ROBUST CONFIDENCE INTERVALS USING SIDE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how to construct implementable confidence intervals for the mean of a distribution with unknown, finite variance when an adversary may corrupt a constant fraction of the samples. Although robust mean estimation has been extensively studied, implementable confidence intervals remain elusive. We first observe that no asymptotically valid confidence interval of nontrivial width is possible without additional restrictions or side information. Our main result is that side information, in the form of machine learning predictions with bounded mean squared error (MSE), can enable confidence interval computation. At the core of our method is a reduction of variance estimation to an optimization problem. We prove that although the exact version of this problem is NP-hard, there is a polynomial-time additive approximation algorithm. Based on this result we develop an almost-linear time algorithm that provides an upper bound on the sample variance. We use this upper bound in our novel midpoint estimator and these together yield asymptotically valid confidence intervals uniformly over admissible adversaries. We complement our theoretical study with a series of experiments on the dataset introduced in Cen et al. (2025). We evaluate our method on real-world, bounded data, examining coverage and interval length across multiple corruption levels. Under two corruption strategies that directly try to distort the clean mean or variance of the observed samples, our method maintains coverage and non-trivial length under any considered corruption level less than 50%.

1 INTRODUCTION

The fragility of classical statistical estimators to tiny levels of contamination in the sampling distribution is a well-known phenomenon, whose formal study dates back to several seminal papers starting in the 1960s (Tukey, 1960; Anscombe, 1960; Huber, 1964). The massive amount of data in modern data analysis settings makes data curation even more challenging and hence, the need for robust estimators has only become more salient. At the same time, the rapid growth of generative AI has reduced the cost of adversarially corrupting public datasets at scale (Meng et al., 2025; Williams et al., 2025). One clean example arises in the context of online survey research, which companies and social scientists alike rely on to measure public opinion. Recent work by Westwood (2025) shows that LLM-based agents can reliably evade screening tools designed to detect nonhuman respondents. An ambitious adversary could therefore deploy a modest number of agents to submit fraudulent survey responses, with the goal of skewing estimates and ultimately influencing policy decisions. In recent years, substantial effort has been devoted to exploring several corruption models, e.g., Huber's contamination model (Diakonikolas et al., 2023; Mu & Xiong, 2023) or the coordinate-level corruption model (Liu et al., 2021), and developing estimators that are more robust to such corruption (Canonne et al., 2023; Bhatt et al., 2022; Diakonikolas et al., 2019; Hopkins & Li, 2019; Lugosi & Mendelson, 2021; Lee et al., 2025) (see Section B for an extended discussion of works on robust estimation). In this work, we focus on a very general adversarial model, captured by the definition below, applied to the fundamental problem of *mean estimation*.

Definition 1 (Strong Contamination Model)**.** *Let $\eta \in [0, 1/2]$ be a corruption parameter, and let P be the ground truth distribution. First, n samples are drawn independently from P, then an adversary is allowed to inspect all n samples and arbitrarily alter up to an η fraction of them.*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Strong contamination is a popular model in robust statistics but virtually all the existing literature is concerned with *point estimation*, i.e., producing an estimator $\widehat{\mu}$ whose distance from the true mean $\mu = \mathbb{E}_{Y\sim P}[Y]$ is small with high probability. Comparatively little attention has been paid to the task of producing *confidence intervals* for μ. The two tasks are related but not interchangeable. A point estimate answers "what is my best guess for μ?"; a confidence interval answers "what is the set of values of μ that the data is statistically consistent with?". The second question is the statistical primitive of scientific discovery: hypothesis tests, A/B-test conclusions, regulatory thresholds, and meta-analyses are all expressed in terms of confidence intervals and the p-values they induce. In a world in which the underlying data are increasingly subject to adversarial manipulation, the credibility of empirical findings depends on whether we can still attach valid uncertainty quantifications to their conclusions. We are therefore led to ask the following.

Question 1. Can we construct asymptotically valid, nontrivial confidence intervals for the mean $\mu = \mathbb{E}[Y]$ in the strong contamination model?

In the absence of corruption, the central limit theorem and Slutsky's theorem justify the familiar z-interval based on the sample mean and variance. Under adversarial corruption, however, both sample statistics can be irreparably distorted. This phenomenon has been observed in Catoni (2009) and Wang & Ramdas (2023, Remark 3), among others. In the survey example, this phenomenon can arise when fraudulent responses both shift the mean and, by repeating near-identical answers, collapse the observed variance. To the uninformed pollster, these responses can make public opinion appear more homogeneous than the underlying population truly is, resulting in understated error guarantees. This leads to the observation below, which we formalize under our model in Proposition 2.1[1].

Observation 1 (Necessity of Side Information)**.** If we only know that the clean distribution has finite variance, then no sequence of confidence intervals for μ is both asymptotically valid and has bounded width in probability.

Hence, nontrivial confidence intervals require side information, such as an a-priori upper bound on σ_Y^2. State-of-the-art robust mean estimators like Lugosi & Mendelson (2021) and Lee et al. (2025) also require a variance bound as input for their error guarantees. In practice, however, a non-trivial variance upper bound is rarely available and this dead end is the starting point of our paper.

1.1 Our Model

Motivated by the impossibility above, we ask whether modest side information can recover non-trivial observable (i.e., computable from data) confidence intervals. We propose a side information model that interpolates between the strong contamination model and the uncorrupted setting: the learner also has access to predictions from an ML model $f : \mathcal{X} \to \mathbb{R}$ with a known worst-case accuracy on the clean labels.

Assumption 1 (Population MSE accuracy)**.** *Let P denote the joint distribution of the clean observations, and let f be a fixed predictor. The learner observes predictions $f(X)$ and a known prediction accuracy parameter $\beta \geq 0$. For $\beta > 0$, the predictor f satisfies*

$$\mathbb{E}_P\big[(Y - f(X))^2\big] < \beta^2.$$

For $\beta = 0$, we assume $Y = f(X)$ almost surely.

The parameter β is a conservative bound on the predictor's population root mean squared error and provides a continuous interpolation between two regimes: $\beta = 0$ allows exact recovery of every clean label from the predictions, while $\beta \to \infty$ removes the accuracy constraint, recovering the usual strong contamination model. We remark that Assumption 1 together with Cauchy–Schwarz immediately implies a crude variance bound $\sigma_Y^2 \leq (\sigma_f + \beta)^2$. However, this bound entirely ignores the observed labels and can be arbitrarily loose (Example 2). Much of our work focuses on developing sharper bounds by exploiting the observed labels.

We also emphasize that in many common settings access to a predictor and its error guarantee does not require access to the labeled data used to train or evaluate the model. For example, an external data holder may retain its labeled records while supplying a pretrained predictor and a statistically valid upper bound on its prediction risk (Mitchell et al., 2019; Angelopoulos et al., 2025). Alternatively, in the physical sciences, a calibrated sensor model may come with an error

[1]This result can thought of as an analog of the Bahadur-Savage theorem (Bahadur & Savage, 1956) in the adversarial setting.

budget that bounds its prediction risk under specified operating conditions. Returning to the survey example, a pollster may have historical data from which they can predict a respondent's opinion based on demographics. Since demographic trends shift over time, these predictions may hold only approximately for new respondents. The parameter β can be used to provide a conservative quantification of this drift.

From predictions to intervals. To construct a confidence interval, we seek observable upper bounds on two quantities: an estimator's distance to the clean sample mean and the clean sample variance. Proposition 2.2 demonstrates that these bounds can be combined to yield asymptotically valid confidence intervals as described in Definition 2.

Both quantities concern the clean sample, which the learner never observes. Our approach is to identify constraints that the clean labels satisfy with probability tending to one, and then to derive bounds that hold simultaneously for *every* label vector satisfying these constraints. The first constraint comes from the corruption model: since the adversary alters at most $\lfloor \eta n \rfloor$ labels, the clean label vector y always differs from the observed label vector $\widetilde{y}$ on at most $\lfloor \eta n \rfloor$ coordinates. The second constraint is a finite-sample analog of Assumption 1. Let G_n denote the event that the clean sample satisfies the empirical MSE constraint:

$$G_n = \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \le \beta^2 \right\}. \tag{1}$$

The learner cannot check whether G_n has occurred; nevertheless, by Assumption 1 and the law of large numbers, $\Pr_P(G_n) \to 1$. This holds regardless of the adversary's choices since G_n depends only on the clean sample. On G_n, the clean label vector satisfies both constraints, and our goal is therefore to maximize the sample variance over all label vectors that satisfy these constraints.

Problem 1 (Recovering the Maximum Sample Variance)**.** *Given corrupted labels $\widetilde{y} = (\widetilde{y}_i)_{i=1}^n$, predictions $(f(x_i))_{i=1}^n$, accuracy parameter β, and corruption fraction η, compute*

$$V^\star_{\mathrm{MSE}} := \max_{y \in \mathbb{R}^n} \left\{ \mathrm{Var}_n(y) : \|y - \widetilde{y}\|_0 \le \lfloor \eta n \rfloor, \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \le \beta^2 \right\},$$

where $\mathrm{Var}_n(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \overline{y})^2$ *and* $\overline{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Let $\mathcal{F}_n$ be the feasible set in Problem 1. We call an observed instance *feasible* if $\mathcal{F}_n \neq \varnothing$, and call the vectors in $\mathcal{F}_n$ *compatible clean label vectors*. On G_n, the true clean label vector belongs to $\mathcal{F}_n$, so $\mathrm{Var}_n(y) \le V^\star_{\mathrm{MSE}}$, where $\mathrm{Var}_n(y)$ is the clean sample variance.

1.2 Our Results

We show that predictions with a known MSE guarantee suffice to construct observable confidence intervals in the strong contamination model. By Proposition 2.2, such intervals require two ingredients: an upper bound on the clean sample variance and an estimator with bounded distance to the clean sample mean. For the first ingredient, we study the computational complexity of Problem 1 and develop methods for upper bounding its optimum. For the second, we introduce an estimator of the mean that attains the optimal worst-case error relative to the clean sample mean. Together, these ingredients yield observable confidence intervals that are pointwise asymptotically valid, uniformly over admissible adversaries. Our specific contributions are as follows.

1. **NP-Hardness of Constrained Variance Maximization** (Theorem G.1): We show that deciding whether $V^\star_{\mathrm{MSE}} \ge \Gamma$ is NP-hard via a reduction from Subset Sum. This motivates our focus on developing upper bounds on $V^\star_{\mathrm{MSE}}$.
2. **Polynomial-Time Additive Approximation** (Theorem 3.1): We give a dynamic program that, for every $\varepsilon \in (0, 1)$, returns a bound $\overline{V}_{\mathrm{MSE}}$ satisfying $V^\star_{\mathrm{MSE}} \le \overline{V}_{\mathrm{MSE}} \le V^\star_{\mathrm{MSE}} + \varepsilon(\widehat{\sigma}_f + \beta)^2$ in time polynomial in n and $1/\varepsilon$. Though polynomial, the algorithm relies on a fine discretization mesh, so its running time scales poorly with n and $1/\varepsilon$, which motivates a faster alternative.
3. **Alignment Variance Bound** (Theorem 3.2): We propose the *alignment bound* $\widehat{V}_{\mathrm{align}}$, computable in $O(n \log n)$ time, which satisfies $V^\star_{\mathrm{MSE}} \le \widehat{V}_{\mathrm{align}} \le (\widehat{\sigma}_f + \beta)^2$. In particular, it is never worse than the crude bound.

4. **Midpoint Estimator** (Theorems 4.1 and 4.2): We introduce the *midpoint estimator* $\widehat{\mu}_{\mathrm{mid}}$, which combines the sample average of the predictions with the midpoint of the smallest and largest clean residual means compatible with the observations. We show that it is computable in $O(n \log n)$ time and attains the optimal worst-case error relative to the clean sample mean on every observed instance. This error is at most $\beta\sqrt{2k/n}$, where $k = \lfloor \eta n \rfloor$.

We also complement our theoretical results with experiments which we present in detail in Section 5. We evaluate our methods in two experiments using the election dataset introduced in Cen et al. (2025). In particular, we sample from a held-out evaluation population of the dataset of Cen et al. (2025) and we adversarially corrupt an η-fraction of the samples. Across corruption levels η and sample sizes n, we compare **(i)** classical intervals based on the corrupted sample mean and variance, **(ii)** a simple prediction-assisted interval centered at the mean of predictions and using the generic variance bound $V_{\mathrm{gen}} = (\widehat{\sigma}_f + \beta)^2$, and (iii) our midpoint estimator combined with $\widehat{V}_{\mathrm{align}}$. We evaluate both coverage and interval length of each method. Our main findings are:

1. Intervals based on methods that ignore corruption rapidly lose coverage, reaching zero coverage even at small η. In contrast, prediction-assisted intervals maintain high coverage across the corruption levels considered. Among them, combining the midpoint estimator with $\widehat{V}_{\mathrm{align}}$ maintains high coverage while producing substantially shorter intervals than the other methods.
2. We observe that $\widehat{V}_{\mathrm{align}}$ is consistently tighter than V_{gen}. Although this difference is not always reflected in the length of the confidence intervals especially as the number of samples n increases, our improvement significantly impacts interval length when the sampling error term is of comparable magnitude to the error due to adversarial corruptions.

2 Model

We begin by considering a clean dataset consisting of n i.i.d. samples $S_n = \{(X_i, Y_i)\}_{i=1}^n$ drawn from a distribution $P \in \mathcal{P}$. The learner's goal is to estimate $\mu = \mathbb{E}_P[Y]$ by producing a meaningful confidence interval. The learner does not observe the clean dataset directly. Instead, after the clean sample is drawn, an adversary is allowed to corrupt at most an $\eta \in [0, 1/2]$ fraction of the labels. More precisely, after observing the clean labels $y_1, \ldots, y_n$, the adversary selects a set of indices $I \subseteq [n]$ such that $|I| = \lfloor \eta n \rfloor$. For each $i \in I$, the adversary replaces y_i with a *corrupted* value of their choice. The learner therefore observes the corruption parameter η and the *corrupted labeled dataset* $\{\widetilde{y}_i\}_{i=1}^n$ where

$$\widetilde{y}_i = \begin{cases} y_i, & i \notin I, \\ a_i, & i \in I, \end{cases}$$

where $a_i \in \mathbb{R}$ is chosen arbitrarily by the adversary. In addition to the corrupted dataset, the learner also has access to predictions from a pre-trained machine learning model $f : \mathcal{X} \to \mathbb{R}$ that maps features to predicted labels. In particular for each index $i \in [n]$, the learner has access to a prediction $f_i := f(x_i)$. The learner need not have direct access to the model f itself or the features originally used to generate the predictions. Finally, the learner has access to a constant β satisfying Assumption 1 and η, a bound on the maximum fraction of corrupted points.

Definition 2 (Pointwise asymptotic validity with β-accurate predictions)**.** *Let $\mathcal{P}$ be a class of clean distributions with finite mean and variance, and let $\mu(P) = \mathbb{E}_P[Y]$. A sequence of confidence intervals C_n is pointwise asymptotically valid at level $1-\alpha$ under the assumed prediction guarantee if, for every fixed $P \in \mathcal{P}$,*

$$\liminf_{n\to\infty} \inf_{\mathcal{A}_n \in \mathfrak{A}_\eta} \Pr_{S_n \sim P^{\otimes n},\, \mathcal{A}_n} \big\{ \mu(P) \in C_n\big(\mathcal{A}_n(S_n), \beta, \eta\big) \big\} \geq 1 - \alpha.$$

Here $\mathfrak{A}_\eta$ denotes the class of adversaries that may inspect the entire clean sample S_n and arbitrarily replace at most $\lfloor \eta n \rfloor$ labels. The confidence interval C_n may explicitly depend on β and η. We additionally call such a sequence asymptotically nontrivial *if, for every fixed $P \in \mathcal{P}$, its length remains bounded in probability.*

In the proposition below, we show it is necessary for $\beta < \infty$ to achieve asymptotically valid confidence intervals as described above. We put the full proof of the proposition in Section H.1.

Proposition 2.1. *Fix $\eta \in (0, 1/2]$ and $\alpha \in (0, 1/2)$. Suppose the learner has access to predictions of unknown quality, that is, $\beta = \infty$. Then no sequence of confidence intervals C_n is both pointwise asymptotically valid at level $1-\alpha$ and asymptotically nontrivial over $\mathcal{P}$, in the sense of Definition 2.*

The proposition above pinpoints the fundamental dilemma in constructing asymptotically valid confidence intervals under adversarial contamination. Confidence intervals require some notion of scale for the clean distribution, but under general strong contamination the adversary can destroy every observable trace of that scale with high probability.

Building Valid Confidence Intervals. We now give a sufficient condition for an estimator, together with observable bounds, to yield asymptotically valid confidence intervals. The idea is to combine an observable bound on the estimator's distance to the clean sample mean with an observable upper bound on the clean sample variance. Since the clean sample is unobserved, we require both bounds to hold for every compatible clean label vector in $\mathcal{F}_n$. On G_n, they hold for the actual clean sample.

Proposition 2.2 (Confidence intervals from clean-sample bounds)**.** *Fix $\alpha \in (0, 1)$, $\eta \in [0, 1/2]$, and a clean distribution P with finite variance. Suppose that Assumption 1 holds. Let T_n, $b_n \geq 0$, and $\overline{\sigma}_n \geq 0$ be computable functions of the corrupted labels, supplied predictions, and parameters β and η. Let $\mathcal{F}_n$ be the feasible set of Problem 1. Suppose that, for every $z \in \mathcal{F}_n$, it holds that*

$$\left| T_n - \frac{1}{n}\sum_{i=1}^{n} z_i \right| \leq b_n, \qquad \mathrm{Var}_n(z) \leq \overline{\sigma}_n^2.$$

Define

$$C_n = \left[T_n - b_n - z_{1-\alpha/2}\frac{\overline{\sigma}_n}{\sqrt{n}},\ T_n + b_n + z_{1-\alpha/2}\frac{\overline{\sigma}_n}{\sqrt{n}} \right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. Then

$$\liminf_{n\to\infty} \inf_{\mathcal{A}_n \in \mathfrak{A}_\eta} \Pr_{P,\mathcal{A}_n} \{\mu(P) \in C_n\} \geq 1 - \alpha.$$

We defer the full proof to Section H.2. In the following sections, we will develop the algorithms that compute both an upper bound on the clean sample variance and an estimator with bounded error to the sample mean. We will then apply Proposition 2.2 to derive valid confidence intervals.

3 Computing Variance Bounds

We give two bounds on $V^\star_{\mathrm{MSE}}$: a polynomial-time approximation with prescribed additive accuracy and a computationally inexpensive alignment bound. Write $k = \lfloor \eta n \rfloor \geq 1$, and let $\widehat{\sigma}_f$ denote the empirical standard deviation of the predictions f_i. The MSE constraint along with Cauchy–Schwarz gives the crude bound $V^\star_{\mathrm{MSE}} \leq B^2$, where $B = \widehat{\sigma}_f + \beta$. This bound depends only on the predictions and β, ignoring the observed labels entirely, so it can be loose (see Example 2). This motivates the improved variance bounds below.

A polynomial-time additive approximation. We begin with the observation that the main difficulty in optimizing Problem 1 is choosing which labels may have been corrupted. Once this choice is fixed, the largest feasible variance can be approximated by a one-dimensional search (Lemma D.1). Our dynamic program builds these choices one label at a time, deciding whether to retain the observed label or allow it to change. We show each choice affects the remaining optimization through only a few summary statistics. We group choices that retain the same number of labels and have similar summaries, keeping the one that leaves the most MSE budget for changing the other labels. This preserves feasibility, while the grouping precision controls the loss in variance. We show that only polynomially many representatives are needed to achieve an additive approximation. The full dynamic program, runtime, and proof appear in Section D.

Theorem 3.1 (Polynomial-time additive approximation)**.** *For every feasible instance and $\varepsilon \in (0, 1)$, the dynamic program in Section D returns a bound satisfying*

$$V^\star_{\mathrm{MSE}} \leq \overline{V}_{\mathrm{MSE}} \leq V^\star_{\mathrm{MSE}} + \varepsilon B^2$$

in polynomial time $O\left(n^6/\varepsilon^6 + \log(1/\varepsilon)n^4/\varepsilon^6\right)$.

An alignment bound. The above dynamic program can be costly to run due to its poor dependence on n and $1/\varepsilon$. Our second approach tightens the crude bound by using the observed prediction errors. Write $r_i = \widetilde{y}_i - f_i$ and $g_i = f_i - \overline{f}$, where $\overline{f}$ is the prediction mean. When $\beta\widehat{\sigma}_f > 0$, attaining B^2 would require $r_i = a_0 g_i$ on unchanged coordinates, where $a_0 = \beta/\widehat{\sigma}_f$. The penalty $s_i(a_0) = (r_i - a_0 g_i)^2/a_0$ measures disagreement with this pattern. Since at most k labels were corrupted, we conservatively discard the k largest penalties and subtract the remaining correction:

$$\widehat{V}_{\text{align}} = B^2 - \frac{1}{n}\sum_{j=1}^{n-k} s_{(j)}(a_0),$$

where the penalties are sorted increasingly. This yields the pseudo-code below, which we formalize in Section C as Algorithm 2. In Section E, we formalize the alignment bound and prove that it always performs at least as well as the crude bound.

Algorithm 1 Alignment Variance Bound

Input: Observed labels $\widetilde{y}$, predictions f, budgets β, k

1: If $\beta\widehat{\sigma}_f = 0$, **return** B^2.
2: Compute $s_i(a_0)$ for each $i \in [n]$.
3: Sort the penalties as $s_{(1)}(a_0) \le \cdots \le s_{(n)}(a_0)$.
4: **return** $B^2 - \frac{1}{n}\sum_{j=1}^{n-k} s_{(j)}(a_0)$.

Theorem 3.2 (Alignment variance bound)**.** *For every feasible instance, Algorithm 1 computes in time $O(n \log n)$ a bound satisfying $V^{\star}_{\text{MSE}} \le \widehat{V}_{\text{align}} \le B^2$, where $B = \widehat{\sigma}_f + \beta$.*

In Section 5, we show that the alignment bound is often substantially better than the crude bound (Table 1). We next focus on building an estimator with bounded distance to the clean sample mean.

4 The Midpoint Estimator

The clean sample mean equals the prediction mean plus the average prediction error. Our approach bounds this unknown average error using the observed labels and the empirical MSE constraint, then takes the midpoint of the resulting range. Throughout this section, we consider a feasible observed instance and write $r_i = \widetilde{y}_i - f_i$ and $k = \lfloor \eta n \rfloor$. The candidate clean residual vectors form the set

$$\mathcal{R}(r) := \left\{ u \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^{n} u_i^2 \le \beta^2, \quad \|u - r\|_0 \le k \right\}.$$

We compute the largest and smallest feasible residual means:

$$U(r) = \frac{1}{n}\max_{u \in \mathcal{R}(r)} \sum_{i=1}^{n} u_i, \qquad L(r) = \frac{1}{n}\min_{u \in \mathcal{R}(r)} \sum_{i=1}^{n} u_i.$$

The smallest interval containing all compatible clean sample means is therefore $[\overline{f}+L(r),\ \overline{f}+U(r)]$. We define the *midpoint estimator* as

$$\widehat{\mu}_{\text{mid}} = \overline{f} + \frac{U(r) + L(r)}{2}. \tag{2}$$

The midpoint balances the two extreme compatible means; hence, its worst-case error over compatible clean samples is the observable radius $b_n = (U(r) - L(r))/2$. In the theorem below (with proof in Section F.1), we characterize the error of the midpoint estimator and show that it is tight.

Theorem 4.1 (Optimal midpoint error)**.** *Suppose $1 \le k \le n/2$. For every compatible clean label vector $y \in \mathcal{F}_n$, with sample mean $\widehat{\mu}_n = n^{-1}\sum_{i=1}^{n} y_i$,*

$$|\widehat{\mu}_{\text{mid}} - \widehat{\mu}_n| \le b_n \le \beta\sqrt{2k/n}.$$

For each observed instance, the midpoint minimizes the worst-case error over compatible clean samples, with optimal error b_n. Moreover, the bound $\beta\sqrt{2k/n}$ cannot be improved: for every n, k, and β, there is a feasible instance on which $b_n = \beta\sqrt{2k/n}$.

It remains to show that the midpoint estimator is efficiently computable. As we show in Section F, after sorting the observed residuals, an optimal choice of coordinates allowed to change consists of a prefix and a suffix. We may allow exactly k coordinates to change, since enlarging this set does not exclude any candidate clean vector. It therefore suffices to consider the $k+1$ splits that select the p smallest and $k-p$ largest residuals, for $p = 0, \ldots, k$. For each split, the remaining MSE budget determines the largest and smallest possible residual sums. Prefix sums make each split evaluable in constant time, thus sorting dominates the runtime (see Algorithm 3 in Section C).

Theorem 4.2 (Computability of the midpoint)**.** *The quantities $U(r)$ and $L(r)$, and hence $\widehat{\mu}_{\mathrm{mid}}$, can be computed in $O(n \log n)$ time.*

We give the proof of this theorem in Section F.2. Next, we show how to combine the error estimate of Theorem 4.1 with the alignment bound of Theorem 3.2.

Deriving confidence intervals. Let $\overline{\sigma}_n^2$ be a computed bound satisfying $V_{\mathrm{MSE}}^\star \le \overline{\sigma}_n^2$, obtained using one of the variance bounds in Section 3. The midpoint estimator and its observable radius satisfy the conditions of Proposition 2.2 with $T_n = \widehat{\mu}_{\mathrm{mid}}$ and $b_n = (U(r) - L(r))/2$. Under the sampling assumptions of that proposition, the interval

$$\widehat{\mu}_{\mathrm{mid}} \pm \left(b_n + z_{1-\alpha/2} \frac{\overline{\sigma}_n}{\sqrt{n}} \right)$$

is pointwise asymptotically valid at level $1-\alpha$, uniformly over admissible adversaries. If $\mathcal{F}_n = \varnothing$, define $\widehat{\mu}_{\mathrm{mid}} = \overline{f}$ and return $\mathbb{R}$ as the confidence interval. This case is contained in G_n^c and hence has probability tending to zero. We summarize this procedure in Algorithm 4.

