.org/D16-1230.

[19] S.L. 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.

[20] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994. doi: https://doi.org/10.1016/S1573-4412(05)80005-4. URL https://www.sciencedirect.com/science/article/pii/S1573441205800054.

[21] Jerzy 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.

[22] Alexandru Niculescu-Mizil and Rich Caruana. Obtaining calibrated probabilities from boosting. In *UAI*, volume 5, pages 413–20, 2005.

[23] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1238. URL https://aclanthology.org/D17-1238.

[24] Richard Yuanzhe Pang, Vishakh Padmakumar, Thibault Sellam, Ankur Parikh, and He He. Reward gaming in conditional text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vlume 1: Long Papers)*, 2023.

[25] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[27] James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291135.

[28] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint: 2311.09476*, 2024.

[29] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[30] Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

[31] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 09 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt1458. URL https://doi.org/10.1093/mnras/stt1458.

[32] Jeffrey M. Wooldridge. Asymptotic properties of weighted m-estimators for standard stratified samples. *Econometric Theory*, 17:451 – 470, 2001. URL https://api.semanticscholar.org/CorpusID:123148497.

[33] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[34] Donald G. York, J. Adelman, Jr. John E. Anderson, Scott F. Anderson, James Annis, Neta A. Bahcall, J. A. Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, William N. Boroski, Steve Bracker, Charlie Briegel, John W. Briggs, J. Brinkmann, Robert Brunner, Scott Burles, Larry Carey, Michael A. Carr, Francisco J. Castander, Bing Chen, Patrick L. Colestock, A. J. Connolly, J. H. Crocker, István Csabai, Paul C. Czarapata, John Eric Davis, Mamoru Doi, Tom Dombeck, Daniel Eisenstein, Nancy Ellman, Brian R. Elms, Michael L. Evans, Xiaohui Fan, Glenn R. Federwitz, Larry Fiscelli, Scott Friedman, Joshua A. Frieman, Masataka Fukugita, Bruce Gillespie, James E. Gunn, Vijay K. Gurbani, Ernst de Haas, Merle Haldeman, Frederick H. Harris, J. Hayes, Timothy M. Heckman, G. S. Hennessy, Robert B. Hindsley, Scott Holm, Donald J. Holmgren, Chi hao Huang, Charles Hull, Don Husby, Shin-Ichi Ichikawa, Takashi Ichikawa, Željko Ivezić, Stephen Kent, Rita S. J. Kim, E. Kinney, Mark Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, John Korienek, Richard G. Kron, Peter Z. Kunszt, D. Q. Lamb, B. Lee, R. French Leger, Siriluk Limmongkol, Carl Lindenmeyer, Daniel C. Long, Craig Loomis, Jon Loveday, Rich Lucinio, Robert H. Lupton, Bryan MacKinnon, Edward J. Mannery, P. M. Mantsch, Bruce Margon, Peregrine McGehee, Timothy A. McKay, Avery Meiksin, Aronne Merelli, David G. Monet, Jeffrey A. Munn, Vijay K. Narayanan, Thomas Nash, Eric Neilsen, Rich Neswold, Heidi Jo Newberg, R. C. Nichol, Tom Nicinski, Mario Nonino, Norio Okada, Sadanori Okamura, Jeremiah P. Ostriker, Russell Owen, A. George Pauls, John Peoples, R. L. Peterson, Donald Petravick, Jeffrey R. Pier, Adrian Pope, Ruth Pordes, Angela Prosapio, Ron Rechenmacher, Thomas R. Quinn, Gordon T. Richards, Michael W. Richmond, Claudio H. Rivetta, Constance M. Rockosi, Kurt Ruthmansdorfer, Dale Sandford, David J. Schlegel, Donald P. Schneider, Maki Sekiguchi, Gary Sergey, Kazuhiro Shimasaku, Walter A. Siegmund, Stephen Smee, J. Allyn Smith, S. Snedden, R. Stone, Chris Stoughton, Michael A. Strauss, Christopher Stubbs, Mark SubbaRao, Alexander S. Szalay, Istvan Szapudi, Gyula P. Szokoly, Anirudda R. Thakar, Christy Tremonti, Douglas L. Tucker, Alan Uomoto, Dan Vanden Berk, Michael S. Vogeley, Patrick Waddell, Shu i Wang, Masaru Watanabe, David H. Weinberg, Brian Yanny, and Naoki Yasuda. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, sep 2000. doi: 10.1086/301513. URL https://dx.doi.org/10.1086/301513.

