
Efficient and Accurate Causal Validation of Training Data Attribution

Simon Vary^{1*} Alireza Mousavi-Hosseini^{1,2*} Roger Grosse^{2,3}

¹MATS Research

²University of Toronto

³Vector Institute

Abstract

Causal validation of training data attribution (TDA) methods is costly, since evaluation involves retraining a model on many different subsets of the original training set. For example, the Linear Datamodeling Score (LDS), which has become a standard metric for evaluating TDA methods, measures the correlation between TDA predictions and expected retrained model outputs. In this paper, we reduce the amount of computation needed to estimate the LDS along three axes: the number of sampled training subsets, the number of retraining seeds per subset, and the number of models in the attribution ensemble. Each of these factors induces a form of bias in the LDS estimation, and removing these sources of bias significantly reduces the number of retrains needed to match the same accuracy in LDS estimation. On ResNet-9/CIFAR-10, finite-subset debiasing achieves 3.2% relative RMSE with as few as 4 subsets, and 32 subsets match the accuracy of ordinary Spearman using 96 subsets, measured against a 300-subset reference. On ResNet-9/CIFAR-10, BERT/QNLI and GPT-2/WikiText-2, estimating the training-noise variance allows us to reproduce the standard five-seed LDS at the same accuracy with $2\text{--}5\times$ fewer retrains.

1 Introduction

Training Data Attribution (TDA) methods [Koh and Liang, 2017, Pruthi et al., 2020, Grosse et al., 2023, Park et al., 2023, Bae et al., 2024] aim to approximate the causal effect of training data on model behaviour after training. A widely used metric for evaluating the accuracy of their causal predictions is the Linear Datamodeling Score (LDS), which was first introduced by Park et al. [2023, Definition 2.3] in the context of TDA and builds on the datamodel framework of Ilyas et al. [2022]. LDS measures the correlation, across training subsets, between TDA predictions and the expected outputs of models retrained on those subsets, where the expectation is over training randomness. As such, computing LDS is computationally expensive, since each sample used to estimate the correlation requires multiple full model retrains, one for every random training seed. This limits evaluation to small scales. LDS has nevertheless become a standard evaluation practice for TDA, and has been adapted even beyond the supervised learning setting in which it was first introduced [Georgiev et al., 2023, Bae et al., 2024, Deng et al., 2024b, Quirke et al., 2026].

In order to reduce the computational cost of evaluating LDS, we analyze the key sources of error in the LDS estimates and give recipes for reducing each error source at a given level of computation. The first source of error is randomness in the training procedure itself (e.g. from parameter initialization and batch ordering), which results in noise in the ground truth targets themselves. One typically mitigates this noise by averaging over multiple random seeds, but this is expensive as it must be done for *each* data subset. To cheaply reduce this error source, we use multiple random seeds to estimate

*Equal contribution.

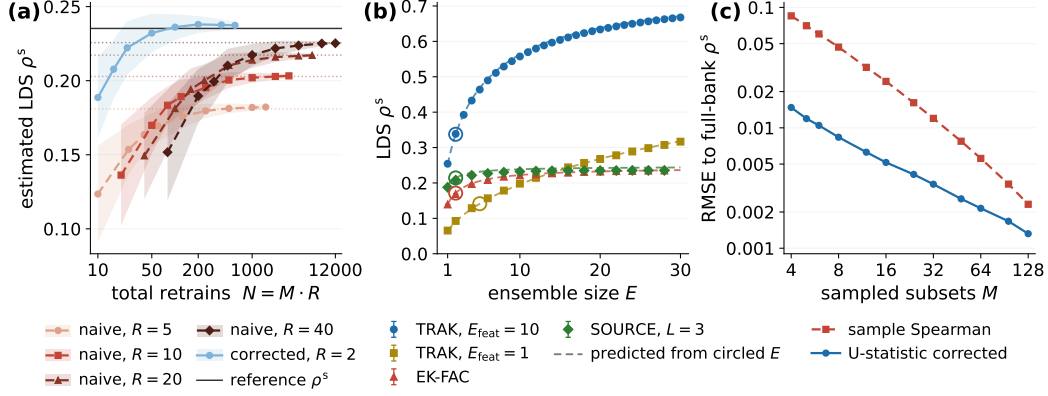


Figure 1: **Three ways to reduce LDS evaluation cost on CIFAR-10/ResNet-9.** (a) Correcting for the training variance converges quicker to a Spearman reference computed from held-out retraining seeds, while naive seed averaging plateaus (EK-FAC). (b) The circled small attribution ensembles predict the LDS of larger ensembles (dashed lines). (c) Correcting finite-subset Spearman bias reduces RMSE against the 300-subset reference (TRAK).