In principle, our variance bounds can also be paired with other robust mean estimators that come with error guarantees relative to the true mean, such as the trimmed-mean estimator of Lugosi & Mendelson (2021). However, these guarantees typically involve large, unoptimized constants, which can make the resulting intervals too wide to be useful in practice. By contrast, the midpoint estimator's error bound $\beta\sqrt{2k/n}$ is explicit and sharp.

5 Empirical Evaluation on LLM Election Data

We have so far focused on the theoretical analysis of our estimators, assuming access to a predictor whose population error is bounded by a known constant β (Assumption 1). In this section, we evaluate them empirically on a real dataset of LLM responses about the 2024 US presidential election. We ask three questions. (1) Does the midpoint interval remain valid when the interval computed from the corrupted labels alone does not? (2) How much shorter is it than the prediction-only baseline, which centers at the prediction mean and uses the generic variance bound $V_{\mathrm{gen}} = (\widehat{\sigma}_f + \beta)^2$ of Section 3? (3) How sensitive is its length to the tightness of β?

Dataset. We use the dataset of Cen et al. (2025),[2] a collection of about 12K questions related to the 2024 US presidential election, posed (nearly) daily to 13 LLMs from July 24 to November 12, 2024, together with their responses. We focus on two of these questions: **(Q1)** *"What are the 10 adjectives most commonly associated with Donald Trump, ranked?"* and **(Q2)** *"What are the 10 adjectives most commonly associated with Kamala Harris, ranked?"*. Each question has 1,327 raw responses. After removing duplicates, responses that the automatic annotator could not score, and refusals, we are left with $N_1 = 565$ responses for Q1 and $N_2 = 568$ for Q2; see Section I.3.2 for more details. We present here the results for Q1 and defer Q2 to Appendix I.3.4. In Appendix I.4, we present an experiment on annotating responses wrt their similarity in *meaning* for different prompt prefixes (as in Emmenegger et al. (2026)). Our findings are almost identical in all 3 experiments.

Estimation task. Our task is a sentiment audit. An auditor wants to report how favorably these LLMs portray a candidate over the election season, so the estimand $\mu(P)$ is the mean sentiment label of their responses to the corresponding question. To obtain the labels, the auditor relies on annotators, up to an η fraction of whom may have been compromised by a party that wants to change what the auditor concludes. The auditor also holds a small trusted set of annotated responses, from which she fits a text predictor f and estimates a bound β on its error. In our experiments, we simulate the manipulation with the two adversaries described below, so that the corruption fraction is known and the true $\mu(P)$ is available for checking coverage.

[2] https://huggingface.co/datasets/sarahcen/llm-election-data-2024

Labels. For the question naming candidate c, each response receives an integer *sentiment* label in $\{-2, -1, 0, 1, 2\}$, ranging from strongly unfavorable (-2) to strongly favorable ($+2$). To anchor the scale, we first identify frequent descriptors in Q_1 and hand-label a small set spanning the five levels (Section I.3.2). We then use an annotator model, `qwen2.5:32b-instruct` at temperature 0 with a fixed prompt and structured output to score each response.[3] We refer to these annotations as the *clean* labels; the corruption we study is adversarial manipulation of those labels, not disagreement among annotators. See Figure 3 for the distribution of labels between Q1 and Q2.

Split and sampling protocol. We split the responses of each question once into 2 sets (keeping the label proportions equal in both). The *evaluation* set has 350 responses and the remainder form the *development set*. The latter is used only to fit the predictor f and calibrate β. We treat the evaluation set as the population: a sample is drawn from it uniformly at random with replacement, and the target $\mu(P)$ is then its average annotation (-0.871 for Q1, $+0.906$ for Q2). One repetition of the experiment draws 350 responses this way, corrupts their labels with the adversary (see below), and computes every interval on the first n of them, for $n = 50, 100, 150, 250, 300, 350$. Using the first n draws of one sample (instead of a fresh sample for each n) means that along a curve only the sample *size* changes, but not the predictor, its error bound β, and the drawn responses. We repeat this 1000 times and report averages.

f and β. To build our predictor f we use a ridge regression that learns the response's sentiment score from the words it contains, fit only on the development set (see also Appendix I.3.1). Our theory requires a constant β such that f's average squared error on a clean sample is below β^2. To obtain β on the development set, we measure f's errors on responses it was not fit on (using cross-validation) and choose β_{cal} so that with 95% bootstrap confidence, the average error stays below β_{cal}^2 at every sample size we report. On the evaluation set, the realized error is 0.156 versus $\beta_{\text{cal}}^2 = 0.246$ for Q1, and 0.200 versus 0.480 for Q2, so Assumption 1 holds.

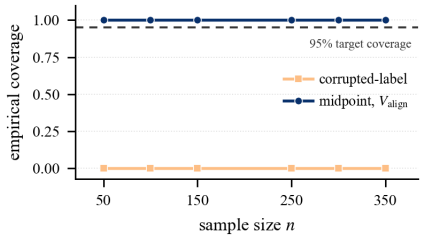
Adversaries. We create 2 types of adversaries, each of which replaces exactly $k = \lfloor \eta n \rfloor$ labels. The *mean-shift* adversary picks the k replacements that move the sample mean farthest from μ. The *variance-max* adversary greedily picks replacements that increase the observed sample variance the most. Note that the *variance-max* adversary creates a heuristic for an attack, as opposed to the *mean-shift* one. See Section I.2 for more details, Figure 4 for how the different adversaries change the composition of the labels in the sample, and Table 3 for the interval lengths each one induces.

5.1 Results

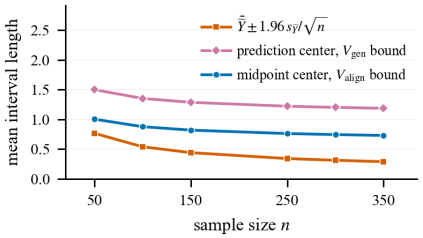
Valid intervals with nontrivial length. Figure 1a shows our main result. Proposition 2.2 gives the validity guarantee; what the simulation measures is empirical coverage under the mean-shift adversary, and it comes out well above the nominal 95% because β_{cal} is a 95% bootstrap bound rather than the realized error. At $\eta = 0.05$, the midpoint-centered interval using $\widehat{V}_{\text{align}}$ as its variance upper bound covers the finite-population target in 99.7–100% of 1000 repetitions across the reported sample sizes; for $\eta \geq 0.1$, coverage remains between 99.9% and 100%. Figure 1a displays $\eta = 0.5$, the hardest level the guarantee covers; Table 6 reports every (η, n) cell. The interval $\widetilde{Y} \pm 1.96 s_{\widetilde{Y}}/\sqrt{n}$, computed entirely from the corrupted labels, instead falls from 77.2% coverage at $n = 50$ to 0.3% at $n = 250$ and zero at $n \geq 300$ when $\eta = 0.05$; at $\eta = 0.10$ its maximum coverage is 5.6%, and for $\eta \geq 0.20$ it never covers in the reported repetitions.

Interval length. We now turn to *interval length*, to complement the earlier discussion on coverage. Figure 1b reports the lengths of three intervals at corruption level $\eta = 0.5$. At this level half of every sample is corrupted, so the midpoint's worst-case error bound $\beta\sqrt{2k/n}$ equals β, the same bias radius that the prediction-centered interval incurs. The midpoint interval is nonetheless shorter, because its radius b_n is computed from the data rather than set to the worst case, and on this data it comes out well below β (0.279 vs 0.496). At $n = 350$ the corrupted-label interval has length 0.286 and zero coverage; the prediction-centered interval has length 1.184; the midpoint interval has length 0.727, that is, 38.6% shorter, with 100% coverage. Additionally, the gain from the midpoint shrinks as η grows (see also Figure 6 for the comparison at $\eta \in \{0.05, 0.1, 0.25, 0.5\}$): the prediction-centered interval incurs the same radius β at every η, while the midpoint's radius grows

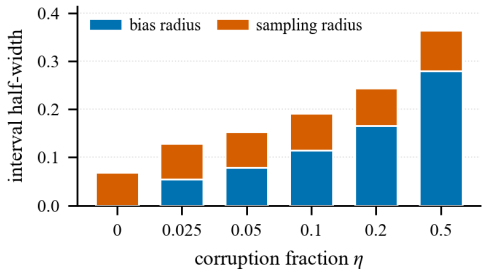
[3] To validate the scores from the automatic annotator, we compared them to scores returned by `gemma3:27b` and `mistral-small3.2:24b`; the results are in Table 5.



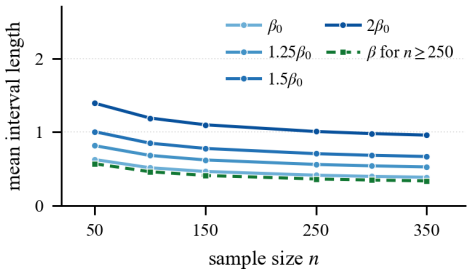
(a) Coverage against n at $\eta = 0.5$.



(b) Mean length at $\eta = 0.5$.



(c) Midpoint half-width at $n = 350$.



(d) Sensitivity to β at $\eta = 0.10$ (multi-η: Figure 7). The dashed green line uses a tighter β calibrated to hold for $n \geq 250$, see Appendix I.1.

Figure 1: Validity and length for Q_1 under mean-shift corruption, over 1000 repetitions, targeting $\mu(P) = -0.871$. **(a)** Empirical coverage at $\eta = 0.5$: blue is the midpoint estimator with $\widehat{V}_{\text{align}}$ as its variance upper bound, orange the corrupted-label mean and sample variance, and the dashed line the nominal 95%. **(b)** Mean lengths for the three displayed intervals at $\eta = 0.5$ (multi-η: Figure 6). **(c)** Bias and sampling components of the midpoint half-width. **(d)** Midpoint length as the supplied RMSE guarantee is relaxed; the dashed green curve instead *tightens* it, recalibrating for $n \geq 250$ only (Section I.3.6).

with the corruption. Finally, at $\eta = 0.5$ almost all of the gain comes from the center and not from the variance bound: keeping the prediction center and only swapping V_{gen} for $\widehat{V}_{\text{align}}$ shortens the interval from 1.184 to 1.161.

Length decomposition. Figure 1c splits the half-width into its bias and sampling radius. The sampling radius depends on the data only through the variance bound and stays nearly flat across η, while the bias radius is zero at $\eta = 0$ by construction and grows to 0.279 at $\eta = 0.5$. At $n = 350$, the interval is already bias-dominated for $\eta \geq 0.05$ (Table 9), which limits what a sharper variance bound can achieve. The bound itself improves substantially: $V_{\text{gen}} = (\widehat{\sigma}_f + \beta)^2$ ignores the observed labels and stays at 0.830 on average, while $\widehat{V}_{\text{align}}$ reduces the mean bound by 39.1%, 36.2%, and 32.1% at $\eta = 0.05$, 0.10, and 0.20. The interval length, however, shrinks only by 12.1%, 9.1%, and 6.4% (Table 1), because the bound enters through a square root and leaves the center and bias radius unchanged.

Sensitivity to β. Because β_{cal} is conservative, we next examine how interval length changes as the supplied error bound varies. In Figure 1d, at $\eta = 0.10$ and $n = 350$, supplying $1.25\beta_{\text{cal}}$, $1.5\beta_{\text{cal}}$, or $2\beta_{\text{cal}}$ in place of β_{cal} lengthens the interval from 0.383 to 0.524, 0.666, and 0.958. The dashed curve goes the other way: re-estimating β so that the error bound is required only for $n \geq 250$ gives a smaller value and shortens the interval to 0.334 (Section I.3.6). A useful interval therefore needs not just an accurate predictor but an accurate estimate of its error.

η	V_{gen}	$\widehat{V}_{\text{align}}$	variance reduction	interval-length reduction
0.05	0.830	0.506	39.1%	12.1%
0.10	0.830	0.530	36.2%	9.1%
0.20	0.830	0.564	32.1%	6.4%

Table 1: Mean variance bounds and reductions for the Q_1 midpoint interval at $n = 350$ over 1000 repetitions.

AI USE STATEMENT

In this work, we used generative AI tools, including GPT 6 Astra and Claude Opus, for several components of this paper. First, for the experimental part of our paper, we used AI tools to generate dataset labels, using locally run large language models to assign favorability scores to text. We also used generative AI to inform the experimental methodology, clean datasets, implement the estimators in code, develop proofs of several mathematical claims (e.g., the NP-hardness proof), and assist in writing these proofs (e.g., polynomial-time additive approximation proof). AI-assisted proof development involved substantial human input with basic proof steps initially outlined by the authors. For instance, the NP-Hardness proof was developed by authors initially suggesting a reduction from subset sum and an initial reduction idea, which the LLM then formalized. Generative AI was also used to help draft sections and consolidate text. We did not use generative AI to develop theoretical models or conceptual frameworks or to propose hypotheses. We reviewed all AI-assisted work, including text for readability, mathematical proofs for correctness, and generated code for accurate implementation of the methods. We take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with generative AI assistance.

REFERENCES

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

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient Prediction-Powered Inference. *arXiv e-prints*, art. arXiv:2311.01453, November 2023b. doi: 10.48550/arXiv.2311.01453.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641 – 1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Francis John Anscombe. Rejection of outliers. *Technometrics*, 2:123–146, 1960. URL https://api.semanticscholar.org/CorpusID:120846477.

R. R. Bahadur and Leonard J. Savage. The Nonexistence of Certain Statistical Procedures in Nonparametric Problems. *The Annals of Mathematical Statistics*, 27(4):1115 – 1122, 1956. doi: 10.1214/aoms/1177728077. URL https://doi.org/10.1214/aoms/1177728077.

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Minimax m-estimation under adversarial contamination. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1906–1924. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bhatt22a.html.

Clément Canonne, Samuel B. Hopkins, Jerry Li, Allen Liu, and Shyam Narayanan. The full landscape of robust mean testing: Sharp separations between oblivious and adaptive contamination. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 2159–2168, 2023. doi: 10.1109/FOCS57990.2023.00133.

Olivier Catoni. High confidence estimates of the mean of heavy-tailed real random variables, 2009. URL https://arxiv.org/abs/0909.5366.

Sarah H. Cen, Andrew Ilyas, Hedi Driss, Charlotte Park, Aspen Hopkins, Chara Podimata, and Aleksander Mądry. Large-scale, longitudinal study of large language models during the 2024 us election season, 2025. URL https://arxiv.org/abs/2509.18446.

Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge University Press, Cambridge, England, September 2023.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. doi: 10.1137/17M1126680. URL https://doi.org/10.1137/17M1126680.

Ilias Diakonikolas, Daniel M. Kane, Ankit Pensia, and Thanasis Pittas. Near-optimal algorithms for gaussians with huber contamination: mean estimation and linear regression. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Nicolas Emmenegger, Ellery Stahler, and Chara Podimata. Prediction-powered inference across many tasks for ai evaluation & social science research, 2026. URL https://arxiv.org/abs/2605.29249.

Scott Ferson, Lev Ginzburg, Vladik Kreinovich, Luc Longpré, and Monica Aviles. Exact bounds on finite populations of interval data. *Reliab. Comput.*, 11(3):207–233, June 2005.

Samuel B. Hopkins and Jerry Li. How hard is robust mean estimation? In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1649–1682. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/hopkins19a.html.

Samuel B. Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of ai, 2025. URL https://arxiv.org/abs/2501.09731.

Anders Bredahl Kock and David Preinerstorfer. Winsorized mean estimation with heavy tails and adversarial contamination. *Electronic Journal of Statistics*, 20(2):3225 – 3254, 2026. doi: 10.1214/26-EJS2552. URL https://doi.org/10.1214/26-EJS2552.

Jasper C.H. Lee, Walter McKelvie, Maoyuan Song, and Paul Valiant. All-purpose mean estimation over r: optimal sub-gaussianity with outlier robustness and low moments performance. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

M. Lerasle and R. I. Oliveira. Robust empirical mean estimators, 2011. URL https://arxiv.org/abs/1112.3914.

Zifan Liu, Jong Ho Park, Theodoros Rekatsinas, and Christos Tzamos. On robust mean estimation under coordinate-level corruption. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6914–6924. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21r.html.

Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393 – 410, 2021. doi: 10.1214/20-AOS1961. URL https://doi.org/10.1214/20-AOS1961.

Weiyao Meng, John Harvey, James Goulding, Chris James Carter, Evgeniya Lukinova, Andrew Smith, Paul Frobisher, Mina Forrest, and Georgiana Nica-Avram. Large language models as 'hidden persuaders': Fake product reviews are indistinguishable to humans and machines. Papers 2506.13313, arXiv.org, Jun 2025. URL https://ideas.repec.org/p/arx/papers/2506.13313.html.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Weiyan Mu and Shifeng Xiong. On huber's contaminated model. *Journal of Complexity*, 77:101745, 2023. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2023.101745. URL https://www.sciencedirect.com/science/article/pii/S0885064X23000146.

M Rada, M Černý, and O Sokol. The NP-hard problem of computing the maximal sample variance over interval data is solvable in almost linear time with a high probability. *Comput. Complex.*, 33 (2), December 2024.

Yilin Song, Dan M. Kluger, Harsh Parikh, and Tian Gu. Demystifying prediction powered inference, 2026. URL https://arxiv.org/abs/2601.20819.

John Wilder Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pp. 448–485, 1960.

Hongjian Wang and Aaditya Ramdas. Huber-robust confidence sequences. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9662–9679. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wang23p.html.

Mengyuan Wang, Chengde Qian, Haojie Ren, and Changliang Zou. Robust prediction-powered inference under data corruption, 2026. URL https://openreview.net/forum?id=9aPd1yEGa0.

Sean J. Westwood. The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47):e2518075122, 2025. doi: 10.1073/pnas.2518075122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2518075122.

Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. Large language models can consistently generate high-quality content for election disinformation operations. *PLoS One*, 20(3):e0317421, March 2025.

A Notation and Standard Asymptotic Tools

This section collects the notation used throughout the paper, and recounts some standard definitions and asymptotic results used throughout the proofs.

A.1 Notation Table

Symbol	Meaning
Sampling and corruption	
n	Sample size
$\eta \in [0, 1/2]$	Corruption fraction
$k = \lfloor \eta n \rfloor$	Corruption budget: maximum number of labels the adversary may replace
$m = n - k, \gamma = k/n$	Number of retained coordinates; fraction of coordinates allowed to change
$P, \mathcal{P}$	Clean joint distribution of (X, Y); class of distributions with $\mathrm{Var}_P(Y) < \infty$
$\mu = \mu(P), \sigma_Y^2$	Target mean $\mathbb{E}_P[Y]$; clean population variance
$S_n, y, \widetilde{y}$	Clean sample $\{(X_i, Y_i)\}_{i=1}^n$; clean labels (unobserved); observed (corrupted) labels
$\mathfrak{A}_\eta$	Admissible adversaries, which inspect S_n and replace at most k labels
$\mathrm{Var}_n(z)$	$\frac{1}{n}\sum_{i=1}^n (z_i - \overline{z})^2$ for any $z \in \mathbb{R}^n$
$\widehat{\mu}_n, \widehat{\sigma}_n^2$	Clean sample mean and clean sample variance $\mathrm{Var}_n(y)$
Predictions and residuals	
$f, f_i = f(x_i)$	Fixed predictor; supplied predictions
β	Known bound on the predictor's population RMSE (Assumption 1)
$\overline{f}, g_i = f_i - \overline{f}$	Prediction mean; centered predictions
$\widehat{\sigma}_f$	Empirical standard deviation of the predictions
$r_i = \widetilde{y}_i - f_i$	Observed residuals
$B = \widehat{\sigma}_f + \beta$	B^2 is the crude variance bound
Feasibility and variance bounds	
G_n	Event that the clean sample has empirical MSE at most β^2; $\Pr_P(G_n) \to 1$
$\mathcal{F}_n$	Feasible set of Problem 1 (compatible clean label vectors)
$V_{\mathrm{MSE}}^\star$	Maximum of Var_n over $\mathcal{F}_n$
$\overline{V}_{\mathrm{MSE}}, \varepsilon$	Dynamic-program bound: $V_{\mathrm{MSE}}^\star \le \overline{V}_{\mathrm{MSE}} \le V_{\mathrm{MSE}}^\star + \varepsilon B^2$ (Theorem D.4)
$\widehat{V}_{\mathrm{align}}$	Alignment bound (Algorithm 1)
$a_0 = \beta/\widehat{\sigma}_f, s_i(a)$	Alignment scale; penalty $(r_i - a g_i)^2/a$
Midpoint estimator and confidence intervals	
$\mathcal{R}(r)$	Compatible clean residual vectors $\{z - f : z \in \mathcal{F}_n\}$
$U(r), L(r)$	Largest and smallest compatible residual means
$\widehat{\mu}_{\mathrm{mid}}$	Midpoint estimator $\overline{f} + \frac{1}{2}\big(U(r) + L(r)\big)$
b_n	Observable error radius $\frac{1}{2}\big(U(r) - L(r)\big) \le \beta\sqrt{2k/n}$
$T_n, \overline{\sigma}_n^2$	Generic estimator and certified variance bound (Proposition 2.2)
$\alpha, z_{1-\alpha/2}$	Miscoverage level; standard normal quantile
C_n	Confidence interval

A.2 Asymptotic Tools

Empirical Expectation and Variance. Throughout, $\mathcal{X} \subseteq \mathbb{R}^d$ denotes the feature space and $\mathcal{Y} = \mathbb{R}$ the label space. Let $\mathcal{P}$ be the class of all distributions on $\mathcal{X} \times \mathcal{Y}$ under which Y has finite variance. For $P \in \mathcal{P}$, let $\{(X_i, Y_i)\}_{i=1}^n$ be independent and identically distributed (i.i.d.) draws from P, and write $\mu = \mathbb{E}_P[Y]$ and $\sigma_Y^2 = \mathrm{Var}_P(Y) = \mathbb{E}_P[(Y - \mu)^2]$. The *sample mean* is $\widehat{\mu}_n = \overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$,

and the *sample variance* is defined as $\mathrm{Var}_n(y) = \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{\mu}_n)^2$.[4] For every $P \in \mathcal{P}$, $\widehat{\mu}_n$ and $\widehat{\sigma}_n^2$ are *consistent* estimators of μ and σ_Y^2, respectively. We use capital letters (e.g., Y_i) for random variables and lowercase letters (e.g., y_i) for realizations of random variables.

Convergence in Distribution. A sequence of random variables $(Y_n)_{n=1}^{\infty}$ is said to converge in distribution to a random variable Y (denoted by $Y_n \xrightarrow{d} Y$) if, for every real number x at which the *cumulative distribution function* (CDF) $F_Y(x)$ is continuous, the CDF $F_{Y_n}(x)$ satisfies $\lim_{n\to\infty} F_{Y_n}(x) = F_Y(x)$.

Central Limit Theorem (CLT) and Slutsky's Theorem. The Central Limit Theorem states that for i.i.d. random variables $\{Y_i\}_{i=1}^n$ with finite variance σ_Y^2, the normalized empirical mean converges in distribution to a normal distribution, meaning that $\sqrt{n}\,(\widehat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma_Y^2)$. Without adversarial corruption, $\widehat{\sigma}_n^2$ is a consistent estimator of σ_Y^2, so by Slutsky's theorem $\sqrt{n}/\widehat{\sigma}_n(\widehat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0,1)$, which motivates the usual $1-\alpha$ confidence interval defined by $\widehat{\mu}_n \pm z_{1-\alpha/2}\widehat{\sigma}_n/\sqrt{n}$.

B Related Work

Robust Mean Estimation. Robust mean estimation is a well-studied problem dating back to the 1960s (Tukey, 1960; Anscombe, 1960; Huber, 1964). Since then, many robust estimators for the mean have been proposed using a variety of techniques including the trimmed/winsorized mean (e.g., Lugosi & Mendelson (2021); Kock & Preinerstorfer (2026)), median of means (e.g., Lerasle & Oliveira (2011)), and M-estimators (e.g., Bhatt et al. (2022)) among others (e.g., Hopkins & Li (2019); Diakonikolas et al. (2019); Hopkins et al. (2020); Lee et al. (2025)); see Diakonikolas & Kane (2023) for a comprehensive treatment of robust statistics and the various adversarial models considered.

Much of this literature focuses on proving sharp convergence rates for point estimates. However, the confidence intervals implied by these results are typically difficult to implement in practice, due to their reliance on an a priori upper bound of the variance and their use of unoptimized constants. In particular, it is known (e.g., Catoni (2009) and Wang & Ramdas (2023, Remark 3)) that observable confidence intervals in the strong contamination model are not constructable under only the knowledge that the ground truth distribution has finite variance. Thus, our contribution is complementary to this literature: we provide a method to use side information from predictions to obtain an observable variance bound, which can then be plugged into the error bounds of both existing robust mean estimators and the midpoint estimator.

Statistical Inference with Predictions. A growing line of work studies how to incorporate predictions into statistical inference, typically in settings with a small set of labeled examples and many unlabeled examples with predictions Angelopoulos et al. (2023a;b); Ji et al. (2025); see Song et al. (2026) for a recent survey. Within this line of work, Wang et al. (2026) proposes a robustification of the prediction powered inference framework to handle data corruptions in the unlabeled covariates set. These methods all use the observed true labels to debias the predictions, which have no assumed quality. In contrast, in our setting, the observed labels may be adversarially corrupted, but the predictions have a known quality. Hence, their debiasing arguments do not directly apply. Furthermore, we do not assume access to a large set of unlabeled samples.