[35] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion 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.

[36] Tijana Zrnic and Emmanuel 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.

[37] Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. *arXiv preprint: 2403.03208*, 2024.

A Proofs

We begin by defining the basic regularity conditions that we assume ℓ_θ satisfies:

Definition A.1 (Regularity conditions of ℓ_θ). *Assume that*

(i) Θ *is a compact subset of* $\mathbb{R}^d$*;*

(ii) The minimizer $\theta^* \in \operatorname{int}(\Theta)$ *is unique with* $\mathbb{E}[\nabla \ell_{\theta^*}(X, Y)] = 0$*;*

(iii) For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$*,* $\ell_\theta(x, y)$ *is twice continuously differentiable on* $\operatorname{int}(\Theta)$*;*

(iv) For all $\theta \in \Theta$*,* $\ell_\theta(X, Y)$*,* $\frac{\partial \ell_\theta(X,Y)}{\partial \theta_i}$*, and* $\frac{\partial^2 \ell_\theta(X,Y)}{\partial \theta_i \partial \theta_j}$ *have finite expectations;*

(v) $\mathbb{E}[\nabla^2 \ell_\theta(X, Y)]$ *is non-singular.*

These regularity conditions are fairly mild—for example, it is straightforward to verify that the loss of the mean estimator, $\frac{1}{2}\|y - \theta\|^2$, satisfies Definition A.1. We note, however, that `PPI++` can be applied to merely "smooth enough" losses, which also include non-continuously differentiable functions losses like the quantile loss. As we primarily focus on mean estimation in this work, for simplicity of analysis we only consider (the still broad class) of losses satisfying Definition A.1, though our results can be expected to readily extend to the more general case following a similar treatment as in [3; 29].

To support our theoretical results, we also provide the following two lemmas.

Lemma A.1 (Slutsky's Theorem, general form). *Let* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$*, where* X_n*,* X*, and* Y_n *are random vectors, and c is a constant vector. Then for any continuous function g,* $g(X_n, Y_n) \xrightarrow{d} g(X, c)$*.*

Proof. By Theorem 2.7 of [29] we have $(X_n, Y_n) \xrightarrow{d} (X, c)$. The continuous mapping theorem then implies that $g(X_n, Y_n) \xrightarrow{d} g(X, c)$. □

Lemma A.2. *Under the regularity conditions of Definition A.1, we have that* $\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}} \xrightarrow{p} \theta^*$*.*

Proof. Under the regularity conditions of Definition A.1 and the fact that $\hat{\lambda}_k \xrightarrow{p} \lambda_k$ where λ_k is constant, the uniform weak law of large numbers can be applied to each term in $L^{\mathrm{SPP}}_{\hat{\lambda}}(\theta)$ so that

$$\sup_{\theta \in \Theta} \left\| L^{\mathrm{SPP}}_{\hat{\lambda}}(\theta) - \mathbb{E}[\ell_\theta(X, Y)] \right\| \xrightarrow{p} 0.$$

As θ^* is unique, Θ is compact, and ℓ_θ is continuous it follows from Theorem 2.1 of [20] that

$$\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}} \xrightarrow{p} \theta^*.$$

□

A.1 Proof of Theorem 2

Proof. For ease of notation, we will define

$$\tilde{L}^f_{N_k}(\theta) = \frac{1}{N_k} \sum_{i=1}^{N_k} \ell_\theta(\tilde{X}_{ik}, f(\tilde{X}_{ik})$$

$$L_{n_k}(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_\theta(X_{ik}, Y_{ik})$$

$$L^f_{n_k}(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_\theta(X_{ik}, f(X_{ik})).$$