the label noise, and then correct for it when estimating LDS. Empirically, this allows us to estimate the noise-free LDS values reasonably accurately with as few as 2 random seeds per subset (Fig. 1a).

A second source of error is the randomness in the TDA method itself (e.g., due to the choice of checkpoint at which to evaluate the TDA criterion). It has been observed that ensembling TDA methods reduces this randomness and improves their performance [Deng et al., 2024a, Bae et al., 2024, Park et al., 2023], but evaluating ensembles can be expensive. By estimating the attribution noise-to-signal ratio and its effect under averaging, we can predict the LDS of large TDA ensembles from as few as two members (Fig. 1b).

Finally, a third source of error comes from estimating the correlation from a finite number of subsets. We observe that a significant part of the LDS error is caused by a shared bias across different collections of training subsets. Correcting this bias substantially reduces the retraining budget needed for accurate evaluation (Fig. 1c).

As a side observation which our analysis depends on but which may be of independent interest, we also observe that the paired TDA predictions and the retrained model outputs are well approximated by a jointly Gaussian distribution. This allows us to convert the corrected Pearson correlation to Spearman correlation, the latter being the standard correlation measure used to compute LDS [Park et al., 2023, Definition 2.1].

Our code and precomputed retraining banks are publicly available².

2 Setup: Linear Datamodeling Score

The Linear Datamodeling Score (LDS) measures the correlation, across training subsets $S \subset D$, between TDA predictions $G_q(S)$ and expected retrained model outputs $F_q(S)$. Here q denotes a query test point, and the expectation is over training randomness.

We sample $S_1, \dots, S_M$ independently from $\pi_S := \text{Unif}\{S \subset D : |S| = \alpha|D|\}$, and let $\hat{\pi}_S$ denote their empirical distribution, which assigns weight $1/M$ to each sampled subset.

Definition 1 (Linear Datamodeling Score). The empirical LDS is the Spearman correlation between the TDA predictions and expected retrained model outputs, averaged over the Q queries:

$$\hat{\rho}_M^s := \frac{1}{Q} \sum_{q=1}^Q \hat{\rho}_M^s(q), \quad \hat{\rho}_M^s(q) := \text{Corr}_{S \sim \hat{\pi}_S}(\text{rank}_M(G_q(S)), \text{rank}_M(F_q(S))), \quad (1)$$

²Code: <https://github.com/simonvany/corrected-lds>.

Retraining banks: <https://huggingface.co/datasets/corrected-lds/data>.

where rank_M denotes ranks among the M sampled subsets.

The corresponding population LDS is

$$\rho^s := \frac{1}{Q} \sum_{q=1}^Q \rho^s(q), \quad \rho^s(q) := \text{Corr}_{S \sim \pi_s} (H_q^G(G_q(S)), H_q^F(F_q(S))), \quad (2)$$

where H_q^G and H_q^F are the cumulative distribution functions of $G_q(S)$ and $F_q(S)$, respectively.

While the standard LDS uses Spearman correlation [Park et al., 2023], we derive the noise corrections using Pearson correlation, whose covariance form makes the contribution of the training noise explicit, and later we convert the estimates to Spearman LDS under a joint Gaussian assumption. We define Pearson LDS as

$$\rho^p := \frac{1}{Q} \sum_{q=1}^Q \rho^p(q), \quad \rho^p(q) := \text{Corr}_{S \sim \pi_s} (G_q(S), F_q(S)), \quad (3)$$

and the empirical counterpart $\hat{\rho}_M^p$ is defined by replacing π_s with $\hat{\pi}_s$.