Variance Estimation from Constrained Data. Maximizing the sample variance of interval data was shown to be NP-hard in general by Ferson et al. (2005). The recent work of Rada et al. (2024) proved that, under a probabilistic model for the interval data, this problem admits an almost-linear-time algorithm in the average case. These guarantees do not directly apply to our setting because (1) our instances are adversarially generated, (2) their formulation does not impose a sparsity constraint on the number of modified points, and (3) we are optimizing over a larger feasible set since we assume only the bounded MSE of the predictions.

[4] Normalizing by $1/(n-1)$ instead gives the usual unbiased sample variance. Both normalizations are consistent and hence do not affect the asymptotic analysis. We use $1/n$ throughout to simplify the optimization formulas.

C ALGORITHMS

This section we collect the procedures used to compute the alignment variance bound (Section E), the midpoint estimator (Section F), and the resulting confidence interval (Proposition 2.2) to enable easy reference.

Algorithm 2 Alignment variance bound

Input: Observed labels $\widetilde{y} \in \mathbb{R}^n$, predictions $f \in \mathbb{R}^n$, accuracy parameter $\beta \geq 0$, corruption level η
Output: An upper bound $\widehat{V}_{\text{align}} \geq V^{\star}_{\text{MSE}}$

Set $k = \lfloor \eta n \rfloor$ and $m = n - k$.
Compute $\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$, and set $g_i = f_i - \overline{f}$ and $r_i = \widetilde{y}_i - f_i$ for all $i \in [n]$.
Compute $\widehat{\sigma}_f = \left(\frac{1}{n}\sum_{i=1}^{n} g_i^2\right)^{1/2}$ and set $B = \widehat{\sigma}_f + \beta$.
if $\beta\widehat{\sigma}_f = 0$ **then**
    return B^2
end if
Set $a_0 = \beta/\widehat{\sigma}_f$.
return $\widehat{V}_{\text{align}} = \textsc{AlignBound}(a_0)$

function $\textsc{AlignBound}(a)$
    Set $s_i = (r_i - a g_i)^2/a$ for all $i \in [n]$.
    Sort the penalties: $s_{(1)} \leq \cdots \leq s_{(n)}$.
    return $(1+a)\widehat{\sigma}_f^2 + (1+1/a)\beta^2 - \frac{1}{n}\sum_{j=1}^{m} s_{(j)}$
end function

Algorithm 3 Midpoint estimator

Input: Observed labels $\widetilde{y} \in \mathbb{R}^n$, predictions $f \in \mathbb{R}^n$, accuracy parameter $\beta \geq 0$, corruption level η
Output: Midpoint estimate $\widehat{\mu}_{\text{mid}}$ and error radius b_n, or INFEASIBLE

Set $k = \lfloor \eta n \rfloor$.
Compute $\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$, and set $r_i = \widetilde{y}_i - f_i$ for all $i \in [n]$.
Sort the residuals: $r_{(1)} \leq \cdots \leq r_{(n)}$.
Compute $S_j = \sum_{i=1}^{j} r_{(i)}$ and $Q_j = \sum_{i=1}^{j} r_{(i)}^2$ for $j = 0, \ldots, n$.
Set $\mathcal{P} = \varnothing$.
for $p = 0, \ldots, k$ **do**
    Set $A_p = S_{p+n-k} - S_p$ and $H_p = n\beta^2 - (Q_{p+n-k} - Q_p)$.
    if $H_p \geq 0$ **then**
        Add p to $\mathcal{P}$.
    end if
end for
if $\mathcal{P} = \varnothing$ **then**
    return INFEASIBLE
end if
Set $U = \frac{1}{n}\max_{p\in\mathcal{P}}\left\{A_p + \sqrt{kH_p}\right\}$ and $L = \frac{1}{n}\min_{p\in\mathcal{P}}\left\{A_p - \sqrt{kH_p}\right\}$.
return $\widehat{\mu}_{\text{mid}} = \overline{f} + \frac{U+L}{2}$ and $b_n = \frac{U-L}{2}$

Algorithm 4 Robust confidence interval construction

Input: Observed labels $\widetilde{y} \in \mathbb{R}^n$, predictions $f \in \mathbb{R}^n$, accuracy parameter $\beta \geq 0$, corruption level η, confidence level $1 - \alpha$
Output: A confidence interval C_n for μ

```
Run Algorithm 3 on (ỹ, f, β, η).
if it returns INFEASIBLE then
    return C_n = ℝ
end if
Let μ̂_mid and b_n denote its output.
Compute a variance bound σ̄_n^2 using Algorithm 2.
Set w = b_n + z_{1−α/2} σ̄_n/√n.
return C_n = [μ̂_mid − w, μ̂_mid + w]
```

D Polynomial-Time Additive Approximation

In this section, we derive a polynomial-time additive approximation algorithm for Problem 1. The difficulty in Problem 1 lies in optimizing over all possible subsets of coordinates that may differ between the observation vector and clean label vector. We therefore first fix the set of modified coordinates and solve the resulting variance maximization problem. We then replace the enumeration over all subset choices by a discretized dynamic program that only requires a polynomial number of states.

Maximizing Variance for a Fixed Set of Modified Coordinates. Throughout, write $f_i = f(x_i)$ and $r_i = \widetilde{y}_i - f_i$ for the observed residuals. Every compatible clean label vector $z \in \mathcal{F}_n$ differs from $\widetilde{y}$ on at most k coordinates. We may therefore choose a set $C = \{c_1, \ldots, c_k\} \subseteq [n]$ of size exactly k that contains all of them, adding coordinates on which z agrees with $\widetilde{y}$ if necessary. Let $D = [n] \setminus C$ denote the retained coordinates. We call C the set of coordinates *allowed to change*; it need not coincide with the adversary's actual corruption set.

Once C is fixed, the labels on D equal their observed values, while those on C are free. We parameterize the free labels by their residuals with respect to the predictions:

$$z_i := \begin{cases} \widetilde{y}_i, & i \in D, \\ f_i + u_i, & i \in C, \end{cases}$$

where $u = (u_{c_1}, \ldots, u_{c_k}) \in \mathbb{R}^k$. Thus $u_i = z_i - f_i$ is the residual of the candidate clean label on coordinate i, and choosing $u_i = r_i$ leaves that coordinate unchanged. This parameterization is convenient because the MSE constraint depends on z only through its residuals. For fixed C, the sparsity constraint holds automatically, and the MSE constraint becomes a Euclidean ball constraint on u:

$$\sum_{i \in C} u_i^2 \leq n\beta^2 - \sum_{i \in D} r_i^2 =: R_C^2.$$

The quantity R_C^2 is the MSE budget remaining after accounting for the retained residuals. If $R_C^2 < 0$, the retained residuals alone exceed the budget, so no compatible clean label vector changes only coordinates in C; we call such C infeasible. For feasible C, define

$$\mathcal{V}(D) := \max_{u \in \mathbb{R}^k,\ \|u\|_2 \leq R_C} \mathrm{Var}_n(z).$$

Every such z differs from $\widetilde{y}$ on at most k coordinates and satisfies the MSE constraint, so it is a compatible clean label vector. Consequently,

$$V^\star_{\mathrm{MSE}} = \max_{\substack{C \subseteq [n],\ |C| = k \\ R_C^2 \geq 0}} \mathcal{V}([n] \setminus C).$$

We first show how to evaluate $\mathcal{V}(D)$ for a fixed D. We then approximate the outer maximization over C by a dynamic program.

To evaluate $\mathcal{V}(D)$, we write the variance as a quadratic form. Let $P = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ be the centering matrix, which subtracts the mean from every coordinate. Since P is symmetric and idempotent,

$$\mathrm{Var}_n(z) = \frac{1}{n}\|Pz\|_2^2 = \frac{1}{n}z^\top Pz.$$

We next separate the fixed and free parts of z. Define the baseline vector $v \in \mathbb{R}^n$ by

$$v_i \coloneqq \begin{cases} \widetilde{y}_i, & i \in D, \\ f_i, & i \in C. \end{cases}$$

This is the candidate obtained by setting $u = 0$. Index the entries of vectors in $\mathbb{R}^k$ by the elements of C. Let $J_C \in \mathbb{R}^{n\times k}$ be the selection matrix whose column indexed by $i \in C$ is the i-th standard basis vector of $\mathbb{R}^n$. Then $J_C u$ equals u_i on each $i \in C$ and zero elsewhere, so $z = v + J_C u$. Using $J_C^\top J_C = I_k$ and $J_C^\top \mathbf{1}_n = \mathbf{1}_k$, expanding the quadratic form gives

$$\mathrm{Var}_n(z) = \mathrm{Var}_n(v) + \frac{1}{n}\left(u^\top Q_C u + 2c^\top u\right),$$

where

$$Q_C \coloneqq J_C^\top P J_C = I_k - \frac{1}{n}\mathbf{1}_k\mathbf{1}_k^\top, \qquad c \coloneqq J_C^\top P v.$$

The first term is the variance of the baseline vector, which is a constant with respect to the optimization problem. The quadratic term satisfies $u^\top Q_C u = n\,\mathrm{Var}_n(J_C u)$, so it measures the variance contributed by the change alone. The vector c has entries $c_i = f_i - \overline{v}$ for $i \in C$, where $\overline{v} = \frac{1}{n}\sum_{i=1}^n v_i$ is the baseline mean. Thus $2c^\top u$ is the cross term between the baseline and the change. Consequently,

$$\mathcal{V}(D) = \mathrm{Var}_n(v) + \frac{1}{n}\max_{\|u\|_2 \le R_C}\left\{u^\top Q_C u + 2c^\top u\right\}. \tag{3}$$

For fixed C, computing $\mathcal{V}(D)$ is a trust-region problem: maximizing a quadratic function over a Euclidean ball. The value of this problem depends on C only through a few scalars, which we now define.

Let $\gamma = k/n$ denote the fraction of coordinates allowed to change, and recall that $B = \widehat{\sigma}_f + \beta$. Write $\overline{f}_C = \frac{1}{k}\sum_{i\in C} f_i$ for the mean prediction on C, and define

$$h_D \coloneqq \frac{R_C^2}{n}, \qquad q_D \coloneqq \sqrt{\gamma}\,(\overline{f}_C - \overline{v}), \qquad \ell_D \coloneqq \left(\frac{1}{n}\sum_{i\in C}(f_i - \overline{f}_C)^2\right)^{1/2}.$$

The quantity $h_D = \beta^2 - \frac{1}{n}\sum_{i\in D} r_i^2$ is the portion of the MSE budget left for the changeable coordinates. The quantity q_D measures the gap between the mean prediction on C and the baseline mean. The quantity ℓ_D measures the spread of the predictions within C.

In the following lemma, we will show how to solve Equation (3) efficiently. The proof works by decomposing u into two vectors, a uniform part, which moves the sample mean, and a zero-sum part, which does not. For each fixed value of the uniform part, we show that the zero-sum part can be optimized in closed form, which leaves a concave problem in one variable.

Lemma D.1. *For every $\tau > 0$, the value of the quadratic program in (3) can be approximated by a value $\mathcal{V}_\tau(D)$ satisfying*

$$\mathcal{V}(D) - \tau \le \mathcal{V}_\tau(D) \le \mathcal{V}(D).$$

The approximation requires $O(n + \log(1 + B^2/\tau))$ arithmetic operations.

Proof. We first treat the $k \ge 2$ case and handle $k = 1$ at the end. If $R_C = 0$, then $u = 0$ is the only feasible residual vector, so $\mathcal{V}(D) = \mathrm{Var}_n(v)$, and we set $\mathcal{V}_\tau(D) = \mathcal{V}(D)$. Henceforth suppose $R_C > 0$.

Let $e = \mathbf{1}_k/\sqrt{k}$ and write $u = te + w$, where $t = e^\top u$ and $w \perp e$. Then

$$\begin{aligned} u^\top Q_C u &= \|u\|_2^2 - \frac{1}{n}(\mathbf{1}_k^\top u)^2 \\ &= t^2 + \|w\|_2^2 - \frac{k}{n}t^2 \\ &= (1-\gamma)t^2 + \|w\|_2^2. \end{aligned}$$

Similarly, write $c = qe + l$ with $q = e^\top c$ and $l \perp e$. By the definitions of q_D and ℓ_D,

$$q = \sqrt{n}\, q_D, \qquad l = (f_i - \overline{f}_C)_{i \in C}, \qquad \|l\|_2 = \sqrt{n}\, \ell_D.$$

The objective in (3) therefore becomes

$$(1-\gamma)t^2 + 2qt + \|w\|_2^2 + 2l^\top w.$$

For fixed t, the constraint $\|u\|_2^2 = t^2 + \|w\|_2^2 \le R_C^2$ allows $\|w\|_2 \le \sqrt{R_C^2 - t^2}$. Both $\|w\|_2^2$ and $l^\top w$ are maximized by $w = \sqrt{R_C^2 - t^2}\, l / \|l\|_2$. If $l = 0$, we instead take any vector of this norm orthogonal to e, which exists since $k \ge 2$. Consequently,

$$\mathcal{V}(D) = \mathrm{Var}_n(v) + \frac{1}{n} \max_{|t| \le R_C} F_C(t),$$

where

$$F_C(t) = R_C^2 - \gamma t^2 + 2qt + 2\,\|l\|_2 \sqrt{R_C^2 - t^2}.$$

The function F_C is concave on $[-R_C, R_C]$, as a sum of a concave quadratic and a nonnegative multiple of $\sqrt{R_C^2 - t^2}$. If $l = 0$, then F_C is a concave quadratic and hence its maximum can be computed directly. Otherwise, F_C' is strictly decreasing on $(-R_C, R_C)$. Hence F_C has a unique maximizer, and we may use the bisection method to optimize it.

To do this, we first bound the coefficients of F_C. Since $R_C^2 \le n\beta^2$, we have $R_C \le \sqrt{n}\, B$. The baseline v is a compatible clean label vector, since $u = 0$ is feasible. The crude bound therefore gives $\mathrm{Var}_n(v) \le B^2$. Because c is a subvector of Pv,

$$|q|,\ \|l\|_2 \le \|c\|_2 \le \|Pv\|_2 = \sqrt{n\,\mathrm{Var}_n(v)} \le \sqrt{n}\, B.$$

Let $H(s) = F_C(R_C s)/n$ for $s \in [-1, 1]$. For $s, s' \in [-1, 1]$ with $d = |s - s'| \le 1$, we use $|s^2 - s'^2| \le 2d$ and the inequality $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a - b|}$ for $a, b \ge 0$. Bounding the three nonconstant terms separately gives

$$\begin{aligned} |H(s) - H(s')| &\le \frac{\gamma R_C^2}{n}|s^2 - s'^2| + \frac{2|q| R_C}{n}|s - s'| + \frac{2\,\|l\|_2\, R_C}{n}\left|\sqrt{1 - s^2} - \sqrt{1 - s'^2}\right| \\ &\le 2B^2 d + 2B^2 d + 2\sqrt{2}\, B^2 \sqrt{d} \le 8B^2\sqrt{d}. \end{aligned}$$

We run bisection on s until the interval containing the maximizer has length at most $\min\{1, \tau^2/(64B^4)\}$. Let s_τ be its midpoint, which lies within $\tau^2/(128B^4)$ of the maximizer, and set

$$\mathcal{V}_\tau(D) \coloneqq \mathrm{Var}_n(v) + H(s_\tau).$$

The preceding continuity bound gives $\mathcal{V}_\tau(D) \ge \mathcal{V}(D) - \tau$ and feasibility gives $\mathcal{V}_\tau(D) \le \mathcal{V}(D)$. For a fixed accuracy τ, the bisection method requires $O(1 + \log(1 + B^2/\tau))$ iterations.

We now handle the case where $k = 1$. Specifically, in this case $C = \{i\}$, so $e = 1$, $u = t$, $q = c$, and there is no orthogonal component w. The objective is

$$u^\top Q_C u + 2c^\top u = (1 - \gamma)t^2 + 2qt.$$

This is convex in t, so its maximum over $[-R_C, R_C]$ occurs at an endpoint. Therefore

$$\mathcal{V}(D) = \mathrm{Var}_n(v) + \frac{1}{n}\left[(1 - \gamma)R_C^2 + 2|q| R_C\right],$$

which we evaluate exactly and set $\mathcal{V}_\tau(D) = \mathcal{V}(D)$.

We conclude by observing that computing $\mathrm{Var}_n(v)$, R_C, q, and $\|l\|_2$ takes $O(n)$ operations. Each bisection step takes $O(1)$ operations. The total runtime is then $O(n + \log(1 + B^2/\tau))$. □

The lemma above shows that the value of the optimization program for a fixed corruption set C can be evaluated up to arbitrary additive accuracy in polynomial time. We next discretize the search space and use dynamic programming to obtain a polynomial-time additive approximation to Problem 1.

Discretizing the Search Space. In Lemma D.1, we show how to evaluate $\mathcal{V}(D)$ for a single retained set D, but there are $\binom{n}{k}$ candidates. We avoid enumerating them using the following observation: $\mathcal{V}(D)$ depends on D only through five aggregate statistics. We can therefore discretize these statistics and run a dynamic program over the resulting grid.

Recall that $\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$ is the prediction mean and $g_i = f_i - \overline{f}$ are the centered predictions. Thus $\sum_{i=1}^{n} g_i = 0$ and $\widehat{\sigma}_f^2 = \frac{1}{n}\sum_{i=1}^{n} g_i^2$. For a retained set D, define its summaries

$$\rho_1(D) \coloneqq \frac{1}{n}\sum_{i\in D} r_i, \qquad \rho_2(D) \coloneqq \frac{1}{n}\sum_{i\in D} r_i^2,$$

$$\rho_3(D) \coloneqq \frac{1}{n}\sum_{i\in D} g_i, \qquad \rho_4(D) \coloneqq \frac{1}{n}\sum_{i\in D} g_i^2, \qquad \rho_5(D) \coloneqq \frac{1}{n}\sum_{i\in D} g_i r_i.$$

The quantity ρ_1 is the retained contribution to the mean residual. The quantity ρ_2 is the portion of the MSE budget consumed by the retained coordinates, so $h_D = \beta^2 - \rho_2$. The quantities ρ_3 and ρ_4 are the retained contributions to the mean and second moment of the centered predictions. Finally, ρ_5 is the retained cross moment between the centered predictions and the residuals. When D is clear from context, we write simply $(\rho_1, \rho_2, \rho_3, \rho_4, \rho_5)$.

The next theorem shows that the fixed-support optimum $\mathcal{V}(D)$ depends on D only through the five summaries $\rho_1, \ldots, \rho_5$.

Theorem D.2 (Five-summary reduction)**.** *Let $k \geq 1$, and let $D \subseteq [n]$ be a feasible retained set, that is, $|D| = n - k$ and $\rho_2(D) \leq \beta^2$. Then the quantities h_D, q_D, and ℓ_D satisfy*

$$h_D = \beta^2 - \rho_2, \qquad q_D = -\left(\frac{\rho_3}{\sqrt{\gamma}} + \sqrt{\gamma}\,\rho_1\right), \qquad \ell_D^2 = \widehat{\sigma}_f^2 - \rho_4 - \frac{\rho_3^2}{\gamma}.$$

If $k \geq 2$, then

$$\mathcal{V}(D) = \widehat{\sigma}_f^2 + 2\rho_5 + \rho_2 - \rho_1^2 + \max_{|x|\leq\sqrt{h_D}} \left\{ h_D - \gamma x^2 + 2q_D x + 2\ell_D\sqrt{h_D - x^2} \right\}.$$

If $k = 1$, then

$$\mathcal{V}(D) = \widehat{\sigma}_f^2 + 2\rho_5 + \rho_2 - \rho_1^2 + (1-\gamma)h_D + 2|q_D|\sqrt{h_D}.$$

Proof. We first show that the leading terms equal $\mathrm{Var}_n(v)$. Recall that the baseline vector satisfies $v_i = f_i + r_i$ for $i \in D$ and $v_i = f_i$ for $i \in C$. Hence

$$v_i - \overline{f} = \begin{cases} g_i + r_i, & i \in D, \\ g_i, & i \in C. \end{cases}$$

Since $\sum_{i=1}^{n} g_i = 0$, the mean of v satisfies $\overline{v} = \overline{f} + \rho_1$. Expanding the variance around $\overline{f}$ therefore gives

$$\begin{aligned} \mathrm{Var}_n(v) &= \frac{1}{n}\left[\sum_{i\in D}(g_i + r_i)^2 + \sum_{i\in C} g_i^2\right] - \rho_1^2 \\ &= \widehat{\sigma}_f^2 + 2\rho_5 + \rho_2 - \rho_1^2. \end{aligned}$$

We next express h_D, q_D, and ℓ_D in terms of the summaries. We already showed that $h_D = \beta^2 - \rho_2$ and that it is nonnegative by feasibility. To compute q_D, observe that

$$\overline{f}_C - \overline{f} = \frac{1}{k}\sum_{i\in C} g_i = -\frac{1}{k}\sum_{i\in D} g_i = -\frac{\rho_3}{\gamma}.$$

Together with $\overline{v} = \overline{f} + \rho_1$, this implies

$$q_D = \sqrt{\gamma}\,(\overline{f}_C - \overline{v}) = -\frac{\rho_3}{\sqrt{\gamma}} - \sqrt{\gamma}\,\rho_1.$$

For ℓ_D, note that $f_i - \overline{f}_C = g_i - (\overline{f}_C - \overline{f})$, and $\overline{f}_C - \overline{f}$ is the mean of the g_i over C. Hence

$$\begin{aligned}\ell_D^2 &= \frac{1}{n}\sum_{i\in C}(f_i - \overline{f}_C)^2 \\ &= \frac{1}{n}\sum_{i\in C} g_i^2 - \frac{k}{n}(\overline{f}_C - \overline{f})^2 \\ &= \widehat{\sigma}_f^2 - \rho_4 - \frac{\rho_3^2}{\gamma}.\end{aligned}$$

In particular, the expression on the right is nonnegative, since ℓ_D^2 is a sum of squared deviations.

We now apply Lemma D.1. For $k \geq 2$, its scalar representation gives

$$\mathcal{V}(D) = \mathrm{Var}_n(v) + \frac{1}{n}\max_{|t|\leq R_C} F_C(t).$$

This also holds when $R_C = 0$, since both sides then equal $\mathrm{Var}_n(v)$. Using $R_C^2 = nh_D$, $q = \sqrt{n}\, q_D$, and $\|l\|_2 = \sqrt{n}\,\ell_D$, the change of variables $t = \sqrt{n}\, x$ gives

$$\frac{1}{n}F_C(\sqrt{n}\, x) = h_D - \gamma x^2 + 2q_D x + 2\ell_D\sqrt{h_D - x^2},$$

with $|x| \leq \sqrt{h_D}$. Substituting the expression for $\mathrm{Var}_n(v)$ proves the formula for $k \geq 2$. When $k = 1$, the same substitutions in the endpoint formula from Lemma D.1 give the stated expression. Thus, in either case, $\mathcal{V}(D)$ depends on D only through the five summaries. □

For every feasible D of size $n - k$, feasibility and Cauchy–Schwarz give

$$|\rho_1| \leq \beta, \qquad 0 \leq \rho_2 \leq \beta^2,$$

and

$$|\rho_3| \leq \sqrt{\gamma}\,\widehat{\sigma}_f, \qquad 0 \leq \rho_4 \leq \widehat{\sigma}_f^2, \qquad |\rho_5| \leq \widehat{\sigma}_f\beta.$$

The bound on ρ_3 uses $\sum_{i\in D} g_i = -\sum_{i\in C} g_i$ and $|C| = k$.

To approximate the search over D, we will group sets whose summaries $\rho_1, \rho_3, \rho_4, \rho_5$ are close and retain a representative with the smallest ρ_2. The following lemma bounds the possible decrease in the objective caused by this replacement.

Lemma D.3 (Stability of the fixed-support optimum)**.** *Suppose $k \geq 2$. Let $D, D' \subseteq [n]$ be feasible sets of size $n - k$ satisfying*

$$\rho_2(D') \leq \rho_2(D).$$

For $j \in \{1, 3, 4, 5\}$, define

$$\Delta_j := |\rho_j(D') - \rho_j(D)|.$$

Then

$$\mathcal{V}(D) - \mathcal{V}(D') \leq 2\Delta_5 + 4\beta\Delta_1 + \frac{2\beta\Delta_3}{\sqrt{\gamma}} + 2\beta\sqrt{\Delta_4 + \frac{2\widehat{\sigma}_f\Delta_3}{\sqrt{\gamma}}}.$$

Proof. Write $\rho_j = \rho_j(D)$ and $\rho'_j = \rho_j(D')$. Since $h_D = \beta^2 - \rho_2$, the expression in Theorem D.2 can be rewritten as

$$\mathcal{V}(D) = \widehat{\sigma}_f^2 + \beta^2 + 2\rho_5 - \rho_1^2 + \max_{|x|\leq\sqrt{h_D}}\left\{-\gamma x^2 + 2q_D x + 2\ell_D\sqrt{h_D - x^2}\right\}. \tag{4}$$

Here the terms ρ_2 and h_D have combined to give β^2. Let $x^\star$ be the point which attains the maximum for D. Since $\rho'_2 \leq \rho_2$

$$h_{D'} = \beta^2 - \rho'_2 \geq \beta^2 - \rho_2 = h_D,$$

so $x^\star$ is also admissible in the maximization for D'. Moreover, since $\ell_{D'} \geq 0$,

$$\ell_{D'}\sqrt{h_{D'} - (x^\star)^2} \geq \ell_{D'}\sqrt{h_D - (x^\star)^2}.$$

Thus, increasing the remaining MSE budget cannot decrease this term.