We will also use the following shorthand for the conditional gradients:

$$\nabla \ell_{k,\theta} := \nabla \ell_\theta(X, Y) \mid X \in \mathcal{A}_k \quad \text{and}$$

$$\nabla \ell^f_{k,\theta} := \nabla \ell_\theta(X, f(X)) \mid X \in \mathcal{A}_k.$$

As samples are i.i.d. within each stratum, the CLT gives

$$\sqrt{N_k}\Big(\nabla\tilde{L}^f_{N_k}(\theta^*)-\mathbb{E}[\nabla\ell^f_{k,\theta^*}]\Big)\xrightarrow{d}\mathcal{N}\left(0,\mathrm{Cov}\left(\nabla\ell^f_{k,\theta^*}\right)\right), \tag{22}$$

and

$$\sqrt{n_k}\begin{bmatrix}\nabla L_{n_k}(\theta^*)-\mathbb{E}[\nabla\ell_{k,\theta^*}]\\ \nabla L^f_{n_k}(\theta^*)-\mathbb{E}[\nabla\ell^f_{k,\theta^*}]\end{bmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\mathrm{Cov}\left(\begin{bmatrix}\nabla\ell_{k,\theta^*}\\ \nabla\ell^f_{k,\theta^*}\end{bmatrix}\right)\right). \tag{23}$$

Since samples S_n and $\tilde{S}_N$ are independent from each other, we also have that (22) and (23) converge jointly. Applying Lemma A.1 for the following continuous function of $\hat{\lambda}_k\xrightarrow{p}\lambda_k$ then gives

$$\begin{aligned}\sqrt{n}&\left(\hat{\lambda}_k\cdot\mathcal{N}\Big(\nabla\tilde{L}^f_{N_k}(\theta^*)-\mathbb{E}[\nabla\ell^f_{k,\theta^*}]\Big)+\begin{bmatrix}1\\-\hat{\lambda}_k\end{bmatrix}^\top\begin{bmatrix}\nabla\tilde{L}_n(\theta^*)-\mathbb{E}[\nabla\ell_{k,\theta^*}]\\ \nabla\tilde{L}^f_{n_k}(\theta^*)-\mathbb{E}[\nabla\ell^f_{k,\theta^*}]\end{bmatrix}\right)\\ &\xrightarrow{d}\mathcal{N}\left(0,\lambda_k^2\frac{n}{N_k}\mathrm{Cov}(\nabla\ell^f_{k,\theta^*})\right)+\mathcal{N}\left(0,\frac{n}{n_k}\mathrm{Cov}(\nabla\ell_{k,\theta^*}-\lambda_k\nabla\ell^f_{k,\theta^*})\right) \qquad (24)\\ &=\mathcal{N}\left(0,\frac{r}{\tilde{\rho}_k}\cdot V^{\lambda_k}_{k,f,\theta^*}+\frac{1}{\rho_k}V^{\lambda_k}_{k,\Delta,\theta^*}\right). \qquad (25)\end{aligned}$$

Since θ^* satisfies $\mathbb{E}[\nabla\ell_{\theta^*}]=0$, by the law of total expectation

$$\sum_{k=1}^K w_k\mathbb{E}[\nabla\ell_{k,\theta^*}]=\mathbb{E}[\nabla\ell_{\theta^*}]=0. \tag{26}$$

Using this fact, we can write

$$\nabla L^{\mathrm{SPP}}_{\hat{\lambda}}(\theta)=\sum_{k=1}^K w_k\left(\nabla\left(\hat{\lambda}_k\tilde{L}^f_{N_k}(\theta)+L_{n_k}(\theta)-\hat{\lambda}_kL^f_{n_k}(\theta)\right)-\mathbb{E}[\nabla\ell_{k,\theta^*}]\right). \tag{27}$$

Since samples across the K (where here K is a constant) stratas are independent, combining the results of (25) with (27) yields

$$\sqrt{n}\nabla L^{\mathrm{SPP}}_{\hat{\lambda}}(\theta^*)\xrightarrow{d}\mathcal{N}(0,B^\lambda_w). \tag{28}$$