2.1 Statistical Error in LDS Estimation

Stochastic Training Noise. Retraining on the same subset can produce different model outputs because of training randomness. Instead of observing the expected output $F_q(S)$, we observe $f_q(S, \xi)$:

$$f_q(S, \xi) = F_q(S) + \varepsilon_q(S, \xi), \quad \text{with} \quad F_q(S) := \mathbb{E}_\xi[f_q(S, \xi)|S] \quad \text{and} \quad \mathbb{E}[\varepsilon_q(S, \xi)|S] = 0, \quad (4)$$

where $\varepsilon_q(S, \xi)$ is the noise arising from the training randomness ξ , e.g., data ordering or random initialization. Note, that a single training run is shared by all queries, so the noises ε_q are correlated across q . It is common practice to use $\bar{f}_{R,q}(S) = F_q(S) + \frac{1}{R} \sum_{r=1}^R \varepsilon_q(S, \xi_r)$ as a proxy for $F_q(S)$, with, for example, $R = 5$ [Park et al., 2023, Appendix C.4] or higher [Bae et al., 2024, Appendix B.2].

Attribution Method Noise. The attribution predictions are also stochastic. The computed scores $g_q(S, \zeta)$ depend on the TDA method’s own randomness ζ (e.g. randomness of the trained model, random projections in TRAK [Park et al., 2023], MC sampling in EK-FAC [Grosse et al., 2023]). Namely we have

$$g_q(S, \zeta) = G_q(S) + \gamma_q(S, \zeta), \quad \text{with} \quad \mathbb{E}_\zeta[g_q(S, \zeta)|S] = G_q(S) \quad \text{and} \quad \mathbb{E}[\gamma_q(S, \zeta)|S] = 0, \quad (5)$$

where ζ is the randomness of the attribution method resulting in the TDA noise $\gamma_q(S, \zeta)$. As a result, it is a common practice to compute E independently random instances of the same TDA method and average their attribution scores $\bar{g}_{E,q}(S) = G_q(S) + \frac{1}{E} \sum_{e=1}^E \gamma_q(S, \zeta_e)$ to reduce the effect of the noise.

Finite-Subset Error. In practice, using a finite number M of subsets introduces an error between $\hat{\rho}_M^s$ and ρ^s . To control the error, naive LDS estimation requires a large M . When combining all finite-sample effects, we end up with the following estimate

$$\hat{\rho}_{R,E,M}^s(q) = \text{Corr}_{S \sim \hat{\pi}_s} (\text{rank}_M(\bar{f}_{R,q}(S)), \text{rank}_M(\bar{g}_{E,q}(S))).$$

3 Improved LDS Estimation

We first estimate LDS against expected retrained outputs while holding a finite attribution ensemble fixed. We then consider two related targets: the LDS of a conventional finite-seed evaluation and the training-noise-free LDS of a larger attribution ensemble. Finally, we address the bias from estimating Spearman correlation using finitely many subsets.

3.1 Correcting for the Variance of the Training and the Attribution Noise

We show how estimating the variances of training and attribution noise can reduce the cost of LDS evaluation.

In the following we assume the randomness of the training and the TDA method are independent.

Assumption 1 (Independent Training Seeds). For each subset S , the R retraining seeds $\xi_1, \dots, \xi_R$ are i.i.d. and are further independent of the E i.i.d. ensemble seeds $\zeta_1, \dots, \zeta_E$.

In practice, it is always possible to design such an experiment by using different random seeds for the training and the TDA method and thus this assumption is not restrictive.

Correcting training noise. For a single query, we omit the index q to ease the notation and consider Pearson LDS computed from M subsets, R training seeds per subset, and a TDA ensemble with E members:

$$\hat{\rho}_{R,E,M}^p = \frac{\text{Cov}_{S \sim \pi_S}(\bar{f}_R(S), \bar{g}_E(S))}{\sqrt{\text{Var}_{S \sim \pi_S}(\bar{f}_R(S)) \text{Var}_{S \sim \pi_S}(\bar{g}_E(S))}}. \quad (6)$$

Later, in equation (11), we give the formula that converts the population Pearson LDS to Spearman LDS under Gaussianity assumption.