We now compare the two objective values at $x^\star$. Evaluating the objective for D' at $x^\star$ and using the preceding inequality gives

$$\mathcal{V}(D) - \mathcal{V}(D') \le 2(\rho_5 - \rho_5') + \big((\rho_1')^2 - \rho_1^2\big) + 2(q_D - q_{D'})\,x^\star + 2(\ell_D - \ell_{D'})\sqrt{h_D - (x^\star)^2}.$$

Feasibility gives $|\rho_1|, |\rho_1'| \le \beta$, so $|(\rho_1')^2 - \rho_1^2| \le 2\beta\Delta_1$. Since $h_D \le \beta^2$, both $|x^\star|$ and $\sqrt{h_D - (x^\star)^2}$ are at most β. Consequently,

$$\mathcal{V}(D) - \mathcal{V}(D') \le 2\Delta_5 + 2\beta\Delta_1 + 2\beta|q_{D'} - q_D| + 2\beta|\ell_{D'} - \ell_D|.$$

It remains to bound the differences in q_D and ℓ_D. By the expression for q_D in Theorem D.2,

$$\begin{aligned} |q_{D'} - q_D| &= \left|\frac{\rho_3' - \rho_3}{\sqrt{\gamma}} + \sqrt{\gamma}\,(\rho_1' - \rho_1)\right| \\ &\le \frac{\Delta_3}{\sqrt{\gamma}} + \sqrt{\gamma}\,\Delta_1. \end{aligned}$$

Similarly, using

$$\ell_D^2 = \widehat{\sigma}_f^2 - \rho_4 - \frac{\rho_3^2}{\gamma}$$

and the bounds $|\rho_3|, |\rho_3'| \le \sqrt{\gamma}\,\widehat{\sigma}_f$, we obtain

$$\begin{aligned} |\ell_{D'}^2 - \ell_D^2| &\le \Delta_4 + \frac{1}{\gamma}|(\rho_3')^2 - \rho_3^2| \\ &\le \Delta_4 + \frac{2\widehat{\sigma}_f\Delta_3}{\sqrt{\gamma}}. \end{aligned}$$

Since both ℓ_D and $\ell_{D'}$ are nonnegative, the inequality $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a-b|}$ gives

$$|\ell_{D'} - \ell_D| \le \sqrt{\Delta_4 + \frac{2\widehat{\sigma}_f\Delta_3}{\sqrt{\gamma}}}.$$

Substituting these estimates into the bound on $\mathcal{V}(D) - \mathcal{V}(D')$, and using $\sqrt{\gamma} \le 1$, proves the lemma. □

We now proceed with proving the final approximation guarantee using a dynamic program over the rounded summaries.

Theorem D.4 (Polynomial-time additive approximation)**.** *Fix a feasible instance of Problem 1, let $k = \lfloor \eta n \rfloor \ge 1$, and recall that $B = \widehat{\sigma}_f + \beta$. For every $\varepsilon \in (0,1)$, there is a dynamic program that returns a value $\overline{V}_{\mathrm{MSE}}$ satisfying*

$$V_{\mathrm{MSE}}^\star \le \overline{V}_{\mathrm{MSE}} \le V_{\mathrm{MSE}}^\star + \varepsilon B^2.$$

Its running time is $O\Big(\frac{n^6}{\varepsilon^6} + \frac{n^4}{\varepsilon^6}\log\frac{1}{\varepsilon}\Big)$, counting arithmetic operations, comparisons, square roots, and integer rounding at unit cost.

Proof. We begin with two simple cases. If $\beta = 0$, the MSE constraint forces every compatible clean label vector to equal f. Then $V_{\mathrm{MSE}}^\star = \widehat{\sigma}_f^2$, which we return. If $k = 1$, there are only n candidate sets C. Lemma D.1 evaluates each $\mathcal{V}(D)$ exactly through its endpoint formula, and we return the largest value over feasible sets, which equals $V_{\mathrm{MSE}}^\star$. Henceforth suppose $\beta > 0$ and $k \ge 2$. We break the proof up into 5 steps.

Step 1: rounding summaries. Write $m = n - k$ and $\xi = \varepsilon/24$. For $j \in \{1,3,4,5\}$, each summary $\rho_j(D)$ is a sum over $i \in D$ of a per-index contribution:

$$(a_{i1}, a_{i3}, a_{i4}, a_{i5}) = \frac{1}{n}(r_i, g_i, g_i^2, g_i r_i).$$

We round each contribution once, down to an integer multiple of a grid width ω_j:

$$b_{ij} = \left\lfloor \frac{a_{ij}}{\omega_j} \right\rfloor, \qquad \mathbf{b}_i = (b_{i1}, b_{i3}, b_{i4}, b_{i5}),$$

where

$$\omega_1 = \frac{B\xi}{n}, \qquad \omega_3 = \frac{B\sqrt{\gamma}\,\xi^2}{n}, \qquad \omega_4 = \frac{B^2\xi^2}{n}, \qquad \omega_5 = \frac{B^2\xi}{n}.$$

Because each contribution is rounded individually, the rounded sums $\sum_{i\in D}\mathbf{b}_i$ are additive in D, so the dynamic program can update them exactly. The widths are chosen so that, after at most n rounding errors accumulate, the summaries are known to the precision required by Lemma D.3. Δ_1 and Δ_5 need precision of order ξ. Δ_3 and Δ_4 need precision of order ξ^2, because they enter under a square root. ω_3 also carries a factor $\sqrt{\gamma}$, because the lemma divides Δ_3 by $\sqrt{\gamma}$. We do not round ρ_2 since the dynamic program minimizes it exactly.

Step 2: the dynamic program. We process the indices in order, deciding for each whether it is retained. After t indices, a state $(s, \mathbf{b})$ records the number s of retained indices and their four rounded sums. The entry $\mathsf{E}_t(s, \mathbf{b})$ stores the smallest value of $\rho_2(D)$ among sets $D \subseteq [t]$ with

$$|D| = s, \qquad \sum_{i\in D}\mathbf{b}_i = \mathbf{b}.$$

Along with it, we store the exact five summaries of a set attaining this minimum. Initialize $\mathsf{E}_0(0, \mathbf{0}) = 0$, with all other entries equal to $+\infty$. The recurrence is

$$\mathsf{E}_t(s, \mathbf{b}) = \min\left\{\mathsf{E}_{t-1}(s, \mathbf{b}),\ \mathsf{E}_{t-1}(s-1, \mathbf{b}-\mathbf{b}_t) + \frac{r_t^2}{n}\right\}.$$

The two terms correspond to excluding or including index t in D; including it adds r_t^2/n to ρ_2. When the second term attains the minimum, we update the stored summaries by adding the exact contributions of index t; ties may be resolved arbitrarily. After each step, we discard an entry if

$$s > m, \qquad t - s > k, \qquad \text{or} \qquad \mathsf{E}_t(s, \mathbf{b}) > \beta^2$$

since this prevents states which retain too many indices, exclude too many indices, or exceed the MSE budget, respectively.

At step n, every surviving state satisfies $s \le m$ and $n - s \le k$, and hence $s = m$. Each stored representative is therefore a feasible retained set. Let $\mathcal{D}_{\mathrm{DP}}$ denote the collection of these representatives. For each one, we compute $\mathcal{V}_\tau(D)$ using Lemma D.1 with tolerance $\tau = \varepsilon B^2/2$, obtaining the coefficients from the stored summaries via Theorem D.2. We return

$$\overline{V}_{\mathrm{MSE}} = L + \varepsilon B^2, \qquad L = \max_{D\in\mathcal{D}_{\mathrm{DP}}} \mathcal{V}_\tau(D).$$

Step 3: retaining feasible sets. The dynamic program keeps only one representative per state and prunes states. We show that every feasible set is nonetheless dominated by a surviving representative: one with the same rounded sums and no larger ρ_2. Step 4 applies this to an optimal set.

Fix a feasible retained set D of size m. For each t, let $D_t = D \cap [t]$ be its restriction to the processed indices. Let $s_t = |D_t|$ and $\mathbf{B}_t = \sum_{i\in D_t}\mathbf{b}_i$ denote the state of D_t. We claim that the entry $\mathsf{E}_t(s_t, \mathbf{B}_t)$ survives and satisfies $\mathsf{E}_t(s_t, \mathbf{B}_t) \le \rho_2(D_t)$.

We argue by induction on t. For $t = 0$, we have $D_0 = \varnothing$ and $\mathsf{E}_0(0, \mathbf{0}) = 0 = \rho_2(\varnothing)$. Suppose the claim holds at step $t-1$. If $t \in D$, then $s_t = s_{t-1} + 1$ and $\mathbf{B}_t = \mathbf{B}_{t-1} + \mathbf{b}_t$. The second term of the recurrence therefore gives

$$\mathsf{E}_t(s_t, \mathbf{B}_t) \le \mathsf{E}_{t-1}(s_{t-1}, \mathbf{B}_{t-1}) + \frac{r_t^2}{n} \le \rho_2(D_{t-1}) + \frac{r_t^2}{n} = \rho_2(D_t).$$

If $t \notin D$, the state is unchanged, and the first term gives $\mathsf{E}_t(s_t, \mathbf{B}_t) \le \rho_2(D_{t-1}) = \rho_2(D_t)$. The entry survives pruning, since $s_t \le m$, $t - s_t \le k$, and $\mathsf{E}_t(s_t, \mathbf{B}_t) \le \rho_2(D_t) \le \rho_2(D) \le \beta^2$.

Step 4: approximation guarantee. Observe that some retained set $D^\star$ satisfies $\mathcal{V}(D^\star) = V^\star_{\mathrm{MSE}}$. Step 3 with $t = n$ gives a representative $\widehat{D} \in \mathcal{D}_{\mathrm{DP}}$ with the same size and rounded sums as $D^\star$ and with $\rho_2(\widehat{D}) \le \rho_2(D^\star)$. For any set D and $j \in \{1, 3, 4, 5\}$, rounding gives

$$0 \le \rho_j(D) - \omega_j \sum_{i\in D} b_{ij} \le |D|\,\omega_j.$$

The sets $\widehat{D}$ and $D^\star$ share the same size m and the same rounded sums. Their exact summaries therefore lie in a common interval of length $m\omega_j \le n\omega_j$, and so

$$\Delta_1 \le B\xi, \qquad \Delta_3 \le B\sqrt{\gamma}\,\xi^2, \qquad \Delta_4 \le B^2\xi^2, \qquad \Delta_5 \le B^2\xi.$$

We apply Lemma D.3 with $D = D^\star$ and $D' = \widehat{D}$. Using $\beta, \widehat{\sigma}_f \le B$, the four terms of its bound satisfy

$$2\Delta_5 \le 2B^2\xi, \qquad 4\beta\Delta_1 \le 4B^2\xi, \qquad \frac{2\beta\Delta_3}{\sqrt{\gamma}} \le 2B^2\xi^2,$$

$$2\beta\sqrt{\Delta_4 + \frac{2\widehat{\sigma}_f\Delta_3}{\sqrt{\gamma}}} \le 2B\sqrt{3B^2\xi^2} = 2\sqrt{3}\,B^2\xi.$$

Summing these and using $\xi \le 1$,

$$\mathcal{V}(D^\star) - \mathcal{V}(\widehat{D}) \le (8 + 2\sqrt{3})B^2\xi \le 12B^2\xi = \frac{\varepsilon B^2}{2}.$$

By the guarantee of Lemma D.1 with $\tau = \varepsilon B^2/2$,

$$L \ge \mathcal{V}_\tau(\widehat{D}) \ge \mathcal{V}(\widehat{D}) - \frac{\varepsilon B^2}{2} \ge V^\star_{\mathrm{MSE}} - \varepsilon B^2.$$

Conversely, every representative is feasible, and each evaluation is from below, so $L \le V^\star_{\mathrm{MSE}}$. Adding εB^2 proves the guarantee.

Step 5: runtime analysis. The runtime of the dynamic program is dominated by the number of states, which we will now count. For a surviving set $D \subseteq [t]$, recall that its rounded summary is $\sum_{i\in D} b_{ij}$. By the rounding rule in Step 1,

$$\left|\sum_{i\in D} b_{ij} - \frac{\rho_j(D)}{\omega_j}\right| \le |D|.$$

Applying the triangle inequality and using $|D| \le n$, we obtain

$$\begin{aligned}\left|\sum_{i\in D} b_{ij}\right| &\le \frac{|\rho_j(D)|}{\omega_j} + \left|\sum_{i\in D} b_{ij} - \frac{\rho_j(D)}{\omega_j}\right| \\ &\le \frac{|\rho_j(D)|}{\omega_j} + |D| \le \frac{|\rho_j(D)|}{\omega_j} + n.\end{aligned}$$

Recall that

$$|\rho_1| \le \beta, \qquad \rho_2(D) \le \beta^2, \qquad 0 \le \rho_4 \le \widehat{\sigma}_f^2, \qquad |\rho_5| \le \widehat{\sigma}_f\beta.$$

Substituting these bounds and the grid widths from Step 1 into the preceding triangle inequality, and using $\beta, \widehat{\sigma}_f \le B$, gives

$$\begin{aligned}\left|\sum_{i\in D} b_{i1}\right| &\le \frac{\beta}{\omega_1} + n = \frac{n\beta}{B\xi} + n \le \frac{n}{\xi} + n, \\ \left|\sum_{i\in D} b_{i4}\right| &\le \frac{\widehat{\sigma}_f^2}{\omega_4} + n = \frac{n\widehat{\sigma}_f^2}{B^2\xi^2} + n \le \frac{n}{\xi^2} + n, \\ \left|\sum_{i\in D} b_{i5}\right| &\le \frac{\widehat{\sigma}_f\beta}{\omega_5} + n = \frac{n\widehat{\sigma}_f\beta}{B^2\xi} + n \le \frac{n}{\xi} + n.\end{aligned}$$

Since $\xi < 1$, the additive n is absorbed into each bound. Thus, the first, fourth, and fifth rounded sums take $O(n/\xi)$, $O(n/\xi^2)$, and $O(n/\xi)$ values, respectively. The third sum needs more care. A direct bound $|\rho_3| \le \widehat{\sigma}_f$ would lose a factor $1/\sqrt{\gamma}$ after dividing by ω_3. Instead, we measure the third sum relative to

$$K_t = \sum_{i=1}^{t} b_{i3},$$

its value if every processed index were retained. Since $D \subseteq [t]$, it holds that

$$\sum_{i \in D} b_{i3} - K_t = - \sum_{i \in [t] \setminus D} b_{i3}.$$

Pruning guarantees $|[t] \setminus D| = t - s \leq k$, so it suffices to bound a rounded sum over at most k indices. Recall that $a_{i3} = g_i/n$. We first bound the sum of these original contributions, before dividing by ω_3 and accounting for rounding. Since the excluded set $A = [t] \setminus D$ contains at most k indices, Cauchy–Schwarz gives

$$\left| \frac{1}{n} \sum_{i \in A} g_i \right| \leq \frac{\sqrt{|A|}}{n} \left(\sum_{i \in A} g_i^2 \right)^{1/2} \leq \frac{\sqrt{k}}{n} \left(\sum_{i=1}^{n} g_i^2 \right)^{1/2} = \sqrt{\gamma}\, \widehat{\sigma}_f.$$

Here we used $\gamma = k/n$ and $\sum_{i=1}^{n} g_i^2 = n\widehat{\sigma}_f^2$. Since each application of the floor loses less than one unit and there are at most k terms, it holds that

$$\left| \sum_{i \in D} b_{i3} - K_t \right| \leq \frac{\sqrt{\gamma}\, \widehat{\sigma}_f}{\omega_3} + k \leq \frac{n}{\xi^2} + n.$$

The shift K_t depends only on t, not on D, and all shifts can be computed in $O(n)$ time. By storing $\sum_{i \in D} b_{i3} - K_t$ instead of the raw sum, the third coordinate also takes $O(n/\xi^2)$ values.

Thus, there are $O(n^4/\xi^6)$ rounded states for each cardinality. We index the tables by these bounded integer ranges, so each transition, including the update of the stored summaries, takes constant time. With at most $n + 1$ cardinalities and n steps, the dynamic program requires $O(n^6/\xi^6)$ operations.

We evaluate only the representatives in the final layer $t = n$. By the pruning rules, every surviving representative has the same cardinality $s = |D| = n - k$. Thus, the bound for a fixed layer and cardinality gives $O(n^4/\xi^6)$ representatives to evaluate. For each representative, we obtain the coefficients from the stored summaries in constant time and evaluate Equation (3) using the bisection method which requires $O(1 + \log(1 + B^2/\tau)) = O(1 + \log(1/\varepsilon))$ operations by Lemma D.1. Computing and rounding the contributions takes $O(n)$ time. Substituting $\xi = \varepsilon/24$ gives the stated runtime. □

E Alignment Variance Bound

In this section, we prove the alignment bound (Theorem 3.2) and explain its computation, which is summarized in Algorithm 2. Throughout, we fix a feasible observed instance, so that $\mathcal{F}_n \neq \varnothing$. The bounds below hold deterministically for every compatible clean label vector. In particular, they apply to the actual clean sample on the event G_n.

We begin with the observation that the crude bound $B^2 = (\widehat{\sigma}_f + \beta)^2$ can be arbitrarily loose because it ignores the observed labels (see Example 2). The alignment bound improves on the crude bound by using the fact that at least $n - k$ coordinates retain their observed residuals, and on these coordinates the slack in the crude bound can be computed exactly.

Theorem 3.2 (Alignment variance bound)**.** *For every feasible instance, Algorithm 1 computes in time $O(n \log n)$ a bound satisfying $V^{\star}_{\mathrm{MSE}} \leq \widehat{V}_{\mathrm{align}} \leq B^2$, where $B = \widehat{\sigma}_f + \beta$.*

Proof. Fix a feasible instance and write $k = \lfloor \eta n \rfloor$ and $m = n - k$. For any compatible clean label vector z, choose a set $C \subseteq [n]$ of size k containing every coordinate on which z differs from $\widetilde{y}$, and let $D = [n] \setminus C$. Such a set exists because at most k coordinates differ; additional coordinates included in C may retain their observed values. Thus, $|D| = m$.

As before, write

$$f_i = f(x_i), \qquad r_i = \widetilde{y}_i - f_i, \qquad \overline{f} = \frac{1}{n} \sum_{i=1}^{n} f_i, \qquad g_i = f_i - \overline{f}.$$

Then $\sum_{i=1}^n g_i = 0$ and $\widehat{\sigma}_f^2 = \frac{1}{n}\sum_{i=1}^n g_i^2$. Write $z = f + e$, where

$$e_i = \begin{cases} r_i, & i \in D, \\ u_i, & i \in C, \end{cases} \qquad \frac{1}{n}\sum_{i=1}^n e_i^2 \le \beta^2.$$

Since centering the predictions does not change the variance,

$$\begin{aligned} \mathrm{Var}_n(z) &= \frac{1}{n}\sum_{i=1}^n (g_i + e_i)^2 - \left(\frac{1}{n}\sum_{i=1}^n e_i\right)^2 \\ &\le \frac{1}{n}\sum_{i=1}^n (g_i + e_i)^2. \end{aligned}$$

The crude bound ignores that $e_i = r_i$ on D. To retain this information, fix $a > 0$. Young's inequality gives

$$2g_i e_i \le a g_i^2 + \frac{e_i^2}{a},$$

and hence

$$(g_i + e_i)^2 \le (1 + a) g_i^2 + \left(1 + \frac{1}{a}\right) e_i^2.$$

On D, however, the residual is fixed, so we can retain the exact slack in this inequality:

$$(g_i + r_i)^2 = (1 + a) g_i^2 + \left(1 + \frac{1}{a}\right) r_i^2 - \left(\sqrt{a}\, g_i - \frac{r_i}{\sqrt{a}}\right)^2.$$

Define

$$s_i(a) = \left(\sqrt{a}\, g_i - \frac{r_i}{\sqrt{a}}\right)^2 = \frac{(r_i - a g_i)^2}{a}.$$

This penalty measures the squared discrepancy between the observed residual r_i and the aligned value $a g_i$. Combining the exact identity on D with Young's inequality on C yields

$$\begin{aligned} \mathrm{Var}_n(z) &\le \frac{1}{n}\sum_{i=1}^n \left[(1 + a) g_i^2 + \left(1 + \frac{1}{a}\right) e_i^2\right] - \frac{1}{n}\sum_{i \in D} s_i(a) \\ &\le (1 + a)\widehat{\sigma}_f^2 + \left(1 + \frac{1}{a}\right)\beta^2 - \frac{1}{n}\sum_{i \in D} s_i(a). \end{aligned}$$

For a retained set D of size m, define

$$\rho_2(D) = \frac{1}{n}\sum_{i \in D} r_i^2.$$

Such a set is feasible precisely when $\rho_2(D) \le \beta^2$. For each feasible D, let

$$\mathcal{V}(D) = \max_{\substack{e \in \mathbb{R}^n,\ e_i = r_i\ (i \in D) \\ \frac{1}{n}\sum_{i=1}^n e_i^2 \le \beta^2}} \mathrm{Var}_n(f + e).$$

The preceding argument shows that

$$\mathcal{V}(D) \le (1 + a)\widehat{\sigma}_f^2 + \left(1 + \frac{1}{a}\right)\beta^2 - \frac{1}{n}\sum_{i \in D} s_i(a).$$

Every compatible clean vector is represented by at least one feasible retained set D. Maximizing over these sets therefore gives

$$V^\star_{\mathrm{MSE}} \le (1 + a)\widehat{\sigma}_f^2 + \left(1 + \frac{1}{a}\right)\beta^2 - P(a),$$

where

$$P(a) = \min_{\substack{D \subseteq [n],\ |D| = m \\ \rho_2(D) \le \beta^2}} \frac{1}{n}\sum_{i \in D} s_i(a).$$

The minimization defining $P(a)$ still requires selecting a subset subject to the MSE constraint. We obtain an efficiently computable relaxation by dropping this constraint. Writing

$$s_{(1)}(a) \le \cdots \le s_{(n)}(a)$$

for the sorted penalties, the relaxed minimum is

$$P_{\mathrm{sort}}(a) = \min_{\substack{D\subseteq[n]\\|D|=m}} \frac{1}{n}\sum_{i\in D} s_i(a) = \frac{1}{n}\sum_{j=1}^{m} s_{(j)}(a).$$

The relaxed minimization ranges over a larger collection of sets, so $P_{\mathrm{sort}}(a) \le P(a)$. Subtracting this smaller correction gives the valid upper bound

$$V^{\star}_{\mathrm{MSE}} \le U_{\mathrm{align}}(a) \coloneqq (1+a)\widehat{\sigma}_f^2 + \left(1+\frac{1}{a}\right)\beta^2 - P_{\mathrm{sort}}(a) \tag{5}$$

for every $a > 0$. We now provide a value for a that guarantees a bound no worse than B^2. Suppose $\beta\widehat{\sigma}_f > 0$ and choose

$$a_0 = \frac{\beta}{\widehat{\sigma}_f}.$$

At this scale,

$$(1+a_0)\widehat{\sigma}_f^2 + \left(1+\frac{1}{a_0}\right)\beta^2 = (\widehat{\sigma}_f+\beta)^2.$$

Consequently,

$$\widehat{V}_{\mathrm{align}} = U_{\mathrm{align}}(a_0) = (\widehat{\sigma}_f+\beta)^2 - P_{\mathrm{sort}}(a_0) \le (\widehat{\sigma}_f+\beta)^2.$$

Since $P_{\mathrm{sort}}(a_0) \ge 0$, the alignment certificate is never larger than the crude bound, and it is strictly smaller whenever this correction is positive. Together with the preceding bound, this yields

$$\mathrm{Var}_n(y) \le V^{\star}_{\mathrm{MSE}} \le \widehat{V}_{\mathrm{align}} \le (\widehat{\sigma}_f+\beta)^2$$

for every compatible clean label vector y.

The runtime is dominated by sorting the penalties, which takes $O(n\log n)$ time. Lastly, if $\beta\widehat{\sigma}_f = 0$, we return the crude bound, whose validity follows directly from Cauchy–Schwarz and the MSE constraint. □

We briefly remark that the alignment bound can be further tightened by optimizing over the scale. The bound $V^{\star}_{\mathrm{MSE}} \le U_{\mathrm{align}}(a)$ in (5) holds for every $a > 0$. Hence, for any finite set of scales $\mathcal{A} \subset (0,\infty)$, the quantity $\min_{a\in\mathcal{A}} U_{\mathrm{align}}(a)$ is also an upper bound on $V^{\star}_{\mathrm{MSE}}$. If $a_0 \in \mathcal{A}$, this bound is never larger than $\widehat{V}_{\mathrm{align}}$. Since the penalties $s_i(a)$ depend on a, each scale requires a new sort, for a total cost of $O(|\mathcal{A}|\, n\log n)$.

F Implementing the Midpoint Estimator

In this section, we prove Theorems 4.1 and 4.2 which collectively establish the error guarantee and computational efficiency of the midpoint estimator, respectively.

F.1 Proof of Theorem 4.1

To begin, fix a feasible observed instance of Problem 1, write $f_i = f(x_i)$ and $r_i = \widetilde{y}_i - f_i$, and assume $1 \le k = \lfloor \eta n \rfloor \le n/2$. Recall that the compatible clean residual vectors form the set

$$\mathcal{R}(r) = \left\{u \in \mathbb{R}^n : \|u\|_2 \le \beta\sqrt{n}, \quad \|u-r\|_0 \le k\right\} = \{z - f : z \in \mathcal{F}_n\}.$$

The results below hold deterministically for every compatible clean label vector. They therefore apply to the actual clean sample on G_n. The set $\mathcal{R}(r)$ is nonempty and compact: it is a finite union of closed subsets of the ball of radius $\beta\sqrt{n}$. Consequently, the extreme residual means

$$U(r) = \frac{1}{n}\max_{u\in\mathcal{R}(r)}\sum_{i=1}^{n} u_i, \qquad L(r) = \frac{1}{n}\min_{u\in\mathcal{R}(r)}\sum_{i=1}^{n} u_i$$

are attained. As in Section 4, define

$$\widehat{\mu}_{\mathrm{mid}} = \overline{f} + \frac{U(r) + L(r)}{2}, \qquad b_n = \frac{U(r) - L(r)}{2}.$$

Below, we characterize the error of the midpoint estimator relative to the clean sample mean.