Applying the mean value theorem, we have

$$\nabla L^{\mathrm{SPP}}_{\hat{\lambda}}(\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}})=\nabla L^{\mathrm{SPP}}_{\hat{\lambda}}(\theta^*)+\nabla^2L^{\mathrm{SPP}}_{\hat{\lambda}}(\tilde{\theta})(\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}-\theta^*) \tag{29}$$

for some $\tilde{\theta}$ between $\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}$ and θ^*. Let $A^\dagger$ represent the pseudoinverse of A. Then

$$\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}-\theta^*=\nabla^2L^{\mathrm{SPP}}_{\hat{\lambda}}(\tilde{\theta})^\dagger\left(\nabla L^{\mathrm{SPP}}_{\hat{\lambda}}(\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}})-\nabla L^{\mathrm{SPP}}_{\hat{\lambda}}(\theta^*)\right). \tag{30}$$

As $\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}\xrightarrow{p}\theta^*$ per Lemma A.2, we have $\tilde{\theta}\xrightarrow{p}\theta^*$. Under the regularity conditions of Definition A.1 and the fact that $\hat{\lambda}_k\xrightarrow{p}\lambda_k$, an application of the uniform weak law of large numbers and the continuous mapping theorem then gives

$$\nabla^2L^{\mathrm{SPP}}_{\hat{\lambda}}(\tilde{\theta})^\dagger\xrightarrow{p}\nabla^2\mathbb{E}[\ell_{\theta*}]^{-1}, \tag{31}$$

which is a constant term. Furthermore, by the law of total expectation,

$$\mathbb{E}[\nabla^2\ell_{\theta^*}]^{-1}=\left(\sum_{k=1}^K w_kH_{k,\theta^*}\right)^{-1}=A_w^{-1}. \tag{32}$$

Finally, the fact that $\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}\xrightarrow{p}\theta^*$ and $\theta^*\in\mathrm{int}(\Theta)$, together with the fact that $\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}}$ is a minimum of $L^{\mathrm{SPP}}_{\hat{\lambda}}$, implies that $\nabla L^{\mathrm{SPP}}_{\hat{\lambda}}(\hat{\theta}^{\mathrm{SPP}}_{\hat{\lambda}})\xrightarrow{p}0$. Combining these results with (28) via Lemma A.1 and the fact that A_w^{-1} is symmetric for twice continuously differentiable ℓ_θ gives

$$\sqrt{n}(\hat{\theta}-\theta^*)\xrightarrow{d}\mathcal{N}(0,A_w^{-1}B^\lambda_w(A_w^{-1})^\top)=\mathcal{N}(0,A_w^{-1}B^\lambda_wA_w^{-1}). \tag{33}$$

□

A.2 Proof of Corollary 2

Proof. The regularity conditions of Definition A.1 allow us to apply Lemma 4.3 of [20] to show that each of the plug-in estimates for each stratum term satisfies

$$\widehat{H}_{k,\hat{\theta}^{\mathrm{SPP}}} \xrightarrow{p} H_{k,\theta^*} \quad \text{and} \quad \widehat{V}^{\lambda_k}_{k,f,\hat{\theta}^{\mathrm{SPP}}} \xrightarrow{p} \widehat{V}^{\lambda_k}_{k,f,\theta^*} \quad \text{and} \quad \widehat{V}^{\lambda_k}_{k,\Delta,\hat{\theta}^{\mathrm{SPP}}} \xrightarrow{p} \widehat{V}^{\lambda_k}_{k,\Delta,\theta^*}$$

which implies that the weighted plug-in estimate for $\hat{\Sigma}^{\hat{\lambda}}_w$ satisfies $\hat{\Sigma}^{\hat{\lambda}}_w \xrightarrow{p} \Sigma^{\lambda}_w$. (14) is thus equivalent to taking the $\left(\frac{\alpha}{2}, 1-\frac{\alpha}{2}\right)$ quantiles of the asymptotic normal distribution of $\hat{\theta}^{\mathrm{SPP}}$, implying (15). □