To isolate the effect of training noise, fix the attribution ensemble $\bar{g}_E$, and consider the *population correlation* over the training subsets $S \sim \pi_S$ and retraining randomness. Under Assumption 1, the zero-mean training noise leaves the covariance unchanged, while averaging independent retraining seeds reduces its variance:

$$\text{Cov}_{S, \xi_{1:R}}(\bar{g}_E, \bar{f}_R) = \text{Cov}_S(\bar{g}_E, F), \quad \text{Var}_{S, \xi_{1:R}}(\bar{f}_R) = \sigma_F^2 + \frac{\sigma_\varepsilon^2}{R},$$

where $\sigma_F^2 = \text{Var}_S(F)$ and $\sigma_\varepsilon^2 = \text{Var}_{S, \xi}(\varepsilon)$. Writing $\lambda = \sigma_\varepsilon^2 / \sigma_F^2$ for the training-noise to signal ratio, we obtain

$$\rho_{R,E}^p := \text{Corr}_{S, \xi_{1:R}}(\bar{g}_E(S), \bar{f}_R(S)) = \frac{\rho_E^p}{\sqrt{1 + \lambda/R}}, \quad (7)$$

where $\rho_E^p = \text{Corr}_S(\bar{g}_E(S), F(S))$ is the population LDS without retraining noise, for the same attribution ensemble. Equation (7) shows that training noise attenuates the population Pearson LDS towards zero by a factor of $(1 + \lambda/R)^{-1/2}$. The empirical score $\hat{\rho}_{R,E,M}^p$ estimates this attenuated correlation, so increasing M alone cannot remove the bias.

Independent retrains allow us to estimate the noise variance and subtract its contribution to the observed readout variance. Algorithm 1 computes the sample variance across retraining seeds for each subset, then averages these variances to obtain $\hat{\sigma}_\varepsilon^2$. The estimated signal variance is therefore $\hat{\sigma}_F^2 = \widehat{\text{Var}}_M(\bar{f}_R) - \hat{\sigma}_\varepsilon^2/R$. This applies the classical correction for attenuation [Spearman, 1904] to the retraining outputs, with the attribution predictor held fixed.

At a fixed budget of the number of retrains $N = MR$, naive averaging with more seeds, i.e., larger R , reduces the training noise bias, but leaves fewer subsets for estimating the correlation. The following proposition quantifies this trade-off.

Proposition 1 (RMSE of naive seed averaging). *Fix the attribution ensemble $\bar{g}_E$. Suppose Assumption 1 holds, the observations are i.i.d. across subsets, $\bar{g}_E, F, \varepsilon$ have finite fourth moments, and $\text{Var}_S(\bar{g}_E), \text{Var}_S(F) > 0$.³ Then the sample Pearson correlation $\hat{\rho}_{R,E,M}^p$ of $(\bar{g}_E(S_m), \bar{f}_R(S_m))_{m=1}^M$ satisfies*

$$\mathbb{E}[(\hat{\rho}_{R,E,M}^p - \rho_E^p)^2]^{1/2} = \mathcal{O}\left(\sqrt{\frac{R}{N}} + \frac{\lambda}{R}\right),$$

with a constant that is independent of $M \geq 2$ and $R \geq 1$. For fixed $\lambda > 0$, allocating $R \asymp \lambda^{2/3} N^{1/3}$ gives RMSE $\mathcal{O}(\lambda^{1/3} N^{-1/3})$ as $N \rightarrow \infty$.

Balancing the sampling and the noise attenuation terms in Proposition 1 gives the $\mathcal{O}(N^{-1/3})$ bound for naive averaging; the proof is given in Section A.1. However, the variance correction in Algorithm 1 achieves the following $\mathcal{O}(N^{-1/2})$ bound with a fixed number of seeds $R \geq 2$.