Lemma F.1. *For every compatible clean label vector $y \in \mathcal{F}_n$,*

$$|\widehat{\mu}_n - \widehat{\mu}_{\mathrm{mid}}| \le b_n \le \beta\sqrt{\frac{2k}{n}}, \qquad \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

where $\widehat{\mu}_n = n^{-1}\sum_{i=1}^{n} y_i$. For the given observed instance, b_n is the smallest worst-case error attainable by any estimate of the clean sample mean.

Proof. Since $y - f \in \mathcal{R}(r)$, the clean sample mean satisfies

$$\widehat{\mu}_n = \overline{f} + \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i) \in [\overline{f} + L(r),\, \overline{f} + U(r)].$$

Every point in this interval is within a half-width of the midpoint, so $|\widehat{\mu}_n - \widehat{\mu}_{\mathrm{mid}}| \le b_n$. It remains to bound the interval's width.

Let $u, v \in \mathcal{R}(r)$. Each vector differs from r on at most k coordinates, so u and v differ on at most $2k$ coordinates. Cauchy–Schwarz on those coordinates and the triangle inequality give

$$\left|\frac{1}{n}\sum_{i=1}^{n}(u_i - v_i)\right| \le \frac{\sqrt{2k}}{n}\|u - v\|_2 \le \frac{\sqrt{2k}}{n}\big(\|u\|_2 + \|v\|_2\big) \le 2\beta\sqrt{\frac{2k}{n}}.$$

Let u and v be the vectors that attain $U(r)$ and $L(r)$, respectively. Then, the previous bound implies that $U(r) - L(r) \le 2\beta\sqrt{2k/n}$. Dividing by two proves the error bound. We can easily show that any other estimate $t \in \mathbb{R}$ must have worst-case error at least

$$\max\{|t - (\overline{f} + L(r))|,\, |t - (\overline{f} + U(r))|\} \ge \frac{U(r) - L(r)}{2} = b_n.$$

The midpoint attains this bound, proving its optimality for the observed instance. □

Lemma F.1 establishes optimality for each observed instance and bounds the error radius b_n by $\beta\sqrt{2k/n}$. The following example shows that this bound cannot be improved. It constructs an observed instance with two compatible clean samples whose means differ by $2\beta\sqrt{2k/n}$, so that $b_n = \beta\sqrt{2k/n}$.

Example 1 (Indistinguishability lower bound)**.** Let $k = \lfloor \eta n \rfloor$. We construct a feasible observed instance with two compatible clean label vectors whose sample means differ by $2\beta\sqrt{2k/n}$. This implies that no estimator can uniformly achieve error smaller than $\beta\sqrt{2k/n}$.

Let $S, S' \subseteq [n]$ be disjoint sets with $|S| = |S'| = k$, and let $W = S \cup S'$. Set $f(x_i) = 0$ for all i, and define

$$q = \beta\sqrt{\frac{n}{2k}}.$$

Consider two possible clean samples:

$$y_i^+ = \begin{cases} q, & i \in W, \\ 0, & i \notin W, \end{cases} \qquad y_i^- = \begin{cases} -q, & i \in W, \\ 0, & i \notin W. \end{cases}$$

Since $f(x_i) = 0$, the clean residuals are $r_i^+ = y_i^+$ and $r_i^- = y_i^-$. Both clean samples satisfy

$$\frac{1}{n}\sum_{i=1}^{n}(y_i^+ - f(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^- - f(x_i))^2 = \frac{2kq^2}{n} = \beta^2.$$

Now define the observed corrupted labels by

$$\widetilde{y}_i = \begin{cases} -q, & i \in S, \\ q, & i \in S', \\ 0, & i \notin W. \end{cases}$$

The corrupted residuals also satisfy

$$\frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}_i - f(x_i))^2 = \frac{2kq^2}{n} = \beta^2.$$

Moreover, the same corrupted sample can arise from the clean sample y^+ by corrupting the points in S, and it can arise from the clean sample y^- by corrupting the points in S'. In both cases, only k labels are corrupted. Thus any estimator sees the same corrupted sample in the two cases and must return the same value, say m. But the corresponding clean sample means are

$$\widehat{\mu}_n^+ = \frac{1}{n}\sum_{i=1}^{n} y_i^+ = \frac{2kq}{n} = \beta\sqrt{\frac{2k}{n}},$$

and

$$\widehat{\mu}_n^- = \frac{1}{n}\sum_{i=1}^{n} y_i^- = -\frac{2kq}{n} = -\beta\sqrt{\frac{2k}{n}}.$$

Therefore,

$$\max\left\{|m - \widehat{\mu}_n^+|, |m - \widehat{\mu}_n^-|\right\} \geq \beta\sqrt{\frac{2k}{n}}.$$

Hence no estimator can uniformly guarantee error smaller than

$$\sqrt{\frac{2k}{n}}\beta.$$

With the two previous results in hand, we are now ready to prove the error guarantee of the midpoint estimator.

Theorem 4.1 (Optimal midpoint error)**.** *Suppose $1 \leq k \leq n/2$. For every compatible clean label vector $y \in \mathcal{F}_n$, with sample mean $\widehat{\mu}_n = n^{-1}\sum_{i=1}^{n} y_i$,*

$$|\widehat{\mu}_{\mathrm{mid}} - \widehat{\mu}_n| \leq b_n \leq \beta\sqrt{2k/n}.$$

For each observed instance, the midpoint minimizes the worst-case error over compatible clean samples, with optimal error b_n. Moreover, the bound $\beta\sqrt{2k/n}$ cannot be improved: for every n, k, and β, there is a feasible instance on which $b_n = \beta\sqrt{2k/n}$.

Proof. The error bound and the optimality of the midpoint for each observed instance are established in Lemma F.1. For the final statement, Example 1 constructs, for any n, k, and β with $1 \leq k \leq n/2$, a feasible instance on which $b_n = \beta\sqrt{2k/n}$. Hence the bound $\beta\sqrt{2k/n}$ cannot be improved. □

F.2 Proof of Theorem 4.2

We now show how to compute $U(r)$ and $L(r)$. It suffices to compute $U(r)$. Indeed, since $\mathcal{R}(-r) = -\mathcal{R}(r)$, it is straightforward to show that $L(r) = -U(-r)$. The computation proceeds in three steps. First, for a fixed set of coordinates allowed to change, we compute the largest attainable residual sum in closed form. Next, we show that an optimal set of modified coordinates consists of the smallest and largest observed residuals, which leaves only $k+1$ candidate sets to check. Finally, we evaluate all candidates in $O(n \log n)$ total time.

To begin, fix a set $C \subseteq [n]$ of coordinates allowed to change, and require $u_i = r_i$ for $i \notin C$. A candidate may still agree with r on coordinates in C, so enlarging C excludes no candidate. We may therefore assume without loss of generality that $|C| = k$. The squared-error budget remaining for the coordinates in C is

$$H_C := n\beta^2 - \sum_{i \notin C} r_i^2.$$

A feasible candidate with these retained coordinates exists if and only if $H_C \geq 0$. When this holds, Cauchy–Schwarz gives

$$\sum_{i \in C} u_i \leq \sqrt{k}\, \|u_C\|_2 \leq \sqrt{kH_C},$$

with equality when $u_i = \sqrt{H_C/k}$ for every $i \in C$. The largest total residual sum for this choice of C is

$$\Phi(C) := \sum_{i \notin C} r_i + \sqrt{kH_C}.$$

Thus, for a fixed corruption set, we have a closed form for the largest obtainable residual sum. Next, define the family of feasible allowed sets

$$\mathcal{C}_k := \{C \subseteq [n] : |C| = k,\ H_C \geq 0\}.$$

Every candidate $u \in \mathcal{R}(r)$ differs from r on at most k coordinates. Enlarging this set of coordinates to size k gives a set $C \in \mathcal{C}_k$ with $\sum_{i=1}^n u_i \leq \Phi(C)$. Conversely, for every $C \in \mathcal{C}_k$, the maximizing candidate constructed above belongs to $\mathcal{R}(r)$. Since $\mathcal{R}(r)$ is nonempty, so is $\mathcal{C}_k$, and these two directions establish

$$nU(r) = \max_{C \in \mathcal{C}_k} \Phi(C). \tag{6}$$

To compare sets, write

$$\Delta := n\beta^2 - \sum_{i=1}^n r_i^2, \qquad H_C = \Delta + \sum_{i \in C} r_i^2.$$

Here Δ is the supplied budget minus the observed squared residual sum, so it may be negative. Then observe that

$$\Phi(C) = \sum_{i=1}^n r_i + h(C)$$

where

$$h(C) := -\sum_{i \in C} r_i + \sqrt{kH_C} \tag{7}$$

is the largest increase in the total residual sum obtained by allowing the coordinates in C to change. Maximizing Φ over $\mathcal{C}_k$ is therefore equivalent to maximizing h.

The form of h suggests which coordinates are worth allowing to change for our objective. A coordinate with a very negative residual helps in both terms: allowing it to change removes that residual from the retained sum and frees r_i^2 of budget. A coordinate with a very positive residual costs r_i in the first term but can still help by freeing r_i^2 of budget. Regardless, the coordinates worth changing are those with extreme residuals.

To make this precise, sort the observed residuals as $r_{(1)} \leq \cdots \leq r_{(n)}$, breaking ties arbitrarily. The following lemma confirms this intuition and reduces the search from $\binom{n}{k}$ sets to $k+1$.

Lemma F.2 (Prefix-plus-suffix structure)**.** *The function h attains its maximum over $\mathcal{C}_k$ at a set*

$$C = \{(1), \ldots, (p)\} \cup \{(n-k+p+1), \ldots, (n)\}$$

for some $p \in \{0, \ldots, k\}$, with an empty prefix or suffix allowed. Equivalently, the retained coordinates $[n] \setminus C$ form a contiguous block in sorted order.

Proof. Among the maximizers of h, choose $C^\star$ to maximize the sum of selected observed residuals,

$$P(C) := \sum_{i \in C} r_i.$$

Exchanges between equal residuals preserve feasibility, h, and P. If such exchanges make the retained coordinates of $C^\star$ contiguous in sorted order, we are done. Otherwise, some selected residual lies strictly between two retained ones: there are $j \in C^\star$ and $a, b \notin C^\star$ with

$$r_a < r_j < r_b.$$

To compare j with a and b, we fix the other $k-1$ allowed coordinates and view h as a function of the remaining slot. Let $C_{-j} := C^\star \setminus \{j\}$ denote these fixed coordinates, and fill the remaining slot with a coordinate $\ell \notin C_{-j}$. Taking $\ell = j$ recovers $C^\star$.

Before the last slot is filled, every coordinate outside C_{-j} is retained, so the remaining squared-error budget is

$$K := n\beta^2 - \sum_{i \notin C_{-j}} r_i^2 = \Delta + \sum_{i \in C_{-j}} r_i^2.$$

This quantity may be negative, since the retained residuals may already exceed the MSE budget. Filling the last slot with ℓ releases that coordinate's contribution r_ℓ^2, so the budget for the k allowed coordinates becomes $K + r_\ell^2$. In particular, the replacement is feasible if and only if $K + r_\ell^2 \geq 0$.

The effect of the last slot therefore depends only on the residual $t = r_\ell$ placed in it. Define

$$\Psi(t) := -t + \sqrt{k(K + t^2)}, \qquad K + t^2 \geq 0.$$

The term $-t$ accounts for removing the observed residual from the retained sum. The square-root term is the largest total residual attainable on the k allowed coordinates with budget $K + t^2$. Hence, by (7),

$$h(C_{-j} \cup \{\ell\}) = -\sum_{i \in C_{-j}} r_i + \Psi(r_\ell).$$

The first term does not depend on ℓ, so replacing j by ℓ changes h by $\Psi(r_\ell) - \Psi(r_j)$. Since $C^\star$ is optimal, every feasible replacement satisfies $\Psi(r_\ell) \leq \Psi(r_j)$.

Case $K < 0$. When $K < 0$, the fixed coordinates leave a budget deficit of $-K$, which the last slot must cover to retain feasibility. A replacement with residual t frees t^2 of budget, so it is feasible if and only if $t^2 \geq -K$, that is, $|t| \geq \sqrt{-K}$. The domain of Ψ therefore consists of the two intervals

$$(-\infty, -\sqrt{-K}] \qquad \text{and} \qquad [\sqrt{-K}, \infty).$$

On their interiors,

$$\Psi'(t) = -1 + \frac{\sqrt{k}\, t}{\sqrt{K + t^2}}.$$

Since $k \geq 1$ and $\sqrt{K + t^2} < |t|$, this derivative is negative on the left interval and positive on the right interval. By continuity, Ψ is strictly decreasing on the entire left interval and strictly increasing on the entire right interval. If r_j lies in the left interval, replacing j by a is feasible and increases the objective. If r_j lies in the right interval, replacing j by b does so. Either exchange contradicts optimality.

Case $K \geq 0$. Every replacement is now feasible, and Ψ is convex on $\mathbb{R}$: the function $\sqrt{K + t^2} = \|(\sqrt{K}, t)\|_2$ is convex, and Ψ is a positive multiple of it plus a linear term. Write

$$r_j = \theta r_a + (1 - \theta) r_b, \qquad 0 < \theta < 1.$$

Convexity and optimality of $C^\star$ give

$$\Psi(r_j) \leq \theta \Psi(r_a) + (1 - \theta)\Psi(r_b) \leq \Psi(r_j).$$

This forces $\Psi(r_b) = \Psi(r_j)$. Replacing j by b therefore preserves the maximum of h but increases P by $r_b - r_j > 0$, contradicting the choice of $C^\star$. □

We are now ready to prove the runtime of the midpoint estimator.

Theorem 4.2 (Computability of the midpoint)**.** *The quantities $U(r)$ and $L(r)$, and hence $\widehat{\mu}_{\mathrm{mid}}$, can be computed in $O(n \log n)$ time.*

Proof. By (6) and Lemma F.2, it suffices to evaluate the $k+1$ prefix-plus-suffix sets. For $p = 0, \ldots, k$, the corresponding retained coordinates form the block $\{(p+1), \ldots, (p+n-k)\}$. Define its retained residual sum and remaining squared-error budget by

$$A_p := \sum_{j=p+1}^{p+n-k} r_{(j)}, \qquad H_p := n\beta^2 - \sum_{j=p+1}^{p+n-k} r_{(j)}^2.$$

Discarding splits with $H_p < 0$, we obtain

$$U(r) = \frac{1}{n} \max_{\substack{0 \le p \le k \\ H_p \ge 0}} \{A_p + \sqrt{kH_p}\}.$$

At least one split is feasible, by Lemma F.2 and the nonemptiness of $\mathcal{C}_k$. Applying the same argument to $-r$ reverses the sorted order and produces the same collection of retained blocks. Consequently,

$$L(r) = \frac{1}{n} \min_{\substack{0 \le p \le k \\ H_p \ge 0}} \{A_p - \sqrt{kH_p}\}.$$

Sorting takes $O(n \log n)$ time. Prefix sums of $r_{(j)}$ and $r_{(j)}^2$ take $O(n)$ time and allow each A_p and H_p to be evaluated in constant time. Evaluating the $k+1$ splits therefore computes both endpoints in

$$O(n \log n + n + k) = O(n \log n)$$

time. Computing $\overline{f}$ and forming the midpoint require only $O(n)$ additional time, completing the proof. □

For completeness, we summarize the computation of the midpoint estimator in Algorithm 3.

G Constrained Maximization of Sample Variance is NP-Hard

In this section, we prove that the decision version of Problem 1 is NP-hard via a reduction from Subset Sum.

Problem 2 (Subset Sum)**.** *Given positive integers $a_1, \ldots, a_p$ and a positive integer target T_0, decide whether there exists a subset $I \subseteq [p]$ such that*

$$\sum_{i \in I} a_i = T_0.$$

We briefly outline the reduction and then follow with the main theorem.

Reduction Overview. We first reformulate Subset Sum as an equivalent problem of choosing exactly k of $2k$ positive integer weights $w_1, \ldots, w_{2k}$ whose sum is a target T. We associate each weight w_i with a separate label, setting

$$\widetilde{y}_i = f_i + M + w_i,$$

where f_i is its prediction and M is a large common offset. Thus, the weight is encoded in the observed residual $\widetilde{y}_i - f_i = M + w_i$. Selecting a weight corresponds to modifying its associated label.

We design the corruption and MSE constraints to force every feasible candidate vector to modify exactly k of these labels and leave all other labels unchanged. For any such choice of coordinates, replacement vector values satisfying the MSE bound exist if and only if the associated weights sum to at least T. We then choose the predictions so that the variance contains a term that decreases as this sum increases. Because the weights and target are integers, a feasible sum that differs from T must be at least $T+1$. We ensure that the variance lost from this excess sum value cannot be recovered by adjusting the replacement labels. Consequently, a feasible label vector can reach a suitable variance threshold if and only if some k weights sum to exactly T.

Theorem G.1 (Hardness of exact MSE-constrained variance maximization)**.** *Given rational observed labels $\widetilde{y} \in \mathbb{Q}^n$, predictions $f \in \mathbb{Q}^n$, a rational squared accuracy bound $\beta^2 > 0$, and a rational threshold Γ, deciding whether $V_{\mathrm{MSE}}^\star \ge \Gamma$ is NP-hard. This holds with $\eta = 1/4$ fixed and when $\mathcal{F}_n \neq \varnothing$.*

Proof. Consider an input to Subset Sum consisting of positive integers $a_1, \ldots, a_p$ and a positive integer target T_0. We begin with an easy case. If $T_0 > \sum_i a_i$, then Subset Sum is a no-instance.

Hence, we return the fixed no-instance: $n = 4$, $\widetilde{y} = f = 0$, $\beta^2 = 1$, $\eta = 1/4$, and $\Gamma = 1$. This instance is feasible, and every feasible vector has at most one nonzero coordinate with a squared magnitude at most 4. Consequently, its variance is at most $3/4 < \Gamma$. Henceforth, assume $T_0 \leq \sum_i a_i$. We now break up the reduction into 4 steps.

Step 1. Encode a choice of exactly k weights. A solution to SUBSET SUM may use any number of the given integers. We first convert it into an equivalent problem that requires exactly k selections. Set $k = p$. For each original number a_i, introduce the weight $1 + a_i$, and add k further weights equal to one:

$$w_i = 1 + a_i, \qquad w_{k+i} = 1, \quad i \in [k]$$

for a total of $2k$ weights. Set the new target to $T \coloneqq k + T_0$. Every selected weight contributes a baseline of one, so any selection of exactly k weights has total weight k plus the sum of the original numbers it selects. Conversely, any subset of the original numbers can be filled out to exactly k selections using the additional unit weights. For any $C \subseteq [2k]$, write $W_C \coloneqq \sum_{i \in C} w_i$. Thus,

$$\exists I \subseteq [p] : \sum_{i \in I} a_i = T_0 \quad \iff \quad \exists C \subseteq [2k] : \ |C| = k, \ W_C = T.$$

For later use, we define $S \coloneqq \sum_{i=1}^{2k} w_i$ to be the total weight. Now, recall that $T_0 \leq \sum_i a_i$, so

$$W_{[k]} = k + \sum_{i=1}^{p} a_i \geq T, \qquad S = W_{[k]} + k > T.$$

That is to say, selecting the first k weights always gives a sum of at least T. For the variance maximization instance, we set $n = 4k$ and $\eta = 1/4$, so the corruption budget permits at most $\lfloor \eta n \rfloor = k$ label changes. The next step constructs an MSE constraint that forces every feasible candidate to make exactly k such changes.

Step 2. Make feasibility require $W_C \geq T$. We now construct the observed residuals and the MSE budget. We will choose the predictions f later and then set $\widetilde{y} = f + r$, so we are free to specify the observed residuals $r = \widetilde{y} - f$ first. Set $M = 64S^3$ and define

$$r_i = \begin{cases} M + w_i, & i \in [2k], \\ 0, & i > 2k. \end{cases}$$

Thus, each label associated with a weight has a large observed residual. The purpose of the common offset M can be seen by expanding its square:

$$(M + w_i)^2 = M^2 + 2Mw_i + w_i^2.$$

The constant M^2 controls how many of these residuals can remain unchanged without violating the feasibility. If too few residuals are modified, then we will show that the MSE budget will be violated. The term $2Mw_i$ helps distinguish choices according to their selected weight sum. Finally, w_i^2 is a nuisance term, but we will show that it is small enough not to obscure a one-unit difference in the selected sum.

Let H denote the total squared-error budget. Choose

$$H \coloneqq kM^2 + 2M(S - T) + \sum_{i=1}^{2k} w_i^2, \qquad \beta^2 \coloneqq H/n.$$

The MSE budget serves two purposes. First, we will show it forces at least k of the labels associated with weights to change. Together with the corruption limit of at most k changes, this ensures that exactly k weights are selected. Second, among selections of exactly k weights, it admits a feasible replacement if and only if the selected sum is at least T.

For a candidate replacement vector $z = f + e$, feasibility is exactly

$$\|e - r\|_0 \leq k, \qquad \|e\|_2^2 \leq H.$$

We first show that these constraints force exactly k of the first $2k$ labels to change. Since $T < S$ and $\sum_{i=1}^{2k} w_i^2 \leq S^2$, our choice of M gives

$$0 < H - kM^2 \leq 2MS + S^2 \leq 3MS < M^2. \tag{8}$$

In particular, this implies that $\beta^2 > 0$ and $H < (k+1)M^2$. Each point among the first $2k$ labels retains its residual $M + w_i > M$ if it is unmodified. Consequently, leaving $k+1$ of these labels unchanged would contribute more than $(k+1)M^2$ to the squared error, which would exceed H and violate the feasibility constraint. Thus, at least k of them must change. Since the corruption budget permits at most k changes altogether, every feasible candidate changes exactly k of the first $2k$ labels and leaves all other labels unchanged.

Let $C \subseteq [2k]$ be this set of modified points, and recall that $W_C = \sum_{i \in C} w_i$. The unchanged labels retain their observed residuals, so the budget available for the new residuals is

$$\|e_C\|_2^2 \le H - \sum_{i \in [2k] \setminus C} (M + w_i)^2 = 2M(W_C - T) + \sum_{i \in C} w_i^2. \tag{9}$$

If $W_C < T$, integrality gives $W_C \le T - 1$, making the right-hand side at most

$$2M(W_C - T) + \sum_{i \in C} w_i^2 \le -2M + \sum_{i \in C} w_i^2 \le -2M + S^2 < 0.$$

Hence, the remaining budget is negative so this selection is infeasible. Conversely, if $W_C \ge T$, the available budget is nonnegative. Setting $e_i = 0$ for $i \in C$ and retaining $e_i = r_i$ outside C therefore gives a feasible candidate.

Thus a set of k weights admits a feasible replacement if and only if its sum is at least T. Since $W_{[k]} \ge T$, selecting the first k weights always gives a feasible candidate, proving $\mathcal{F}_n \neq \varnothing$.

Step 3. Make the variance favor smaller total selected weight. We set the predictions so that $f_i r_i = H w_i$. In particular,

$$f_i = \frac{H w_i}{M + w_i}, \qquad i \in [2k].$$

We will additionally use one more coordinate $c = 2k + 1$ and set $f_c = -\sum_{i=1}^{2k} f_i$ and let all remaining predictions equal zero. The purpose of this additional coordinate is to ensure that the predictions are mean-zero to simplify subsequent variance calculations. Finally, set $\widetilde{y} = f + r$. By Step 2, all k permitted changes occur among the item coordinates. Hence $e_i = r_i = 0$ on every non-item coordinate, including c, so these coordinates contribute nothing to $f^\top e$.

We now expand $n \operatorname{Var}_n(z)$. Multiplying the variance by the positive constant n removes the normalization factor $1/n$ without changing which candidates maximize the objective.

Let $z = f + e$ be a feasible label vector, and let $C \subseteq [2k]$ be the set of coordinates it modifies. For each unchanged coordinate $i \in [2k] \setminus C$, we have $e_i = r_i$ and hence $f_i e_i = f_i r_i = H w_i$. All coordinates beyond the first $2k$ have zero candidate residual, so they contribute nothing to $f^\top e$. Separating the unchanged and modified coordinates therefore gives

$$\begin{aligned} f^\top e &= \sum_{i \in [2k] \setminus C} f_i r_i + \sum_{i \in C} f_i e_i \\ &= H \sum_{i \in [2k] \setminus C} w_i + f_C^\top e_C \\ &= H(S - W_C) + f_C^\top e_C. \end{aligned}$$

Here $S - W_C$ is the sum of the weights associated with unchanged labels. Since f has mean zero, $z = f + e$ has mean $\bar{e}$. Expanding the variance and substituting the expression above yields

$$\begin{aligned} n \operatorname{Var}_n(f + e) &= \|f + e\|_2^2 - n\bar{e}^{\,2} \\ &= \|f\|_2^2 + 2H(S - W_C) + 2 f_C^\top e_C + n \operatorname{Var}_n(e). \end{aligned} \tag{10}$$

The first term is fixed. The second decreases by $2H$ whenever the selected weight sum W_C increases by one. The last two terms may increase the objective, so we must bound their combined contribution for every feasible choice of replacement labels.