A.3 Proof of Proposition 1

Proof. In the special case of the square loss, the Hessian is the identity matrix. Therefore, simplifying and applying the linearity of the trace, we have

$$\begin{aligned}
\mathrm{Tr}(\Sigma^{\lambda}_w) &= \sum_{k=1}^{K} w_k^2 \mathrm{Tr}\left(\frac{r}{\tilde{\rho}_k} \cdot V^{\lambda_k}_{k,f,\theta^*} + \frac{1}{\rho_k} V^{\lambda_k}_{k,\Delta,\theta^*}\right) && (34)\\
&= \sum_{k=1}^{K} w_k \mathrm{Tr}\left(r \cdot V^{\lambda_k}_{k,f,\theta^*} + V^{\lambda_k}_{k,\Delta,\theta^*}\right) && (35)\\
&= r\sum_{k=1}^{K} w_k \mathrm{Tr}\left(V^{\lambda_k}_{k,f,\theta^*}\right) + \sum_{k=1}^{K} w_k \mathrm{Tr}\left(V^{\lambda_k}_{k,\Delta,\theta^*}\right) && (36)\\
&= r\sum_{j=1}^{d}\sum_{k=1}^{K} w_k V^{\lambda_k}_{k,f,\theta^*,jj} + \sum_{j=1}^{d}\sum_{k=1}^{K} w_k V^{\lambda_k}_{k,\Delta,\theta^*,jj} && (37)
\end{aligned}$$

where we include the subscript jj to index the diagonal variance terms of both covariance matrices. For ease of notation, denote the conditional variances as

$$\begin{aligned}
&V_{f,j} \mid Z = k := \lambda^2 \nabla \ell_{\theta_j^*}(X, f(X)) \mid X \in \mathcal{A}_k \quad \text{and}\\
&V_{\Delta,j} \mid Z = k := \nabla \ell_{\theta_j^*}(X, Y) - \lambda \nabla \ell_{\theta_j^*}(X, f(X)) \mid X \in \mathcal{A}_k.
\end{aligned}$$

Then we can write (37) as

$$r\sum_{j=1}^{d}\sum_{k=1}^{K} w_k V^{\lambda_k}_{k,f,\theta^*,jj} + \sum_{j=1}^{d}\sum_{k=1}^{K} w_k V^{\lambda_k}_{k,\Delta,\theta^*,jj} = \sum_{j=1}^{d} r\mathbb{E}[\mathrm{Var}(V_{f,j} \mid Z)] + \mathbb{E}[\mathrm{Var}(V_{\Delta,j} \mid Z)], \tag{38}$$

and apply the law of total variance to get

$$\begin{aligned}
\mathrm{Tr}(\Sigma^{\lambda}_w) &= \sum_{j=1}^{d} r\mathbb{E}[\mathrm{Var}(V_{f,j} \mid Z)] + \mathbb{E}[\mathrm{Var}(V_{\Delta,j} \mid Z)] && (39)\\
&= \sum_{j=1}^{d} r(\mathrm{Var}(V_{f,j}) - \mathrm{Var}(\mathbb{E}[V_{f,j} \mid Z])) + (\mathrm{Var}(V_{\Delta,j}) - \mathrm{Var}(\mathbb{E}[V_{\Delta,j} \mid Z])) && (40)\\
&\leq \sum_{j=1}^{d} r\mathrm{Var}(V_{f,j}) + \mathrm{Var}(V_{\Delta,j}) && (41)\\
&= \mathrm{Tr}(\Sigma^{\lambda}). && (42)
\end{aligned}$$

Finally, (41) holds with equality iff both $\mathrm{Var}(\mathbb{E}[V_{f,j} \mid Z]) = 0$ and $\mathrm{Var}(\mathbb{E}[V_{\Delta,j} \mid Z]) = 0$, which, combined with the assumption that $\mathbb{P}(Z = k) = w_k > 0$ for all k, is only satisfied when both $\mathbb{E}[V_{f,j} \mid Z]$ and $\mathbb{E}[V_{\Delta,j} \mid Z]$ are constants. □