Proposition 2 (RMSE of training-noise correction). *Let $\hat{\rho}_{R,E,M}^{p, \text{corr}}$ denote the estimate returned by Algorithm 1 with the fixed attribution predictor $\bar{g}_E$. Under the assumptions of Proposition 1, for any fixed number of seeds $R \geq 2$,*

$$\mathbb{E}[(\hat{\rho}_{R,E,M}^{p, \text{corr}} - \rho_E^p)^2]^{1/2} = \mathcal{O}(N^{-1/2}) \quad \text{as } N = MR \rightarrow \infty, \quad (8)$$

where the constant can depend on R .

³See Section A for the full assumptions, estimator definitions, and finite-sample bounds.

The proof is given in Section A.2. Although these bounds apply to Pearson LDS, in (11) we provide an additional assumption needed to convert it to Spearman LDS.

Reconstructing finite-seed LDS from fewer seeds.

To compare with existing evaluations, we may wish to reproduce the LDS targeted by a naive R -seed averaging protocol, which includes the attenuation from the training noise. By estimating the noise-free Pearson LDS and the noise-to-signal ratio λ from as few as $k \geq 2$ seeds per subset, we can then reconstruct the finite-seed target $\rho_{R,E}^p$ as

$$\hat{\rho}_{R,E,M}^{\text{p,rec}} = \frac{\hat{\rho}_{k,E,M}^{\text{p,corr}}}{\sqrt{1 + \hat{\lambda}/R}}. \quad (9)$$

For $k < R$, this uses fewer retrains than directly averaging R seeds per subset. Estimator details are given in Section C.1.

Predicting LDS of TDA ensembles. Attribution ensembling has the same $1/E$ noise-variance scaling. This motivates a forecast analogous to (7), with (R, λ) replaced by (E, c) , where $c = \mathbb{E}_\zeta[\text{Var}_S(\gamma(S, \zeta))]/\text{Var}_S(G)$ is the attribution noise-to-signal ratio⁴. Using R retraining seeds on M subsets, we can forecast the training-noise-free LDS of an E -member ensemble from $K \leq E$ observed attribution members as

$$\hat{\rho}_{R,E,M}^{\text{p,pred}} := \hat{\rho}_{R,K,M}^{\text{p,corr}} \sqrt{\frac{1 + \hat{c}/K}{1 + \hat{c}/E}}, \quad E \geq K \geq 2. \quad (10)$$

Here $\hat{\rho}_{R,K,M}^{\text{p,corr}}$ is the training-noise-corrected Pearson LDS of the observed K -member ensemble. We estimate c from the variance across subsets of differences between independent attribution members; see Section C.2. Thus, a small attribution ensemble can forecast the LDS of larger ensembles, as tested in Figs. 1b and 2.

Converting Pearson LDS to Spearman LDS. To express the corrections above in terms of Spearman LDS, we use a Gaussian assumption.

Assumption 2 (Gaussian signal pair). The pair $(G(S), F(S))$ is jointly Gaussian under $S \sim \pi_S$.

Under Assumption 2, the population Spearman and Pearson correlations satisfy

$$\rho^s(q) = \frac{6}{\pi} \arcsin(\rho^p(q)/2). \quad (11)$$

For finite- R or finite- E targets, the assumption requires joint Gaussianity of the corresponding averaged prediction-readout pair. We assess the accuracy of the conversion in Fig. 3 and examine the validity of the Gaussian assumption in Fig. 5 in the appendices.

Numerical tests. Figure 2 tests the corrections on ResNet-9/CIFAR-10, BERT/QNLI and GPT-2/WikiText-2, using TRAK, EK-FAC [Grosse et al., 2023], SOURCE [Bae et al., 2024] and TracIn [Pruthi et al., 2020]. Taking the median over queries on the full banks gives estimated retraining noise-to-signal ratios of 3.43, 0.19 and 0.37 for CIFAR-10, QNLI and WikiText-2, respectively. We specify the experimental setup in Section D.