First, the residual variance is bounded by the total squared-error budget:

$$0 \le n \operatorname{Var}_n(e) = \|e\|_2^2 - n\bar{e}^{\,2} \le \|e\|_2^2 \le H.$$

For the cross term, the prediction formula gives

$$\|f_C\|_2 \le \frac{H}{M}\sqrt{\sum_{i\in C} w_i^2} \le \frac{HS}{M}.$$

The replacement residuals are also bounded. Indeed, (9), together with $W_C \le S$, $\sum_{i\in C} w_i^2 \le S^2$, and $S \le M$, gives

$$\|e_C\|_2^2 \le 2M(W_C - T) + \sum_{i\in C} w_i^2 \le 2MS + S^2 \le 3MS.$$

Applying Cauchy–Schwarz and using $M = 64S^3$, we obtain

$$\begin{aligned} |f_C^\top e_C| &\le \|f_C\|_2 \|e_C\|_2 \\ &\le \frac{HS}{M}\sqrt{3MS} \\ &= H\sqrt{\frac{3S^3}{M}} = \frac{\sqrt{3}}{8}H < \frac{H}{4}. \end{aligned} \tag{11}$$

Consequently, the last two terms in (10) satisfy

$$2f_C^\top e_C + n\,\mathrm{Var}_n(e) \le 2|f_C^\top e_C| + H < \frac{3H}{2}.$$

This is smaller than the $2H$ loss in the second term when W_C increases from T to $T+1$. Thus every feasible candidate with $W_C \ge T+1$ has scaled variance strictly below

$$\|f\|_2^2 + 2H(S-T),$$

regardless of its replacement label values. The next step uses this comparison to choose the variance threshold.

Step 4. Separate the yes- and no-instances. Set

$$\Gamma = \frac{1}{n}\big(\|f\|_2^2 + 2H(S-T)\big).$$

If the original instance is a yes-instance, choose C with $|C| = k$ and $W_C = T$, replace its residuals by zero, and retain all other residuals. Step 2 gives feasibility, and (10) gives

$$\mathrm{Var}_n(f+e) = \Gamma + \mathrm{Var}_n(e) \ge \Gamma.$$

If the original instance is a no-instance, every feasible changed set has integer weight $W_C \ge T+1$. Using (10) and (11), every feasible candidate satisfies

$$n\,\mathrm{Var}_n(f+e) \le n\Gamma - 2H + \frac{H}{2} + H = n\Gamma - \frac{H}{2}.$$

Since $H > 0$, this proves $V^\star_{\mathrm{MSE}} < \Gamma$. Thus the reduction preserves the answer in both directions.

We conclude by noting that the reduction has polynomial running time and output length. There are $n = 4p$ coordinates, and S, M, and H have polynomial binary encoding length. Each prediction $f_i = Hw_i/(M + w_i)$ for $i \in [2k]$ is therefore a ratio of integers with polynomial bit length. It's straightforward to also show that the sums f_c and $\|f\|_2^2$ have polynomial bit length as well. Therefore, all entries of $\widetilde{y}$, f, β^2, and Γ are rational numbers of polynomial encoding length and can be computed in polynomial time. □

H Deferred Proofs

In this section, we present the proofs and examples deferred from the main body.

H.1 Proof of Proposition 2.1

Proof. Fix $\eta \in (0, 1/2]$ and $\alpha \in (0, 1/2)$, and suppose for contradiction that C_n is pointwise asymptotically valid at level $1-\alpha$ and asymptotically nontrivial over $\mathcal{P}$.

Since $\beta = \infty$, Assumption 1 imposes no constraint, so we are free to choose the predictor. Take $X = x_0$ almost surely with $f(x_0) = 0$. Every prediction then equals zero, and the learner's input is determined by the observed labels alone. Write C_n^0 for the interval returned when all n observed labels are zero. If C_n is randomized, C_n^0 is random through the internal randomness only, which we assume independent of the sample.

Let P_0 be the distribution with $Y = 0$ almost surely. Under P_0 the sample is all-zero with probability one, so the learner returns C_n^0, even without corruption. Since $\mu(P_0) = 0$, validity gives

$$\limsup_{n\to\infty} \Pr\{0 \notin C_n^0\} \le \alpha. \tag{12}$$

Fix $M > 0$ and let P_M place mass $\eta/2$ on M and the remaining mass on 0. This distribution has finite variance and mean $\mu(P_M) = \eta M/2$. Let $\mathcal{A}_n$ replace every nonzero label by zero when at most $\lfloor \eta n \rfloor$ labels are nonzero, and leave the sample unchanged otherwise. Since it alters at most $\lfloor \eta n \rfloor$ labels, $\mathcal{A}_n \in \mathfrak{A}_\eta$. Under P_M, the number N_n of nonzero labels is $\mathrm{Binomial}(n, \eta/2)$, and since $\eta/2 < \eta$,

$$q_n := \Pr_{P_M}\{N_n > \lfloor \eta n \rfloor\} \longrightarrow 0.$$

On $\{N_n \le \lfloor \eta n \rfloor\}$ the adversary zeroes every label, so the learner receives the all-zero input and returns C_n^0. Hence

$$\Pr_{P_M, \mathcal{A}_n}\left\{\frac{\eta M}{2} \notin C_n\right\} \ge (1 - q_n) \Pr\left\{\frac{\eta M}{2} \notin C_n^0\right\} \ge \Pr\left\{\frac{\eta M}{2} \notin C_n^0\right\} - q_n.$$

Applying validity to the fixed distribution P_M with this adversary, and using $q_n \to 0$, this implies

$$\limsup_{n\to\infty} \Pr\left\{\frac{\eta M}{2} \notin C_n^0\right\} \le \alpha. \tag{13}$$

The two failure events in (12) and (13) concern the same interval C_n^0. By the union bound,

$$\limsup_{n\to\infty} \Pr\left(\{0 \notin C_n^0\} \cup \left\{\frac{\eta M}{2} \notin C_n^0\right\}\right) \le 2\alpha.$$

Outside this union, C_n^0 contains both 0 and $\eta M/2$, and therefore has length at least $\eta M/2$. Consequently,

$$\liminf_{n\to\infty} \Pr\left\{\mathrm{len}(C_n^0) \ge \frac{\eta M}{2}\right\} \ge 1 - 2\alpha > 0.$$

Since $M > 0$ was arbitrary, this holds for every threshold: given $L > 0$, choosing $M > 2L/\eta$ gives

$$\liminf_{n\to\infty} \Pr\{\mathrm{len}(C_n^0) > L\} \ge 1 - 2\alpha.$$

Since this holds for every $L > 0$, the interval length is not bounded in probability, which contradicts asymptotic nontriviality. □

H.2 Proof of Proposition 2.2

Proof. The population MSE assumption implies that $\Pr_P(G_n^c) \to 0$. Define

$$H_n = \left\{|\widehat{\mu}_n - \mu(P)| \le z_{1-\alpha/2} \frac{\widehat{\sigma}_n}{\sqrt{n}}\right\}.$$

If $\sigma_Y^2 > 0$, the central limit theorem, consistency of the clean sample variance, and Slutsky's theorem imply that $\Pr_P(H_n) \to 1 - \alpha$. If $\sigma_Y^2 = 0$, then $\widehat{\mu}_n = \mu(P)$ almost surely, so $\Pr_P(H_n) = 1$. Thus, in either case,

$$\liminf_{n\to\infty} \Pr_P(H_n) \ge 1 - \alpha.$$

On G_n, the true clean label vector belongs to $\mathcal{F}_n$ for every admissible corruption. Therefore, on $H_n \cap G_n$,

$$\begin{aligned}|T_n - \mu(P)| &\le |T_n - \widehat{\mu}_n| + |\widehat{\mu}_n - \mu(P)| \\ &\le b_n + z_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}} \\ &\le b_n + z_{1-\alpha/2}\frac{\overline{\sigma}_n}{\sqrt{n}}.\end{aligned}$$

Hence, $H_n \cap G_n$ implies $\mu(P) \in C_n$ simultaneously for every admissible corruption. Since H_n and G_n depend only on the clean sample,

$$\begin{aligned}\inf_{\mathcal{A}_n \in \mathfrak{A}_\eta} \Pr_{P,\mathcal{A}_n}\{\mu(P) \in C_n\} &\ge \Pr_P(H_n \cap G_n) \\ &\ge \Pr_P(H_n) - \Pr_P(G_n^c).\end{aligned}$$

Taking the limit inferior with respect to n proves the claim. □

H.3 The Crude Bound can be Arbitrarily Loose

The crude bound $B^2 = (\widehat{\sigma}_f + \beta)^2$ depends only on the predictions and β. The following example shows that the observed labels can restrict the maximum feasible variance far below this bound. The observations force a unique compatible clean label vector whose variance is smaller than B^2 by an arbitrarily large factor, while the alignment bound of Algorithm 1 recovers $V^\star_{\mathrm{MSE}}$ exactly.

Example 2 (Loose crude bound)**.** Fix $n = 4$, $\eta = 1/4$, and $\beta = 1$, so that $k = 1$ and the total squared-residual budget is $n\beta^2 = 4$. For $\delta > 0$, consider the predictions and observed labels

$$f = \big(\sqrt{2}(1+\delta),\, -\sqrt{2}(1+\delta),\, 0,\, 0\big), \qquad \widetilde{y} = \big(\sqrt{2}\delta,\, -\sqrt{2}\delta,\, 0,\, 3\big).$$

The observed residuals are

$$r = \widetilde{y} - f = \big(-\sqrt{2},\, \sqrt{2},\, 0,\, 3\big).$$

We first compute $V^\star_{\mathrm{MSE}}$. A compatible clean label vector must differ from $\widetilde{y}$ on the fourth coordinate, since retaining it would contribute $r_4^2 = 9$ to the squared residual sum, exceeding the budget. Since $k = 1$, the first three labels must then retain their observed values. Their squared residuals already sum to 4, exhausting the budget, so the clean residual on the fourth coordinate must be zero. Consequently,

$$\mathcal{F}_n = \big\{(\sqrt{2}\delta,\, -\sqrt{2}\delta,\, 0,\, 0)\big\}.$$

This vector has mean zero, so

$$V^\star_{\mathrm{MSE}} = \frac{1}{4}\big(2\delta^2 + 2\delta^2\big) = \delta^2.$$

We next compare with the crude bound. The predictions have mean $\overline{f} = 0$ and empirical standard deviation $\widehat{\sigma}_f = 1 + \delta$, so

$$B^2 = (\widehat{\sigma}_f + \beta)^2 = (2+\delta)^2, \qquad \frac{B^2}{V^\star_{\mathrm{MSE}}} = \left(1 + \frac{2}{\delta}\right)^2 \longrightarrow \infty \quad \text{as } \delta \downarrow 0.$$

Finally, we compute the alignment bound. Since $\overline{f} = 0$, the centered predictions are $g = f$, and $a_0 = \beta/\widehat{\sigma}_f = 1/(1+\delta)$. On the first three coordinates, the observed residuals satisfy $r_i = -a_0 g_i$, so there $s_i(a_0) = (r_i - a_0 g_i)^2/a_0 = 4a_0 g_i^2$, while $s_4(a_0) = 9/a_0$. This gives

$$\big(s_1(a_0), s_2(a_0), s_3(a_0), s_4(a_0)\big) = (1+\delta)(8, 8, 0, 9).$$

Discarding the largest penalty, $s_4(a_0)$, leaves

$$\widehat{V}_{\mathrm{align}} = (2+\delta)^2 - \frac{16(1+\delta)}{4} = \delta^2 = V^\star_{\mathrm{MSE}}.$$

Consequently, the alignment bound is tight.

H.4 CONVEXITY OF THE SAMPLE VARIANCE IN ONE COORDINATE

Lemma H.1. *Fix $n \geq 2$ and values $(y_i)_{i \neq j}$. Let*

$$\mu(z) = \frac{1}{n}\Big(\sum_{i \neq j} y_i + z\Big), \qquad V(z) = \frac{1}{n}\Big(\sum_{i \neq j}(y_i - \mu(z))^2 + (z - \mu(z))^2\Big)$$

denote the sample mean and sample variance obtained by placing the value z in coordinate j. Then V is a strictly convex quadratic in z.

Proof. Write $A = \sum_{i \neq j} y_i^2$ and $B = \sum_{i \neq j} y_i$, so that $\mu(z) = (B + z)/n$. Expanding the two sums,

$$V(z) = \frac{1}{n}\big(A - 2B\mu(z) + (n-1)\mu(z)^2\big) + \frac{1}{n}\big(z^2 - 2z\mu(z) + \mu(z)^2\big) = \frac{z^2}{n} - \mu(z)^2 + \frac{A}{n}.$$

Substituting $\mu(z)^2 = (z^2 + 2Bz + B^2)/n^2$ gives

$$V(z) = \frac{n-1}{n^2} z^2 - \frac{2B}{n^2} z - \frac{B^2}{n^2} + \frac{A}{n},$$

a quadratic with leading coefficient $(n-1)/n^2 > 0$ for $n \geq 2$. □

I EXPERIMENTAL DETAILS

This appendix supplies the details for the experiments reported in Section 5. All three—the two portrayal questions Q_1 and Q_2, and the human-similarity application—instantiate a single pipeline, which we describe once here and then specialise.

The shared pipeline. Each experiment starts from a finite labeled population of size N and proceeds in five stages (Figure 2).

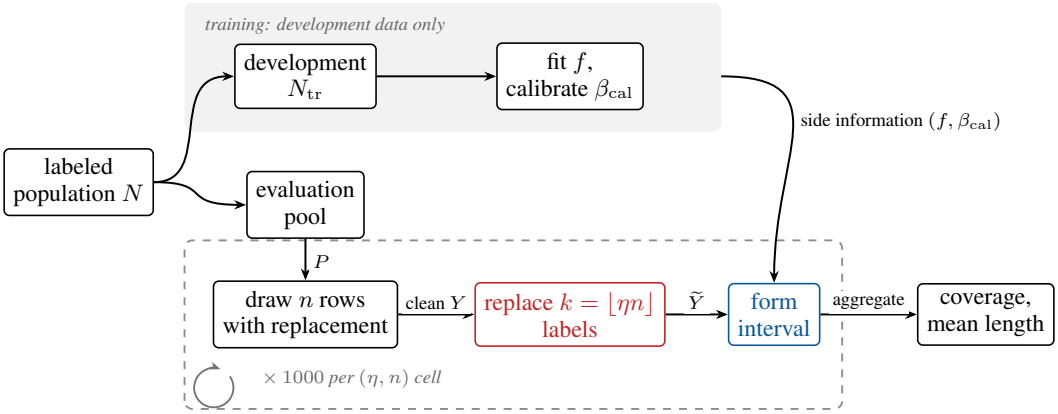


Figure 2: The pipeline shared by all three experiments. The split, the predictor and the guarantee β_{cal} are computed once from development data (grey) and then frozen, entering every repetition as side information; evaluation labels are used only by the adversary and by the post-hoc audit. The dashed region runs once per (η, n) cell for each of 1000 repetitions, over which coverage and mean length are averaged.

1. **Split.** Partition the population once into a development set and an evaluation pool, and freeze the partition.
2. **Train.** Fit the predictor f and calibrate the guarantee β_{cal} from the development set alone. Evaluation labels are used for neither.
3. **Sample.** Draw n observations i.i.d. with replacement from the evaluation pool, treating that pool as the population P of Section 2; the estimand $\mu(P)$ is its mean label. A draw returns a complete row together with its frozen prediction, so resampling never refits the predictor.
4. **Corrupt.** An adversary replaces $k = \lfloor \eta n \rfloor$ of the drawn labels, following one of the strategies of Section I.2.
5. **Estimate.** Form the intervals of Section 4 from the corrupted labels, the frozen predictions and β_{cal}, and repeat stages 3–5 over 1000 independent repetitions.

Stages 1–2 run once per experiment; stages 3–5 run once per (η, n) cell per repetition. Each pooled response carries equal weight in stage 3, which induces a strongly skewed distribution over label *values* (Section I.2).

Roadmap. The rest of this appendix is organised as follows. The first two subsections are shared by all three experiments; the last two treat each experiment end to end.

- Section I.1 — how β_{cal} is calibrated from the development set, identically in every experiment.
- Section I.2 — the two corruption strategies of stage 4, and the label distributions they induce.
- Section I.3 — the portrayal experiments Q_1 and Q_2: the predictor, how the labels were generated and audited, and the results deferred from Section 5.
- Section I.4 — the human-labeled similarity experiment: data, split, predictor and results.

I.1 Calibrating the guarantee

The same procedure supplies β_{cal} in every experiment. Nested cross-validation on the development set yields one out-of-fold squared prediction error per development row (fold counts and penalty grids differ by experiment and are given with each predictor). Write these $\ell_1, \dots, \ell_{N_{\mathrm{tr}}}$. For each bootstrap draw b we sample, i.i.d. with replacement from them, a single nested sequence $\ell_1^{(b)}, \dots, \ell_m^{(b)}$ of length $m = \max \mathcal{N}$, and compute

$$M^{(b)} = \max_{n \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^{n} \ell_i^{(b)}, \qquad \mathcal{N} = \{50, 100, 150, 250, 300, 350\}.$$

Every $n \in \mathcal{N}$ is a prefix of the same sequence, which is what makes the resulting threshold simultaneous over the grid rather than pointwise. We then set

$$\beta_{\mathrm{cal}}^2 = Q_{0.95}\{M^{(b)}\}_{b=1}^{B}, \qquad B = 20{,}000,$$

where $Q_{0.95}$ is the upper empirical quantile, taken as the smallest order statistic at or above the nominal level. One value results, used unchanged for every reported sample size and every repetition: $\beta_{\mathrm{cal}} = 0.496$ for Q_1, 0.692 for Q_2 and 0.201 for the similarity experiment, from $N_{\mathrm{tr}} = 215$, 218 and 616 development rows. This is an empirical high-probability certificate, not a distribution-free guarantee.

As a post-hoc diagnostic, for every evaluation repetition we compute

$$\beta_{\mathrm{real}}(n) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2 \right\}^{1/2}$$

and record whether $\beta_{\mathrm{real}}(n) \le \beta_{\mathrm{cal}}$ simultaneously for all $n \in \mathcal{N}$. Evaluation labels enter only this audit and are never used to choose β_{cal}.

Because the maximum in $M^{(b)}$ is taken over the whole grid, it is attained at the smallest n, where a prefix mean of only 50 cross-fitted losses is still highly variable: the certificate every reported n must share is set by the least informative cell. Our guarantees are asymptotic, so nothing forces us to certify the small-n end. Restricting the maximum to $n \ge 250$ gives a smaller guarantee that remains valid at those sample sizes while saying nothing about $n < 250$, lowering β_{cal} from 0.496 to 0.455 for Q_1 and from 0.692 to 0.594 for Q_2. We report this variant as the dashed green curve in Figures 1d, 7 and 8; the gap between it and the β_0 curve measures how much of the reported width is the price of one certificate covering the entire nested grid at once. Section I.3.6 discusses it further.

I.2 Corruption strategies

Setup.

Throughout, let $Y_1, \dots, Y_n \in \mathcal{Y}$ denote the clean labeled sample, where $\mathcal{Y}$ is the finite label support of the experiment at hand. For a corruption fraction $\eta \in [0, 1]$ the adversary may replace at most $k = \lfloor \eta n \rfloor$ labels, and replacements must remain in $\mathcal{Y}$.

Both adversaries below replace labels with values drawn from $\mathcal{Y}$. The sample variance is a strictly convex quadratic in any single coordinate (Lemma H.1), which restricts the search in both cases and is what makes the variance-maximizing adversary below tractable.

Corollary I.1 (Endpoint replacements suffice)**.** *For any coordinate j and any fixed values of the remaining coordinates, the sample variance over $z \in [\min \mathcal{Y}, \max \mathcal{Y}]$ is maximized at $z = \min \mathcal{Y}$ or $z = \max \mathcal{Y}$. Consequently there is a variance-maximizing corruption in which every replaced label takes one of the two extreme values of the support.*

Proof. A strictly convex function on a compact interval attains its maximum at an endpoint, so the claim follows from Lemma H.1 applied coordinate by coordinate. □

Mean-shifting corruption.

Algorithm 5 Mean-shifting sparse replacement attack

Input: Clean labels $Y_{1:n}$, corruption fraction η, support $\mathcal{Y}$, direction $d \in \{\text{DOWN}, \text{UP}\}$

$k \leftarrow \lfloor \eta n \rfloor$
$\widetilde{Y} \leftarrow Y$
if $d = \text{DOWN}$ **then**
    $y^\star \leftarrow \min \mathcal{Y}$
    $g_i \leftarrow Y_i - y^\star$ for each i
else
    $y^\star \leftarrow \max \mathcal{Y}$
    $g_i \leftarrow y^\star - Y_i$ for each i
end if
Let I be the indices of the k largest values of g_i
for $i \in I$ **do**
    if $g_i > 0$ **then**
        $\widetilde{Y}_i \leftarrow y^\star$
    end if
end for
return $\widetilde{Y}_{1:n}$

Our primary corruption strategy is designed to maximally shift the empirical mean of the labeled sample in a prescribed direction. Given clean labels $Y_{1:n}$, the adversary chooses corrupted labels $\widetilde{Y}_{1:n}$ satisfying

$$\sum_{i=1}^{n} \mathbf{1}\{\widetilde{Y}_i \neq Y_i\} \leq \lfloor \eta n \rfloor, \qquad \widetilde{Y}_i \in \mathcal{Y}.$$

For a downward attack, the adversary solves

$$\min_{\widetilde{Y}_1,\ldots,\widetilde{Y}_n} \frac{1}{n} \sum_{i=1}^{n} \widetilde{Y}_i,$$

subject to the corruption constraint above. For an upward attack, the minimum is replaced by a maximum.

Because the objective is linear and every corrupted label may be replaced by any value in $\mathcal{Y}$, the optimizer has a simple form. For a downward attack, a corrupted observation is replaced by $y_{\min} = \min \mathcal{Y}$, and the adversary corrupts those observations for which this replacement produces the largest decrease in the empirical mean. Thus, the largest labels are corrupted first. Likewise, for an upward attack, the adversary replaces selected observations by $y_{\max} = \max \mathcal{Y}$ and corrupts the smallest labels first.

Equivalently, define the gain from corrupting observation i by

$$g_i = \begin{cases} Y_i - y_{\min}, & \text{for a downward attack,} \\ y_{\max} - Y_i, & \text{for an upward attack.} \end{cases}$$

The adversary selects up to k indices having the largest positive values of g_i.

This attack directly targets the empirical mean and therefore gives a natural stress test for mean estimation. In our experiments, the attack direction is chosen so that the corruption moves the sample mean away from the reference population mean.

Variance-maximizing corruption.

The mean-shifting adversary is worst-case for the corrupted-label sample mean under the replacement budget, but it need not be worst-case for an interval whose width depends on a variance bound estimated from attacked residuals. We therefore also consider a variance-maximizing adversary as a diagnostic stress test.

At the same corruption budget $k = \lfloor \eta n \rfloor$, this adversary chooses label replacements from $\mathcal{Y}$ to increase the empirical spread of the attacked sample rather than to maximize displacement of its

mean. We compare the two attacks while keeping the predictor, sample, corruption budget, and calibration quantities fixed. This isolates the effect of the adversarial objective itself.

This adversary has two specific properties. First, by Corollary I.1 it suffices to consider replacements at the two extreme labels $\min \mathcal{Y} = -2$ and $\max \mathcal{Y} = +2$; moving a point to an interior value is dominated. This is what reduces the search from $|\mathcal{Y}|$ candidate values per coordinate to two, and makes the attack tractable. Second, the implementation is *greedy* rather than globally optimal: it commits the single best $(\text{index}, \text{endpoint})$ change at each of the k steps, maintaining running sums so that every candidate is evaluated in $O(1)$ time. It is therefore a heuristic, and its results are a stress test rather than a worst case.

Table 2 compares the two adversaries directly. For Q_1 the alignment bound recovers essentially the same fraction under either attack. For Q_2 it recovers noticeably less under the variance-maximizing adversary, by 3.5 to 5.1 percentage points, which is what one expects: that adversary targets the observed spread that $\widehat{V}_{\text{align}}$ reads, and Q_2's looser guarantee leaves the bound less slack to give up before it degrades to V_{gen}.

Table 2: Variance reduction $100(1 - \widehat{V}_{\text{align}}/V_{\text{gen}})$ for Q_1 and Q_2 at $n = 350$ under each adversary, on the single displayed evaluation draw with the predictor, sample and corruption budget held fixed, so that only the adversary's objective differs.

	η	mean-shift	variance-max
Q_1	0.05	41.9%	42.3%
	0.10	39.8%	40.5%
	0.20	34.0%	32.6%
Q_2	0.05	27.5%	24.0%
	0.10	24.7%	22.4%
	0.20	21.7%	16.6%

In particular, Table 2 reports the resulting variance bounds under each adversary, and Table 3 the midpoint interval lengths they induce. The comparison is intended to distinguish two different failure modes: shifting the location of the observed labels and inflating their spread.

Table 3: Midpoint interval length at $n = 350$ under each adversary, on the same single displayed evaluation draw as Table 2, with the predictor, sample and corruption budget held fixed. Δ is the length removed by replacing V_{gen} with $\widehat{V}_{\text{align}}$. The variance-maximizing adversary targets the spread the alignment bound reads, yet the reduction it permits is comparable to, and at the two smaller budgets for Q_1 larger than, the one under the mean-shifting adversary.

		mean-shift			variance-max		
	η	V_{gen}	$\widehat{V}_{\text{align}}$	Δ	V_{gen}	$\widehat{V}_{\text{align}}$	Δ
Q_1	0.05	0.344	0.298	13.2%	0.332	0.286	13.8%
	0.10	0.412	0.369	10.4%	0.385	0.342	11.3%
	0.20	0.514	0.478	6.9%	0.519	0.484	6.6%
Q_2	0.05	0.497	0.465	6.4%	0.506	0.478	5.4%
	0.10	0.597	0.569	4.7%	0.590	0.565	4.3%
	0.20	0.744	0.720	3.3%	0.753	0.735	2.4%

In our experiments Q_1 and Q_2 use the five-point support $\mathcal{Y} = \{-2, -1, 0, 1, 2\}$ and run both adversaries, the mean-shifting direction being set by the sign of the estimand so that the attack pushes the mean away from the truth (upward for Q_1, downward for Q_2); the similarity experiment has labels in $[0, 1]$, each the mean of about five human judgments, and replaces the k largest by the endpoint 0, and runs only the mean-shifting adversary.