A.4 Proof of Proposition 2

Proof. By linearity of B_w^λ, we can rewrite $\Sigma_w^\lambda = A_w^{-1} B_w^\lambda A_w^{-1}$ as

$$\Sigma_w^\lambda = \sum_{k=1}^{K} w_k^2 A_w^{-1} \left(\frac{r}{\tilde{\rho}_k} \cdot V_{k,f,\theta^*}^{\lambda_k} + \frac{1}{\rho_k} V_{k,\Delta,\theta^*}^{\lambda_k} \right) A_w^{-1} \tag{43}$$

$$= \sum_{k=1}^{K} \frac{w_k^2}{\rho_k} A_w^{-1} \left(\frac{r\rho_k}{\tilde{\rho}_k} \cdot V_{k,f,\theta^*}^{\lambda_k} + V_{k,\Delta,\theta^*}^{\lambda_k} \right) A_w^{-1} \tag{44}$$

$$= \sum_{k=1}^{K} \frac{w_k^2}{\rho_k} A_w^{-1} \left(\frac{n_k}{N_k} \cdot V_{k,f,\theta^*}^{\lambda_k} + V_{k,\Delta,\theta^*}^{\lambda_k} \right) A_w^{-1}. \tag{45}$$

By linearity of the trace, we then also have

$$\mathrm{Tr}(\Sigma_w^\lambda) = \sum_{k=1}^{K} \frac{w_k^2}{\rho_k} \mathrm{Tr} \left(A_w^{-1} \left(\frac{n_k}{N_k} \cdot V_{k,f,\theta^*}^{\lambda_k} + V_{k,\Delta,\theta^*}^{\lambda_k} \right) A_w^{-1} \right). \tag{46}$$

As each term in the sum is independent, we can minimize the sum by minimizing each individual term, i.e.,

$$\lambda_k^* = \operatorname*{argmin}_{\lambda_k} \mathrm{Tr} \left(A_w^{-1} \left(\frac{n_k}{N_k} \cdot V_{k,f,\theta^*}^{\lambda_k} + V_{k,\Delta,\theta^*}^{\lambda_k} \right) A_w^{-1} \right) \tag{47}$$

for $k = 1, \ldots, K$. We can then directly apply Proposition 2 of [3] to get

$$\lambda_k^* = \frac{\mathrm{Tr}\left(A_w^{-1} (\mathrm{Cov}(\nabla \ell_{k,\theta^*}, \nabla \ell_{k,\theta^*}^f) + \mathrm{Cov}(\nabla \ell_{k,\theta^*}^f, \nabla \ell_{k,\theta^*})) A_w^{-1} \right)}{2(1 + r_k) \mathrm{Tr}\left(A_w^{-1} \mathrm{Cov}(\nabla \ell_{k,\theta^*}^f) A_w^{-1} \right)}, \tag{48}$$

where $r_k = n_k / N_k$. □

A.5 Proof of Proposition 3

Proof. Following the same derivation as Proposition 2, we can rewrite

$$\mathrm{Tr}(\Sigma_w^\lambda) = \sum_{k=1}^{K} w_k^2 \mathrm{Tr} \left(A_w^{-1} \left(\frac{r}{\tilde{\rho}_k} \cdot V_{k,f,\theta^*}^{\lambda_k} + \frac{1}{\rho_k} V_{k,\Delta,\theta^*}^{\lambda_k} \right) A_w^{-1} \right) \tag{49}$$

$$= r \sum_{k=1}^{K} \frac{w_k^2}{\tilde{\rho}_k} \mathrm{Tr} \left(A_w^{-1} V_{k,f,\theta^*}^{\lambda_k} A_w^{-1} \right) + \sum_{k=1}^{K} \frac{w_k^2}{\rho_k} \mathrm{Tr} \left(A_w^{-1} V_{k,\Delta,\theta^*}^{\lambda_k} A_w^{-1} \right), \tag{50}$$

and therefore can optimize $\tilde{\rho}_k$ and ρ_k independently.