Algorithm 1 R -seed noise-corrected LDS

Require: $\bar{g}_E$; subsets $S_1, \dots, S_M$; $M, R \geq 2$

```

for  $m = 1$  to  $M$  do
  for  $r = 1$  to  $R$  do
     $f_{m,r} \leftarrow f(S_m, \xi_{m,r})$ 
  end for
   $\bar{f}_m \leftarrow \frac{1}{R} \sum_r f_{m,r}$ 
   $s_m^2 \leftarrow \frac{1}{R-1} \sum_r (f_{m,r} - \bar{f}_m)^2$ 
end for
 $\hat{\sigma}_\varepsilon^2 \leftarrow \frac{1}{M} \sum_m s_m^2$ 
 $\hat{\sigma}_f^2 \leftarrow \widehat{\text{Var}}_M(\bar{f}_m)$ 
 $\hat{\sigma}_F^2 \leftarrow \hat{\sigma}_f^2 - \hat{\sigma}_\varepsilon^2/R$ 
 $\hat{\sigma}_g^2 \leftarrow \widehat{\text{Var}}_M(\bar{g}_E(S_m))$ 
 $\hat{C} \leftarrow \widehat{\text{Cov}}_M(\bar{f}_m, \bar{g}_E(S_m))$ 
if  $\hat{\sigma}_F^2 \leq 0$  or  $\hat{\sigma}_g^2 \leq 0$  then
  return 0
end if
return  $\text{clip}_{[-1,1]}(\hat{C}/(\hat{\sigma}_g \hat{\sigma}_F))$ 

```

⁴This is the analogue to the previously defined λ , but here on the attribution arm.

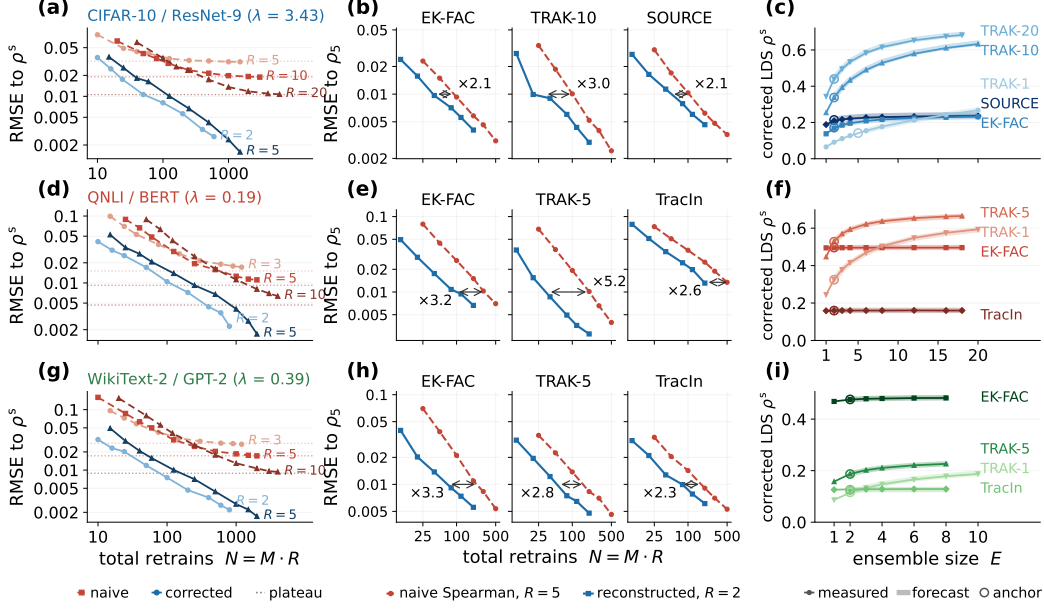


Figure 2: **Noise correction reduces the cost of LDS evaluation.** Rows: CIFAR-10/ResNet-9, QNLI/BERT, WikiText-2/GPT-2. (a, d, g) Naive and corrected Pearson LDS versus seeds per subset R , with predicted attenuation. (b, e, h) Reconstructing the five-seed Spearman target ρ_5 from two seeds matches naive averaging with $2.1\text{--}5.0\times$ fewer retrainings. (c, f, i) Measured noise-corrected Spearman LDS and predictions from the circled small ensembles. The first two columns use $E = 1$ and disjoint retraining seeds for estimation and reference.