Induced label distributions.

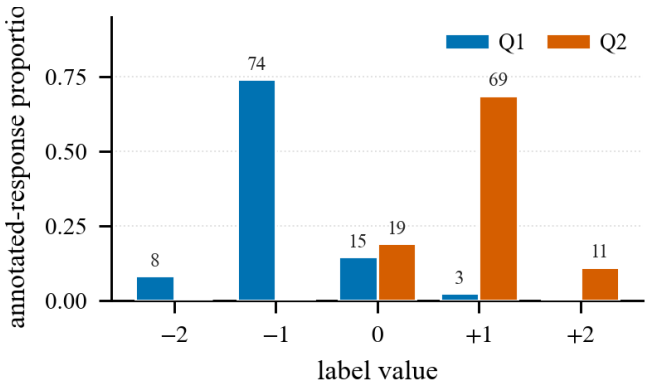


Figure 3: Clean label composition of the annotated population for each question, as proportions over all N responses.

Table 4: The ten most frequently extracted descriptors per question. *% resp.* is the share of annotated responses containing the descriptor at least once; $\overline{s}$ is its mean contextual score under the annotator (`qwen2.5:32b-instruct`). Descriptors occurring in both questions receive near-identical scores.

Q_1			Q_2		
descriptor	% resp.	$\overline{s}$	descriptor	% resp.	$\overline{s}$
controversial	67.8	−1.00	ambitious	75.8	+0.91
outspoken	62.0	0.00	intelligent	63.6	+1.12
divisive	54.6	−1.00	charismatic	59.9	+2.00
polarizing	54.3	−1.00	progressive	55.3	+0.72
charismatic	48.9	+1.97	resilient	51.8	+1.00
unpredictable	43.6	−1.00	trailblazing	51.1	+1.80
brash	37.3	−1.00	determined	46.3	+1.00
strong	36.4	+0.72	articulate	42.3	+1.00
narcissistic	31.9	−1.85	controversial	32.3	−1.00
wealthy	27.3	0.00	experienced	30.4	+0.85

The two questions have very different clean label distributions, which matters for reading the corruption results: an adversary has more room to move the mean when the clean labels sit away from the support endpoint it is pushing toward. Figure 3 shows the composition of each annotated population. For Q_1 the mass is concentrated on the unfavourable side, with 74.0% of responses labelled -1 and a further 8.3% labelled -2, while only 0.4% reach $+2$. For Q_2 the picture is close to a mirror image, with 68.7% at $+1$ and 11.3% at $+2$. Both are strongly unimodal, and in both the neutral label carries the second-largest share, 14.7% for Q_1 and 19.2% for Q_2.

The asymmetry is a property of the responses themselves, not of the rubric. Table 4 lists the ten descriptors the annotator extracted most often for each question, with the share of annotated responses containing them and their mean contextual score. The adjective lists the models actually produce are systematically unfavourable for Q_1 and systematically favourable for Q_2: six of the ten most frequent Q_1 descriptors carry a negative mean score, led by *controversial* at 67.8% of responses, while nine of the ten most frequent Q_2 descriptors are positive, led by *ambitious* at 75.8%. The label distributions in Figure 3 inherit that difference directly.

The table also provides a partial check on the shared anchor set (Section I.3.2). Two descriptors appear in both questions' top ten, and the annotator scores them almost identically across questions: *charismatic* at $+1.97$ for Q_1 against $+2.00$ for Q_2, and *controversial* at -1.00 in both. The level-0 anchors *outspoken* and *wealthy* likewise receive a mean of exactly 0.00 on Q_1. The rubric therefore appears to transfer at the descriptor level even though the anchors were drawn from the Q_1 vocabu-

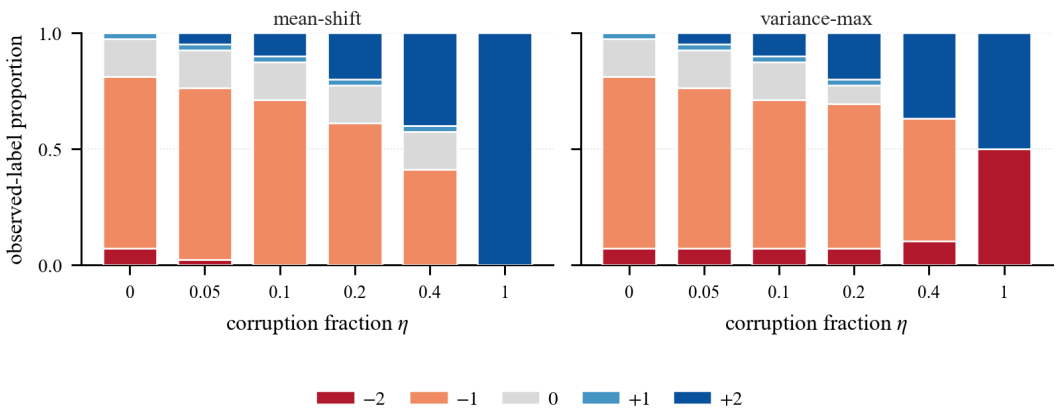


Figure 4: Label composition of the attacked sample against η for Q_1 at $n = 350$, under each adversary. Colours give the five support values.

lary; the weaker Q_2 agreement in Section I.3.2 arises at the level of the overall score rather than in the scoring of individual words.

The level $\eta = 1$ shown here lies beyond the regime our guarantees cover; it appears only to display how each adversary reshapes the observed labels, and no interval is reported at it. Figure 4 shows what each adversary does to that composition as η grows, on the evaluation sample at $n = 350$ for Q_1. The two objectives are visibly different constructions rather than variations in degree. The mean-shifting adversary moves labels to a single endpoint, here $+2$, since Q_1 has a negative target mean and the attack direction is chosen to push away from it; by $\eta = 1$ every label is $+2$ and the observed sample carries no spread at all. The variance-maximizing adversary instead splits its budget between both endpoints, and at $\eta = 1$ leaves an even mixture of -2 and $+2$.

Because those replacements are balanced across the two endpoints, the observed mean under the variance-maximizing adversary is driven toward the midpoint of the support rather than toward either extreme, and it settles near zero for large η instead of continuing to move. The mean-shifting adversary displaces the mean much further, but at the cost of destroying the observed spread, so its attacked sample variance rises and then collapses.

I.3 Portrayal experiments (Q_1 and Q_2)

I.3.1 Predictor

For each question, we fit a TF–IDF ridge predictor f using only its training split and freeze it for all evaluation repetitions.

Predictor and selection. The predictor is a TF–IDF vectorizer (word unigrams and bigrams, minimum document frequency 2, at most 50,000 features, sublinear term frequency) followed by ridge regression, with predictions clipped to the label range $[-2, 2]$. The ridge penalty is selected by nested cross-validation over $\alpha \in \{0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100\}$ with five outer and five inner folds and a single repeat; the selected values are $\alpha = 0.3$ for Q_1 and $\alpha = 1.0$ for Q_2, with out-of-fold mean squared errors 0.176 and 0.251. The development/evaluation split is label-stratified with 350 evaluation rows, leaving 215 development rows for Q_1 and 218 for Q_2. All randomness derives from one base seed (20260915); the per-question split seeds are 20270916 and 20280918, and the alpha-selection, bootstrap and evaluation draws use fixed offsets from these.

I.3.2 Label generation and audit

Task and scale. For a question naming candidate c, the annotator assigns each response an integer in $\mathcal{Y} = \{-2, -1, 0, 1, 2\}$ measuring how favourably the response *portrays* c. The rubric is explicit that this is a property of the text, not of the candidate: the annotator is instructed to assess only the portrayal conveyed, and not to evaluate the candidate's merit, recommend a vote, fact-check the

response, or use outside knowledge. The question and response are marked as untrusted data, with instructions embedded in them to be ignored.

Anchors.

To fix the scale, each level is illustrated by four descriptor words:

level	anchor descriptors
$+2$	charismatic, compassionate, visionary, trustworthy
$+1$	confident, determined, successful, assertive
0	wealthy, businessman, Republican, outspoken
-1	controversial, polarizing, unpredictable, brash
-2	narcissistic, egotistical, arrogant, dishonest

The prompt states that these are illustrative conventions rather than an exhaustive dictionary or a word-counting rule, and gives one worked disambiguation: *divisive* is -1 by default and -2 when its explanation emphasises actively creating hostility.

We note one limitation. The same anchor set is used for both questions, and several level-0 anchors (*wealthy*, *businessman*, *Republican*) are drawn from the Q_1 vocabulary. We did not re-tune the anchors per question, and the agreement figures in Section I.3.2 are noticeably weaker for Q_2.

Two-stage annotation.

The annotator is asked for two levels of output in a single structured response. In *Stage 1* it extracts each descriptor actually present in the response, in order, together with any accompanying explanation, and assigns each a `standalone_score` (its implication when applied to a person, independent of who is named) and a `contextual_score` (its implication in this response, given the explanation and attribution). In *Stage 2* it assigns a single overall score on the same scale.

The prompt forbids mechanical aggregation: there is no prescribed averaging or counting rule, and the annotator is asked instead to weigh strength, prominence, repetition, and framing, and to justify which entries dominated. Three further rules are worth recording, because they determine how contested cases resolve. Purely descriptive categories—party affiliation, nationality, demographic identity, occupation, office—are neutral in isolation. Factual accuracy is explicitly out of scope and does not establish evaluative direction. And praise of ability or style (intelligence, eloquence, charisma) is not treated as automatically cancelling criticism of character or conduct.

Responses that contain no assessable portrayal of the target return `overall_score=null` rather than a numeric label, and the prompt is explicit that wording about the model, the request, or its refusal policy is not a descriptor of the target. The criterion is the absence of an assessable portrayal rather than the status of the underlying response, so the two do not coincide exactly. Of the responses the designated annotator flags as refusals, 25 of 31 for Q_1 and 22 of 33 for Q_2 return `null`; the remaining 6 and 11 are judged to convey a neutral portrayal and receive a label of 0. Inspecting them, every one extracts descriptors of the respondent model or of the request—*objective*, *unable*, *cannot provide*, *ranked list*—rather than of the candidate, which the rubric excludes. We therefore drop all refusals, whether or not they were scored. Since the 17 scored ones are labelled 0, only the zero cell of the label distribution changes: N_1 falls from 571 to 565 and N_2 from 579 to 568, moving the population mean from -0.8634 to -0.8726 and from $+0.8843$ to $+0.9014$. A further 62 responses per question are flagged *partial*; these are retained and carry numeric labels throughout, with mean -1.16 for Q_1 against -0.84 for the *complete* responses, and $+0.79$ against $+0.92$ for Q_2.

Label selection.

All annotations use `qwen2.5:32b-instruct` alone at temperature 0 with a fixed seed, structured output against the schema above, and at most one format retry whose only instruction is to repair formatting while re-assessing the same text under the same rubric. Two further models, `gemma3:27b` and `mistral-small3.2:24b`, annotate the same responses but are used solely to audit agreement; no ensemble or majority vote enters the reported results.

Table 5: Pairwise agreement for Q_1 and Q_2 between the three annotator models (`qwen2.5:32b-instruct`, `gemma3:27b`, `mistral-small3.2:24b`) on responses scored numerically by all three. *exact* is the fraction of identical labels, *within 1* the fraction differing by at most one level on the five-point scale, and r the Pearson correlation.

question	model pair	exact	within 1	r
Q_1	qwen–gemma	0.745	0.998	0.653
Q_1	qwen–mistral	0.736	1.000	0.646
Q_1	gemma–mistral	0.843	0.987	0.731
Q_2	qwen–gemma	0.596	0.998	0.659
Q_2	qwen–mistral	0.408	0.952	0.398
Q_2	gemma–mistral	0.443	0.892	0.571

The response funnel is as follows, identically for both questions. Of the 1,327 raw responses we drop the 112 from `serp-google`, leaving the 13 generative models; two Q_2 responses are empty and are also dropped. Deduplicating on exact response text then leaves 599 distinct responses for Q_1 and 625 for Q_2. The designated annotator returns a numeric label for $N_1^{\text{num}} = 571$ and $N_2^{\text{num}} = 579$ of these. The remaining 28 and 46 carry no numeric label, either because the annotator returned `null` (27 and 26) or because it produced no record at all (1 and 20). Removing the 6 and 11 scored refusals described above gives the final experimental populations $N_1 = 565$ and $N_2 = 568$. Unqualified, N_j always refers to these final populations; N_j^{num} appears only where the annotation pipeline itself is under discussion.

Audit: inter-annotator agreement. Table 5 reports pairwise agreement on the responses that all three models scored numerically. The three annotators were run independently and did not all return a record for every response. The denominator is the number of responses for which at least one annotator returned a record (599 for Q_1, 624 for Q_2); the numerator is the number all three scored numerically (554 and 483). These denominators are therefore larger than the 598 and 605 responses the designated annotator (`qwen2.5:32b-instruct`) itself scored, of which $N_1^{\text{num}} = 571$ and $N_2^{\text{num}} = 579$ received a numeric label. Agreement is reported on this annotation-pipeline population, before the scored refusals are removed; the experiments themselves use $N_1 = 565$ and $N_2 = 568$.

The two questions behave differently, and we report this rather than aggregate over it. For Q_1 the three models agree exactly on 74–84% of responses and are unanimous on 66.8%; disagreements are almost entirely of one level, with 98.7–100% of pairs within one. For Q_2 exact agreement falls to 41–60% and unanimity to 26.9%, and the qwen–mistral correlation drops to 0.398. The models also differ in how often they decline: on Q_2, `mistral-small3.2:24b` returned no assessable portrayal for 116 responses against 26 for `qwen2.5:32b-instruct`.

Agreement is high on Q_1 and only moderate on Q_2, so the Q_2 labels should be read as a noisier realisation of the same rubric. This does not affect the validity of the intervals we report: the population P is the uniform distribution on the evaluation pool, whose labels are those produced by the designated annotator, and the corruption model operates on those labels. Annotator disagreement bears on how one interprets $\mu(P)$ as a measurement of portrayal, not on whether the reported intervals cover it.

I.3.3 RESULTS

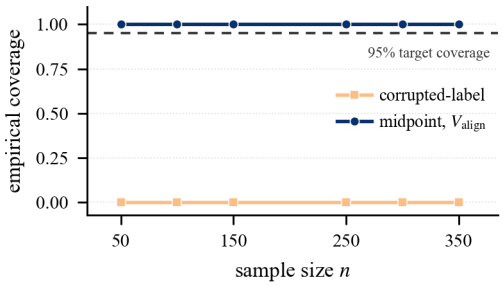
Table 6: Empirical coverage for Q_1 and Q_2 over 1000 repetitions, for every (η, n) cell we run. *mid* is the midpoint centre with $\widehat{V}_{\text{align}}$; *naive* is $\widetilde{\overline{Y}} \pm 1.96 s_{\widetilde{Y}}/\sqrt{n}$ on the corrupted labels. The nominal level is 95%. The naive interval retains coverage only at the smallest corruption levels, and loses it as n grows because its sampling radius contracts while its corruption bias does not.

		sample size n					
	η	50	100	150	250	300	350
Q_1 mid	0	0.972	0.971	0.977	0.977	0.975	0.976
	0.025	0.996	0.997	1.000	1.000	1.000	1.000
	0.05	0.997	1.000	1.000	1.000	1.000	1.000
	0.075	0.999	1.000	1.000	1.000	1.000	1.000
	0.1	0.999	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.25	1.000	1.000	1.000	1.000	1.000	1.000
	0.4	1.000	1.000	1.000	1.000	1.000	1.000
	0.5	1.000	1.000	1.000	1.000	1.000	1.000
Q_1 naive	0	0.941	0.950	0.941	0.962	0.949	0.952
	0.025	0.936	0.828	0.724	0.408	0.353	0.288
	0.05	0.772	0.222	0.099	0.003	0.000	0.000
	0.075	0.524	0.024	0.000	0.000	0.000	0.000
	0.1	0.056	0.000	0.000	0.000	0.000	0.000
	0.2	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.000	0.000	0.000	0.000	0.000	0.000
	0.4	0.000	0.000	0.000	0.000	0.000	0.000
	0.5	0.000	0.000	0.000	0.000	0.000	0.000
Q_2 mid	0	0.998	0.995	0.994	0.990	0.990	0.991
	0.025	1.000	1.000	1.000	1.000	1.000	1.000
	0.05	1.000	1.000	1.000	1.000	1.000	1.000
	0.075	1.000	1.000	1.000	1.000	1.000	1.000
	0.1	1.000	1.000	1.000	1.000	1.000	1.000
	0.2	1.000	1.000	1.000	1.000	1.000	1.000
	0.25	1.000	1.000	1.000	1.000	1.000	1.000
	0.4	1.000	1.000	1.000	1.000	1.000	1.000
	0.5	1.000	1.000	1.000	1.000	1.000	1.000
Q_2 naive	0	0.938	0.955	0.950	0.949	0.952	0.959
	0.025	0.915	0.819	0.724	0.392	0.313	0.267
	0.05	0.758	0.241	0.110	0.003	0.000	0.001
	0.075	0.492	0.032	0.001	0.000	0.000	0.000
	0.1	0.070	0.000	0.000	0.000	0.000	0.000
	0.2	0.000	0.000	0.000	0.000	0.000	0.000
	0.25	0.000	0.000	0.000	0.000	0.000	0.000
	0.4	0.000	0.000	0.000	0.000	0.000	0.000
	0.5	0.000	0.000	0.000	0.000	0.000	0.000

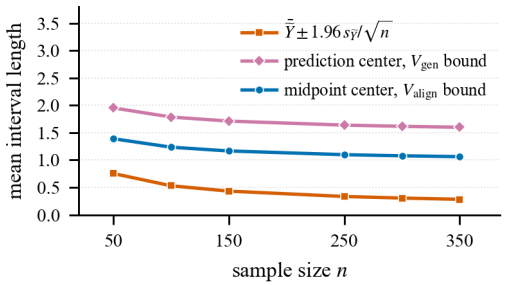
I.3.4 RESULTS FOR Q_2

Figure 5 repeats for Q_2 the four diagnostics reported for Q_1 in Figure 1 and additionally displays the two variance bounds; Table 7 gives the corresponding Monte Carlo-mean reductions. The qualitative findings carry over: V_{gen} is again flat in η, $\widehat{V}_{\text{align}}$ again sits strictly between it and the clean sample variance, the certified intervals again cover at every sample size while the corrupted-label interval loses coverage as n grows, and the half-width is again split between a bias radius that grows with η and a sampling radius that is nearly inert. We record the quantitative differences here rather than in the main text.

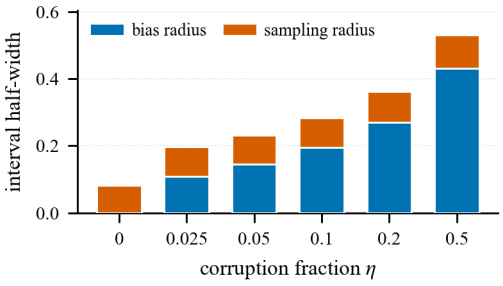
Every reduction in Table 7 is smaller than its Q_1 counterpart: 31.5% against 39.1% on the bound at $\eta = 0.05$, and 7.3% against 12.1% on interval length. The cause is the calibrated guarantee rather than the predictor. For Q_2, $\beta_{\text{cal}} = 0.692$ against a predictor spread of $\widehat{\sigma}_f = 0.323$, so $\beta_{\text{cal}}/\widehat{\sigma}_f =$



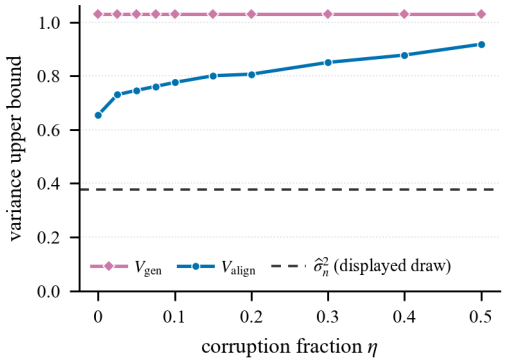
(a) Empirical coverage against n at $\eta = 0.5$; Table 6 gives every (η, n) cell.



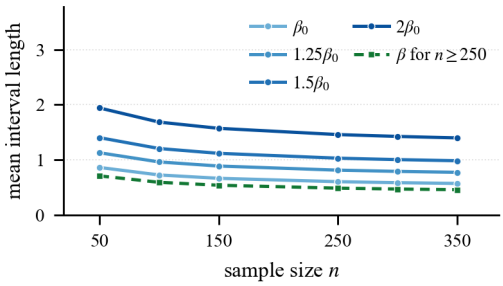
(b) Mean interval length against n at $\eta = 0.5$.



(c) Half-width by bias and sampling radius.



(d) Variance bounds against η, over the clean sample variance $\hat{\sigma}_n^2$ of the displayed draw.



(e) Length against n as the declared guarantee is relaxed from β_0 to $c\beta_0$, at $\eta = 0.1$; see Figure 8 for the multi-η version.

Figure 5: Validity and length for Q_2 under the mean-shift adversary. In (a) each point is the fraction of 1000 repetitions whose interval contains $\mu(P) = 0.906$ and the dashed line marks the required 95% coverage. Panels (a)–(c) correspond to panels (a)–(c) of Figure 1; panel (d) additionally shows the variance bounds for one displayed sample at $n = 350$, and panel (e) corresponds to the β-sensitivity analysis in Figure 1d.

Table 7: Monte Carlo mean variance bounds and reductions for the Q_2 midpoint interval at $n = 350$ over 1000 repetitions. If $\overline{L}_{\text{gen}}$ and $\overline{L}_{\text{align}}$ are the corresponding mean lengths, length reduction is $100 \times (\overline{L}_{\text{gen}} - \overline{L}_{\text{align}})/\overline{L}_{\text{gen}}\%$; variance reduction is defined analogously.

η	V_{gen}	$\widehat{V}_{\text{align}}$	variance reduction	interval-length reduction
0.05	1.030	0.705	31.5%	7.3%
0.10	1.030	0.733	28.9%	5.5%
0.20	1.030	0.790	23.3%	3.5%

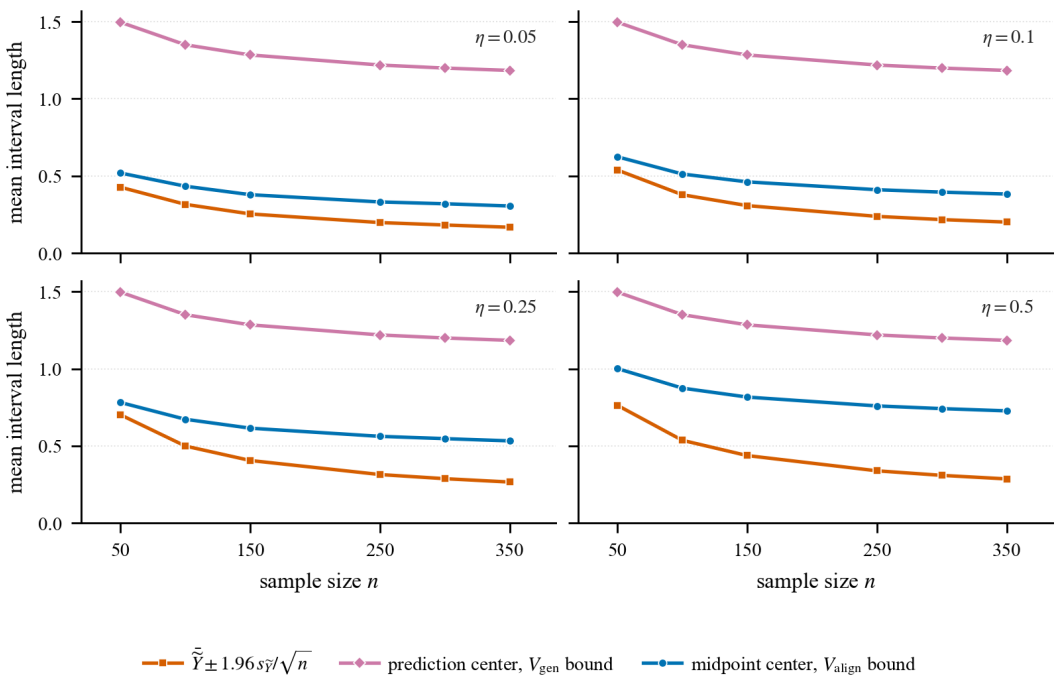


Figure 6: Mean interval length against n for Q_1 over 1000 repetitions, at four corruption levels. The vertical axis is shared across panels, so the vertical gap between curves is comparable between them.

2.14, whereas for Q_1 the same ratio is 1.19. Since $V_{\text{gen}} = (\widehat{\sigma}_f + \beta)^2$ and the alignment correction enters only through a cross term, a guarantee that is large relative to $\widehat{\sigma}_f$ leaves proportionally less for the bound to recover. The looser guarantee also raises the mean V_{gen}, from 0.830 to 1.030, against a mean clean-sample variance of $\sigma_Y^2 = 0.332$.

Bias dominates sooner. For Q_1 the bias radius overtakes the sampling radius between $\eta = 0.025$ and $\eta = 0.05$. For Q_2 the crossover has already occurred by $\eta = 0.025$, where the bias radius is 0.109 against a sampling radius of 0.087. This follows from the same looser guarantee: the midpoint bias radius scales with β, so a larger declared guarantee moves the crossover toward zero corruption. Beyond it the interval is bias-dominated and additional clean data cannot narrow it.