We start with ρ_k. Let $z_k = \mathrm{Tr}\left(A_w^{-1} V_{k,\Delta,\theta^*}^{\lambda_k} A_w^{-1} \right)$. We then solve the constrained optimization problem

$$\text{minimize} \quad \sum_{k=1}^{K} \frac{w_k^2}{\rho_k} z_k \quad \text{s.t.} \quad \sum_{k=1}^{K} \rho_k = 1, \rho_k \geq 0. \tag{51}$$

Turning this into a Lagrangian with additional slack variables $s_k \geq 0$, we have

$$\mathcal{J}(\rho, \mu, \eta, s) = \sum_{k=1}^{K} \frac{w_k^2}{\rho_k} z_k + \mu \left(\sum_{i=1}^{K} \rho_k - 1 \right) + \sum_{k=1}^{K} \eta_k (\rho_k - s_k^2) \tag{52}$$

Setting $\nabla \mathcal{J}(\rho, \mu, \eta) = 0$ and solving for ρ_k then gives:

$$\frac{\partial \mathcal{J}}{\partial \rho_k} = -\frac{w_k^2}{\rho_k^2} z_k + \mu + \eta_k = 0 \tag{53}$$

$$\frac{\partial \mathcal{J}}{\partial \mu} = 1 - \sum \rho_k = 0 \tag{54}$$

$$\frac{\partial \mathcal{J}}{\partial \eta_k} = \rho_k - s_k^2 = 0 \tag{55}$$

$$\frac{\partial \mathcal{J}}{\partial s_k} = -2\eta_k s_k = 0 \tag{56}$$

Assume that the inequality constraint is inactive, and $\eta_k = 0$. Solving for ρ_k^*,

$$\rho_k^* = s_k^2 = \sqrt{\frac{w_k^2 z_k}{\mu}} \implies \sum_{k=1}^{K} \sqrt{\frac{w_k^2 z_k}{\mu}} = 1 \tag{57}$$

$$\implies \sqrt{\mu} = \sum_{k=1}^{K} \sqrt{w_k^2 z_k} \tag{58}$$

$$\implies \rho_k^* = \frac{w_k \sqrt{z_k}}{\sum_{k'=1}^{K} w_{k'} \sqrt{z_{k'}}}. \tag{59}$$

We can now verify that the constraint is inactive, with $\rho_k^* \geq 0$, since $\sum_{k=1}^{K} w_k = 1, w_k \geq 0$ by definition of a valid probability distribution, and we have $z_k \geq 0$ since it is a sum of non-negative variance terms.

The same analysis can then be applied to $\tilde{\rho_k}$, but for $z_k = \operatorname{Tr}\left(A_w^{-1} V_{k,f,\theta^*}^{\lambda_k} A_w^{-1}\right)$. □

B A simplified estimate of the sample allocation for mean estimation

In the setting of Example 2 (i.e., 1-d mean estimation), assume that $c(y \mid x) \approx \mathbb{P}(Y = y \mid X = x)$ is a confidence estimate for label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is discrete. Then, if we define the autorater as

$$f(x) = \sum_{y \in \mathcal{Y}} c(y \mid x) \cdot y \approx \mathbb{E}[Y \mid X = x], \tag{60}$$

the term $\operatorname{Var}(Y - \lambda_k f(X) \mid X \in \mathcal{A}_k)$ further simplifies to $\approx \operatorname{Var}(Y \mid X \in \mathcal{A}_k)$ for any value of λ_k. Applying the law of total variance, we can conveniently write $\operatorname{Var}(Y \mid X \in \mathcal{A}_k)$ as

$$\operatorname{Var}(Y \mid X \in \mathcal{A}_k) = \mathbb{E}[\operatorname{Var}(Y \mid X) \mid X \in \mathcal{A}_k] + \operatorname{Var}(\mathbb{E}[Y \mid X] \mid X \in \mathcal{A}_k), \tag{61}$$

which can be empirically estimated as $\hat{\sigma}_k$ on unlabeled autorater data, $\tilde{X}_{ik}$, $i = 1, \ldots, N_k$, via