The first column of Fig. 2(a, d, g) compares naive and corrected Pearson LDS across seeds per subset R . The naive estimates obtained by averaging R seeds follow attenuation curves computed from the corrected Pearson LDS and noise-to-signal ratio estimated using held-out retraining seeds. On CIFAR-10, which has the largest noise-to-signal ratio, the two-seed correction substantially reduces the bias that remains even with naive $R = 20$ averaging. Figures 6 and 7 further illustrate the effect of correcting training noise when targeting the noise-free LDS, showing RMSE across repeated subsampling draws and estimates from a single draw, respectively.

The second column of Fig. 2(b, e, h) plots RMSE against total retrainings, using a reference for the five-seed Spearman target ρ_5 computed from held-out retraining seeds. Reconstructing this target from $k = 2$ seeds per subset, using the Pearson–Spearman conversion, matches naive five-seed averaging with $2\text{--}5\times$ fewer retrainings ($2.2\text{--}3.3\times$ on WikiText-2). Thus, even the biased finite-seed LDS can be estimated more efficiently via (9).

The third column of Fig. 2(c, f, i) shows that small ensembles can accurately predict the LDS of larger ensembles using the prediction formula in (10).⁵ Apart from TRAK-1, forecasts from just $K = 2$ members are within 0.008 of measured noise-corrected Spearman LDS on CIFAR-10 and QNLI, and within 0.005 on WikiText-2 over the plotted ensemble sizes. TRAK-1 is the main exception with the forecasts underestimating

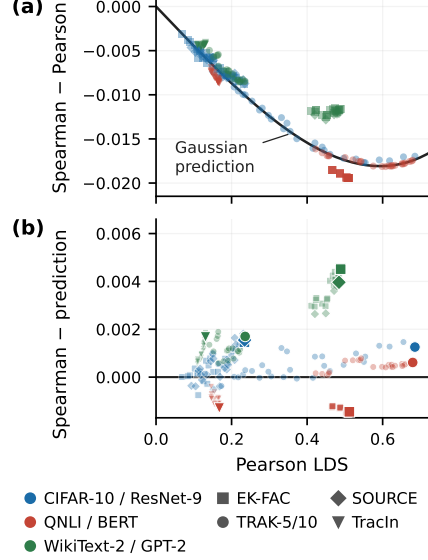


Figure 3: **Spearman versus Pearson LDS.** (a) Spearman minus Pearson LDS; curve: Gaussian prediction. (b) Residual from the finite- M Gaussian expectation. Large markers: full ensemble, all seeds.

⁵For TRAK- k , each member uses k models for featurization [Park et al., 2023], so E members use kE models.

the measured LDS by 0.065 on CIFAR-10 ($K = 5$, $E = 100$) and 0.014 on WikiText-2 ($K = 2$, $E = 50$). We suspect that TRAK-1 attribution scores have higher variance and therefore require more members to estimate the noise-to-signal ratio c reliably.

Figure 3 compares Pearson and Spearman LDS across 220 settings, varying the dataset, attribution method, E and R . Their difference reaches about 0.02 in panel (a), while residuals from the exact finite- M Gaussian expectation in panel (b) remain below 0.006, with the largest departures on WikiText-2. This supports the conversion as an approximation in the tested settings and we give tests of Gaussianity in Fig. 5 in the appendix.

3.2 Finite-Subset Bias

We now address the bias from estimating Spearman correlation using finitely many subsets. This bias is particularly relevant when LDS is averaged over many queries, because, while the effect of averaging can reduce sampling variance, the systematic bias persists. In our experiments, bias dominates the estimation error (Figures 4 and 8). This motivates using an unbiased U -statistic estimator [Hoeffding, 1948], despite its potential variance cost.

Let $\hat{\tau}_M$ denote Kendall’s sample tau-a, averaged over the same queries as $\hat{\rho}_M^s$. For $M \geq 3$ i.i.d. observations with continuous marginals (and hence no ties),

$$\hat{\rho}_M^{s, \text{unb}} := \frac{(M+1)\hat{\rho}_M^s - 3\hat{\tau}_M}{M-2} \quad (12)$$

is an unbiased estimator of ρ^s , i.e. $\mathbb{E}[\hat{\rho}_M^{s, \text{unb}}] = \rho^s$ [Hoeffding, 1948, §9]. The definition of $\hat{\tau}_M$ and the proof are given in Section B.