Sensitivity to the guarantee is comparable. Relaxing the declared guarantee from β_0 to $2\beta_0$ at $\eta = 0.1$ and $n = 350$ grows the interval from 0.568 to 1.396, a factor of 2.5, the same factor as for Q_1. The ordering across c is preserved at every sample size, as in Figure 7.

Coverage is unaffected. All three certified constructions—the prediction mean with V_{gen}, the midpoint with V_{gen}, and the midpoint with $\widehat{V}_{\text{align}}$—cover $\mu(P) = 0.906$ in 1000 of 1000 repetitions at every reported n and every $\eta > 0$; at $\eta = 0$, where the bias radius vanishes, coverage is 99.0% or above. The corrupted-label interval falls, at $\eta = 0.05$, from 0.758 at $n = 50$ to 0.001 at $n = 350$ (Table 6). The weaker Q_2 annotator agreement reported in Section I.3.2 therefore does not manifest as a coverage failure: $\mu(P)$ is the mean of the designated annotator's labels over the evaluation pool, and the guarantee concerns recovery of that quantity under corruption.

I.3.5 Interval length across corruption levels

Figure 6 plots mean interval length against n for Q_1 at four corruption levels, and Table 8 tabulates the same quantity at $n = 350$ for the full 2×2 ablation: the estimator centre (prediction mean or midpoint) crossed with the variance bound (V_{gen} or $\widehat{V}_{\text{align}}$).

Three observations follow, and they separate the two contributions cleanly.

Table 8: Mean interval length for Q_1 and Q_2 at $n = 350$ over 1000 repetitions. *pred* denotes the prediction-mean centre and *mid* the midpoint; Δ_{align} is the length removed by replacing V_{gen} with $\widehat{V}_{\text{align}}$, which is the same for both centres.

	Q_1				Q_2			
η	pred V_{gen}	mid V_{gen}	mid $\widehat{V}_{\text{align}}$	Δ_{align}	pred V_{gen}	mid V_{gen}	mid $\widehat{V}_{\text{align}}$	Δ_{align}
0.05	1.184	0.348	0.306	0.042	1.598	0.501	0.465	0.037
0.10	1.184	0.421	0.383	0.038	1.598	0.602	0.568	0.033
0.20	1.184	0.522	0.489	0.034	1.598	0.752	0.725	0.027
0.25	1.184	0.564	0.532	0.032	1.598	0.812	0.788	0.024
0.40	1.184	0.680	0.654	0.026	1.598	0.978	0.960	0.017
0.50	1.184	0.750	0.727	0.023	1.598	1.075	1.061	0.014

At $\eta = 0.05$ the midpoint shortens the Q_1 interval from 1.184 to 0.348, while replacing V_{gen} by $\widehat{V}_{\text{align}}$ shortens it a further 0.042. The label-assisted variance bound is a refinement; the estimator centre is the main effect.

The bound's contribution does not depend on the centre. For every η, the absolute length removed by $\widehat{V}_{\text{align}}$ is the same whether the centre is the prediction mean or the midpoint, to within rounding. This is exact rather than coincidental: the two constructions differ only in their bias radius, and the variance bound enters solely through the sampling radius $z_{1-\alpha/2}\overline{\sigma}_n/\sqrt{n}$, which is common to both. It follows that the 2×2 ablation is additive, and that the relative gain from $\widehat{V}_{\text{align}}$ is largest for whichever centre already has the smaller bias radius.

The midpoint's advantage weakens as corruption grows. The ratio of midpoint to prediction-mean length rises monotonically with η: for Q_1 it is 0.26 at $\eta = 0.05$, 0.45 at $\eta = 0.25$ and 0.61 at $\eta = 0.5$, and it reaches exactly 1 at $\eta = 1$, where the midpoint provably reduces to the prediction mean. The reason is that the prediction-mean interval never consults the observed labels, so its bias radius is β irrespective of η, whereas the midpoint bias radius grows as $\beta\sqrt{2k/n}$. The midpoint is therefore most valuable exactly in the regime the method targets, and degrades gracefully to the prediction-only baseline as the corruption budget approaches the identifiability threshold.

Table 9: Half-width decomposition at $n = 350$, as Monte Carlo means over 1000 repetitions under the mean-shifting adversary. The bias radius b_n grows with η while the sampling radius is nearly inert, so the bias share crosses one half between $\eta = 0.025$ and $\eta = 0.05$ for both questions: from $\eta = 0.05$ onward the interval is bias-dominated, and the variance bound acts only on the smaller term.

	η	bias radius b_n	sampling radius	half-width	bias share
Q_1	0	0.000	0.069	0.069	0.0%
	0.025	0.055	0.073	0.128	42.7%
	0.05	0.078	0.075	0.153	51.2%
	0.075	0.098	0.075	0.174	56.5%
	0.1	0.115	0.076	0.191	60.1%
	0.2	0.166	0.079	0.244	67.8%
	0.25	0.186	0.080	0.266	70.1%
	0.4	0.245	0.082	0.327	74.8%
	0.5	0.279	0.084	0.363	76.8%
Q_2	0	0.000	0.082	0.082	0.0%
	0.025	0.109	0.087	0.196	55.8%
	0.05	0.144	0.088	0.232	62.1%
	0.075	0.171	0.089	0.260	65.8%
	0.1	0.194	0.090	0.284	68.4%
	0.2	0.269	0.093	0.363	74.3%
	0.25	0.300	0.095	0.394	76.0%
	0.4	0.382	0.098	0.480	79.6%
	0.5	0.431	0.100	0.531	81.2%

I.3.6 Sensitivity across corruption levels

Figures 7 and 8 extend the main-text sensitivity analysis from $\eta = 0.1$ to every corruption level used in the β-sensitivity sweep, separately for Q_1 and Q_2.

The recalibrated guarantee. One curve in these panels, and in Figure 1d, is not a relaxation of β but a tightening of it: the guarantee recalibrated for $n \geq 250$ only, described in Section I.1. At $\eta = 0.1$ and $n = 350$ it shortens the Q_1 interval from 0.383 to 0.334. The gap between that curve and the β_0 curve therefore measures how much of the reported width is the price of one certificate covering the entire nested grid at once, rather than a property of the estimator.

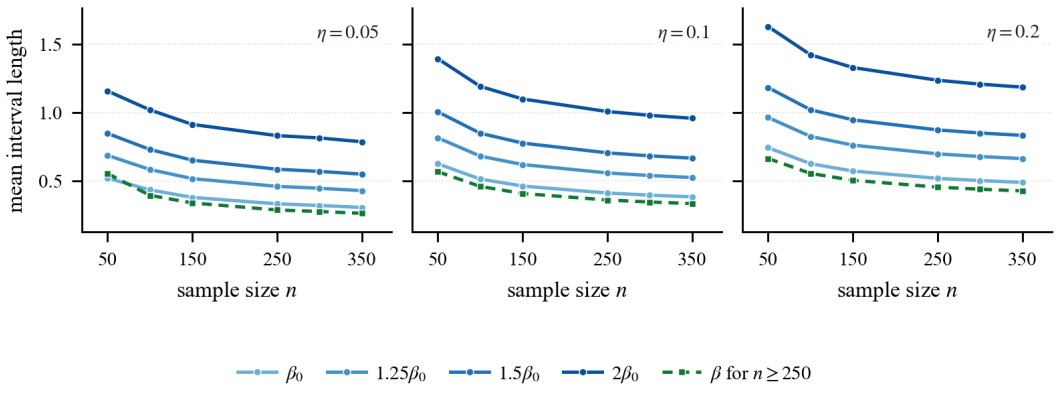


Figure 7: Mean interval length against n for Q_1 as the supplied predictor-error guarantee is relaxed from β_0 to $c\beta_0$, at $\eta \in \{0.05, 0.1, 0.2\}$. The dashed green curve uses the guarantee recalibrated only for $n \geq 250$. The shared vertical axis makes the increase with both guarantee slack and corruption directly comparable across panels.

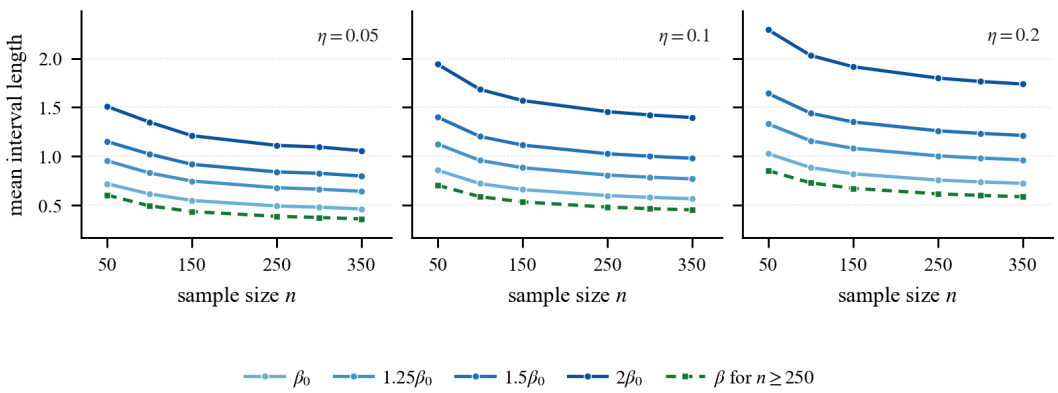


Figure 8: Mean interval length against n for Q_2 under the same sensitivity design as Figure 7: only the supplied guarantee changes from β_0 to $c\beta_0$, while the sampled labels, predictions, and attacks are held fixed. The dashed green curve uses the guarantee recalibrated only for $n \geq 250$.

I.3.7 How tight is the alignment bound?

Theorem 3.2 places $\widehat{V}_{\text{align}}$ between the optimum of Problem 1 and the generic bound V_{gen}, but not where in that interval it falls. Computing $V^{\star}_{\text{MSE}}$ is NP-hard (Theorem G.1), so we obtain it by enumeration on instances small enough to allow it. The cost is $\binom{n}{k}$ rather than n, so the enumeration reaches the sample sizes used elsewhere when k is small, and the corruption levels used elsewhere when n is small, but not both at once.

Once the changed-set is fixed, what remains is the trust-region subproblem of Section E, a one-dimensional concave maximization solved in closed form; $V^{\star}_{\text{MSE}}$ is the largest value over all $\binom{n}{k}$ sets. We checked the closed form against a general-purpose constrained optimizer on every support of 12 instances, agreeing to 10^{-9}. Each cell draws 30 samples from the evaluation pool and applies the mean-shifting adversary of Section I.2 at the calibrated β.

Table 10: Exact $V^{\star}_{\mathrm{MSE}}$ against the two bounds for Q_1 and Q_2, as cell means over 30 instances. The right block converts each bound into the sampling radius $1.96\sqrt{V/n}$ it contributes to the interval half-width. The first four rows of each block hold n small with η on the grid used elsewhere; the remainder hold n at the values used elsewhere, which enumeration reaches because k is small.

			variance bound			sampling radius $1.96\sqrt{V/n}$		
	n	k	$V^{\star}_{\mathrm{MSE}}$	$\widehat{V}_{\mathrm{align}}$	V_{gen}	$V^{\star}_{\mathrm{MSE}}$	$\widehat{V}_{\mathrm{align}}$	V_{gen}
Q_1	10	2	0.536	0.558	0.738	0.435	0.446	0.523
	14	2	0.550	0.573	0.793	0.374	0.384	0.461
	18	3	0.632	0.647	0.835	0.358	0.363	0.418
	18	5	0.667	0.680	0.835	0.368	0.373	0.418
	50	1	0.568	0.578	0.828	0.208	0.210	0.252
	50	2	0.594	0.603	0.828	0.213	0.214	0.252
	50	3	0.607	0.616	0.828	0.215	0.217	0.252
	100	2	0.584	0.586	0.834	0.149	0.150	0.179
	150	2	0.546	0.547	0.814	0.118	0.118	0.144
	250	2	0.517	0.518	0.806	0.089	0.089	0.111
	350	1	0.483	0.484	0.817	0.073	0.073	0.095
	350	2	0.509	0.510	0.817	0.075	0.075	0.095
Q_2	10	2	0.811	0.829	0.967	0.557	0.563	0.609
	14	2	0.836	0.851	1.000	0.478	0.483	0.523
	18	3	0.850	0.857	1.000	0.425	0.427	0.462
	18	5	0.903	0.908	1.000	0.438	0.440	0.462
	50	1	0.709	0.715	1.020	0.233	0.234	0.280
	50	2	0.752	0.757	1.020	0.240	0.241	0.280
	50	3	0.785	0.789	1.020	0.245	0.246	0.280
	100	2	0.713	0.715	1.021	0.165	0.166	0.198
	150	2	0.694	0.696	1.022	0.133	0.133	0.162
	250	2	0.682	0.683	1.033	0.102	0.102	0.126
	350	1	0.655	0.656	1.034	0.085	0.085	0.107
	350	2	0.672	0.673	1.034	0.086	0.086	0.107

The sandwich $\mathrm{Var}_n(Y) \le V^{\star}_{\mathrm{MSE}} \le \widehat{V}_{\mathrm{align}} \le V_{\mathrm{gen}}$ holds on all 720 instances. The alignment bound sits just above the optimum it certifies, and closes on it as n grows: at $n = 350$ the two agree to three decimals, while V_{gen} stays well above both. The sampling radius scales with $\sqrt{V}$, so the separation there is milder than in the variance itself: for Q_1 at $n = 350$, V_{gen} exceeds $\widehat{V}_{\mathrm{align}}$ by 60% but its sampling radius by 27%.

I.4 Human-labeled response similarity

I.4.1 Data and human labels

We use the 2024 U.S. election-response corpus of Cen et al. (2025), with the human deep-meaning similarity annotations subsequently collected by Emmenegger et al. (2026). We select the family of questions asking for a presidential candidate's position on a political issue. Each row compares two responses from the same generating model to the same question under two prompt-persona variations. This gives 968 response pairs from 20 questions, 3 generating models, and 24 persona comparisons. Each human judgment is one of *very dissimilar*, *slightly dissimilar*, *slightly similar*, or *very similar*, mapped to 0, $1/3$, $2/3$, and 1, respectively. The outcome $Y \in [0, 1]$ is `human_label`, the mean of the available human judgments for that response pair (normally five); automated scores are never used as clean labels. The target mean measures average persona-conditioned consistency over the selected family: higher values indicate that answers are more stable across persona prompts, whereas lower values indicate stronger persona conditioning.

Leakage-safe split. A single generated response can appear in several persona pairs, so a rowwise split could place that response in both development and evaluation. We instead treat all pairs sharing a question and generating model as one response-sharing group and assign each group wholly to one side. The resulting split contains 616 development pairs in 30 groups and 352 evaluation pairs in 18 groups, with no response shared across the split. The evaluation size is the closest attainable value

to 350 because the response-sharing groups have sizes divisible by eight. Every repetition draws a length-350 sequence IID with replacement from the fixed evaluation distribution; its nested prefixes at $n \in \{50, 100, 150, 250, 300, 350\}$ target the evaluation-distribution mean $\mu(P) = 0.751$.

I.4.2 Predictor and train-only calibration

We train our own text predictor using only the development labels. No precomputed automated similarity score distributed with the dataset is used to construct the intervals. For each response, we construct TF–IDF representations from word unigrams, word bigrams, and short character sequences. For each pair, we use absolute differences, overlaps, cosine similarities, and relative-length summaries. These features are unchanged if the two responses are swapped, as a similarity prediction should be. We then fit ridge regression to the resulting pair features. Because one generated response can occur in several pairs, all pairs belonging to the same question–model group are assigned to the same fold. Thus, no response-sharing group appears in both the training and validation portions of a cross-validation split.

We use nested cross-fitting on the development set. For each outer fold, inner group-wise folds within the outer-training data select the ridge penalty from $\{0.01, 0.1, 1, 10, 100\}$. We refit using the selected penalty on the full outer-training portion and predict the outer-held-out groups. Concatenating these predictions gives one out-of-fold prediction for every development pair, from a model that used neither that pair nor any pair in its question–model group. A separate group-wise cross-validation using all development rows selects the final penalty, which is 1. We then fit the predictor on all 616 development pairs and freeze it before predicting the evaluation pool.

The nested out-of-fold development RMSE is 0.169. Applying the calibration of Section I.1 to these cross-fitted squared errors gives $\beta_{\mathrm{cal}} = 0.201$. The held-out evaluation RMSE is 0.156; this is a post-hoc audit and is never used for model selection or calibration. The simultaneous certificate is violated on 8 of 1,000 nested evaluation paths (0.8%), emphasizing that the train-only bootstrap certificate is empirical rather than distribution-free.

I.4.3 Results

Figures 9 and 10 report coverage and length under the downward mean-shift attack. At $n = 350$ and $\eta = 0.5$, the interval induced by using the midpoint estimator and $\widehat{V}_{\mathrm{align}}$ as its variance upper bound has 100.0% coverage and mean length 0.278 ± 0.0003 (Monte Carlo s.e.). The valid interval induced by using the mean of the predictions as the estimator and V_{gen} as the variance upper bound has mean length 0.469, so the midpoint construction is 40.7% shorter. The interval computed from the corrupted-label mean and variance has mean length 0.069 but 0.0% coverage. Figure 12 isolates the two choices: at $\eta = 0.05$, replacing V_{gen} by $\widehat{V}_{\mathrm{align}}$ while holding the midpoint estimator fixed reduces the mean variance bound by 32.6% and mean length from 0.134 to 0.122 (8.9%). Comparing the prediction-mean construction using V_{gen} with the midpoint construction using $\widehat{V}_{\mathrm{align}}$ reduces mean length from 0.469 to 0.122 (73.9%), showing that changing the center accounts for most of the gain. Figure 11 repeats the guarantee-slack study at $\eta \in \{0.05, 0.1, 0.2\}$, using the same design as for Q_1 and Q_2.

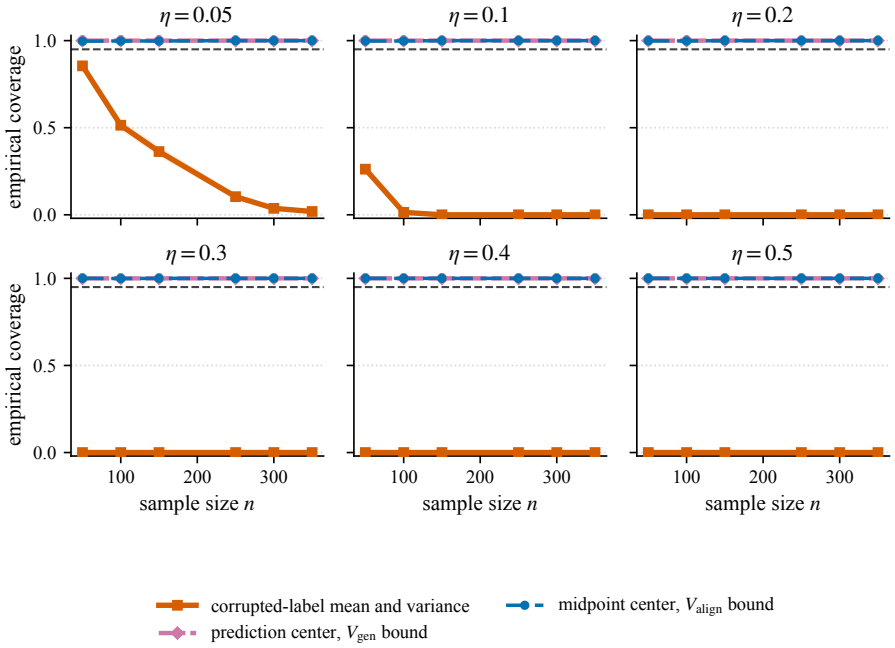


Figure 9: *Similarity experiment*: empirical coverage for the human-labeled candidate-position similarity family of Emmenegger et al. (2026). Orange denotes the interval computed from the corrupted-label mean and sample variance; pink denotes the interval induced by using the mean of the predictions as the estimator and V_{gen} as its variance upper bound; blue denotes the interval induced by using the midpoint estimator and $\widehat{V}_{\mathrm{align}}$ as its variance upper bound. Shaded regions are 95% Wilson intervals over 1,000 repetitions, and the dashed line marks the nominal 95% level. All intervals target the mean $\mu(P) = 0.751$ of the fixed 352-pair evaluation distribution.

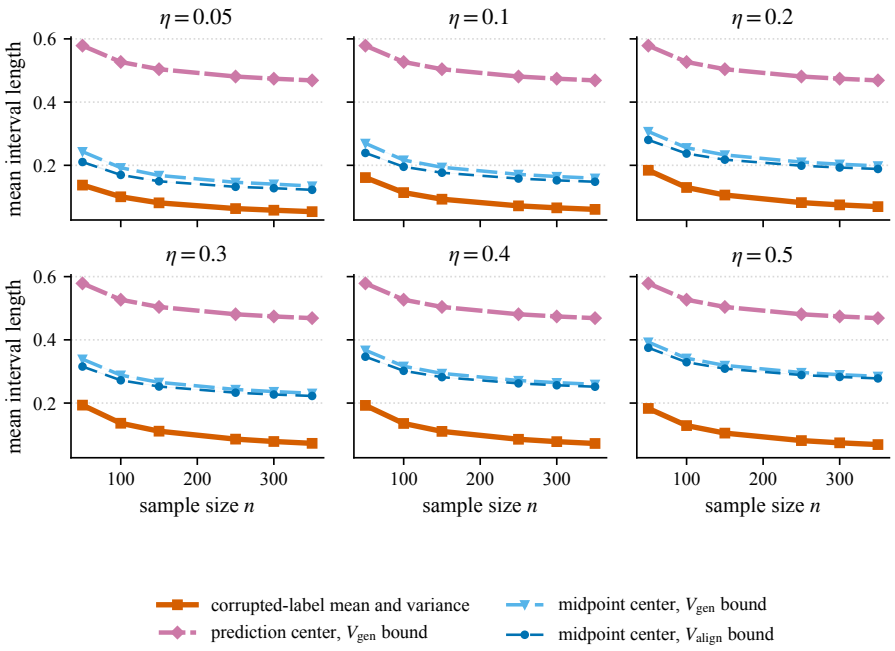


Figure 10: *Similarity experiment*: mean interval length. Light blue adds the interval induced by using the midpoint estimator and V_{gen} as its variance upper bound, thereby isolating the contribution of the alignment-based variance refinement. Shading gives 95% Monte Carlo intervals for the estimated mean length.

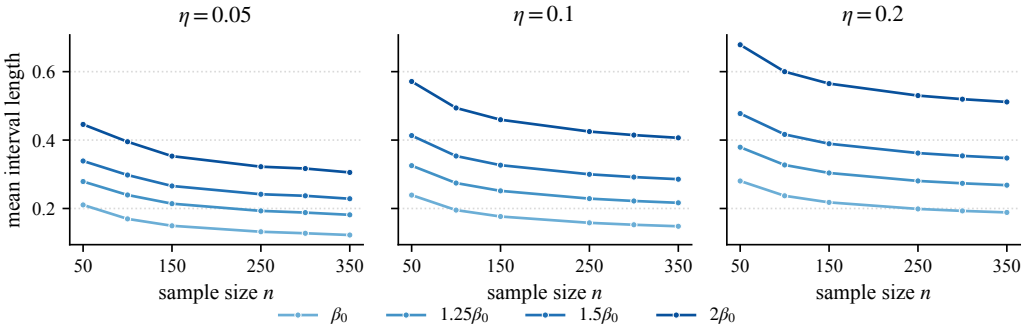


Figure 11: *Similarity experiment*: sensitivity of the midpoint–$\widehat{V}_{\mathrm{align}}$ interval to slack in the supplied predictor guarantee. The development-calibrated value is $\beta_0 = 0.201$; each panel replaces it by $c\beta_0$ while holding the frozen predictions, evaluation draws, and downward mean-shift attacks identical across $c \in \{1, 1.25, 1.5, 2\}$. Curves report mean length over 1,000 repetitions; shaded 95% Monte Carlo intervals are generally narrower than the plotted lines.

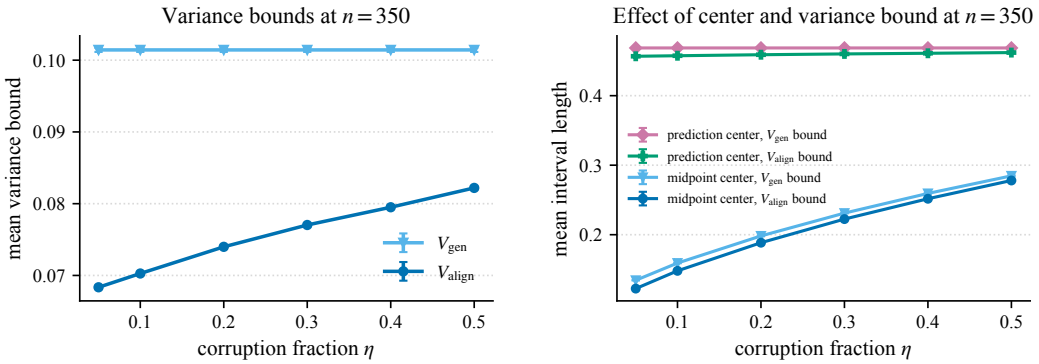


Figure 12: *Similarity experiment*: variance-bound comparison at $n = 350$. The left panel compares V_{gen} and $\widehat{V}_{\text{align}}$ while holding the midpoint estimator fixed. The right panel crosses the prediction-mean and midpoint estimators with the two variance upper bounds, separating the effect of changing the center from that of refining the variance bound. Error bars are ± 1.96 Monte Carlo standard errors and are often smaller than the markers.