$$\hat{\sigma}_k^2 = \widehat{\mathbb{E}}_{N_k}\left[\sum_{y \in \mathcal{Y}} c(y \mid \tilde{X}_{ik}) \cdot y - f(\tilde{X}_{ik}))\right] + \widehat{\operatorname{Var}}_{N_k}(f(\tilde{X}_{ik})), \tag{62}$$

where $\widehat{E}_{N_k}$ and $\widehat{\operatorname{Var}}_{N_k}$ denote the empirical mean and variance over all $\tilde{X}_{ik}$, respectively. Lastly, when $\mathcal{Y}$ is binary (which is the case in all of our experiments in §5.2), this becomes

$$\hat{\sigma}_k^2 = \widehat{\mathbb{E}}_{N_k}\left[f(\tilde{X}_{ik})(1 - f(\tilde{X}_{ik}))\right] + \widehat{\operatorname{Var}}_{N_k}(f(\tilde{X}_{ik})), \tag{63}$$

which can be easily calculated in Python as `np.mean(yhat * (1 - yhat)) + np.var(yhat)`.

C Additional experimental results

We also evaluate `StratPPI` on the Chatbot Arena dataset [10], in which we evaluate the win-rate of `gpt-4-1106-preview` over `claude-2.1`.[4] See Figure 3. Specifically, we use the standard Chatbot

[4] https://huggingface.co/datasets/lmsys/lmsys-arena-human-preference-55k

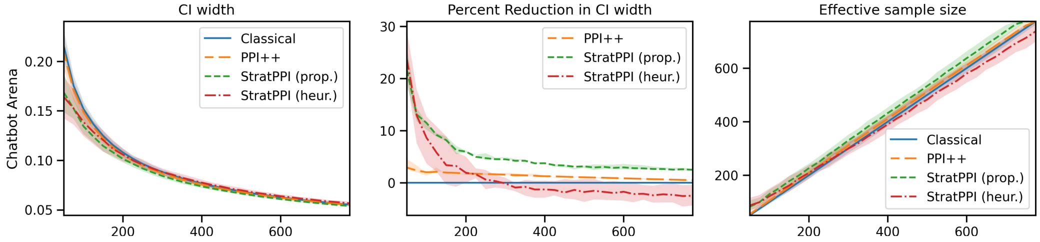

Figure 3: Win-rate experiment on Chatbot Arena for `gpt-4-1106-preview` vs. `claude-2.1`. Scores are based on the average label ('better' = 1 vs. 'worse' = 0) over 10 samples from Gemini Ultra acting as a LLM judge. Interestingly, as these confidence scores are not calibrated, our heuristic becomes overly aggressive at higher n. Future work can explore how to best incorporate additional regularization into the estimated optimal sampling ratios ρ.

Arena auto-eval prompt[5] and map `[[A≫B]]`, `[[A>B]` to 1, `[[A=B]]` to 0.5, and `[[B≫A]]`, `[[B>A]]` to 0. We then do self-consistency sampling and take the average of 10 samples from Gemini Ultra. This is used as both our final autorater estimate and confidence. We find that stratification also helps in this setting. However, in line with prior and contemporary work [e.g., 35; 16] we found confidence scores from the LLM-as-a-judge to be fairly uncalibrated, which makes our heuristic allocation less effective at larger n (though still effective at small labeled sample sizes n). Investigating robust heuristics in the face of miscalibration (e.g., via regularization or recalibration) would likely help, and is a direction for future work.

[5] https://github.com/lm-sys/arena-hard-auto/blob/main/config/judge_config.yaml

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: All claims are supported in the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We address this in the conclusion.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

Justification: Proofs are included in the appendix.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: Results are based on publically available data, and simulations, and all procedures are described.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The data is publicly available, and the methodology is simple to implement. Pseudocode is given. Code may be made available at a future date.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: Detailed information is provided.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: Error bars are provided for the main experimental results.

   Guidelines:

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

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Compute resources required are very light, as no model training is performed.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: No issues.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: The paper discusses the advantages of the method for more effective development of large language models.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: NA

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: NA

    Guidelines:

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

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

13. **New Assets**

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

    Answer: [NA]

    Justification: NA

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: NA

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.