In Fig. 1c, subsets are sampled without replacement from a precomputed retraining bank containing model readouts for 300 training subsets, with 40 retraining seeds per subset. We use the finite-population adjustment in (18) to target the Spearman LDS computed on all 300 subsets. Figure 4 shows how bias is the dominant term in estimating the Spearman correlation for the CIFAR-10 dataset, and how using (12) removes this bias. We test this debiasing for three different TDA methods, namely TRAK, EK-FAC, and SOURCE, on CIFAR-10 and MNIST datasets, where for all instances the model is trained from scratch. The result in Figure 9 shows similar gains in reducing the estimation error. Averaging over queries reduces variance, making finite-subset bias the dominant source of error in these experiments. This is demonstrated in Figure 8 in the appendix.

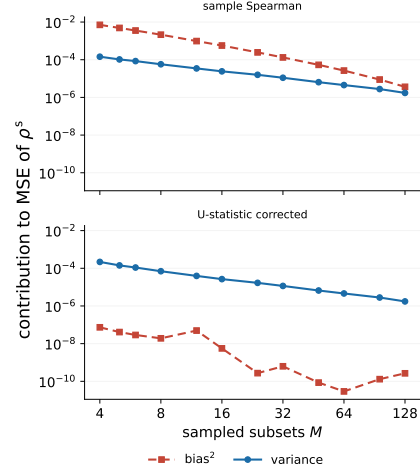


Figure 4: Bias (squared) and variance of estimating TRAK LDS on CIFAR-10 using sample Spearman $\hat{\rho}^s$ and the unbiased estimator $\hat{\rho}^{s, \text{unb}}$. Using the unbiased estimator removes the large bias in estimating LDS.

4 Discussion

Our results show that LDS evaluation can be made substantially cheaper by correcting for training noise and finite-subset bias, and by predicting ensemble performance from a few attribution models. The gains depend on the noise level and on whether the target is a biased finite LDS or the score without training noise.

Acknowledgments

This work was conducted as part of the ML Alignment & Theory Scholars (MATS) Program, with support from MATS.

References

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. Training Data Attribution via Approximate Unrolling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

- Junwei Deng, Ting-Wei Li, Shichang Zhang, and Jiaqi Ma. Efficient Ensembles Improve Training Data Attribution, 2024a. ICML 2024 DMLR Workshop.
- Junwei Deng, Ting-Wei Li, Shiyuan Zhang, Shixuan Liu, Yijun Pan, Hao Huang, Xinhe Wang, Pingbang Hu, Xingjian Zhang, and Jiaqi W. Ma. *dattri*: A Library for Efficient Data Attribution. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024b.
- Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The Journey, Not the Destination: How Data Guides Diffusion Models, 2023.
- Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiuūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying Large Language Model Generalization with Influence Functions, 2023.
- Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. doi: 10.1214/aoms/1177730196.
- Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding Predictions with Data and Data with Predictions. In *International Conference on Machine Learning (ICML)*, pages 9525–9587, 2022.
- Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning (ICML)*, pages 1885–1894, 2017.
- Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing Model Behavior at Scale. In *International Conference on Machine Learning (ICML)*, 2023.
- Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Lucia Quirke, Louis Jaburi, David Johnston, William Z. Li, Gonçalo Paulo, Guillaume Martres, Girish Gupta, Nora Belrose, and Stella Biderman. Bergson: An open source library for data attribution, 2026. URL <https://arxiv.org/abs/2606.11660>.
- C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

A Proofs of the estimation error bounds

Sampling and estimator conventions. We analyse one query and a fixed attribution predictor G , with target $\rho^p = \text{Corr}_{S \sim \pi_S}(G(S), F(S))$, but the same results apply *mutatis mutandis* also to a finite attribution ensemble $\bar{g}_E$, provided Assumption 1 holds, i.e., the ensemble is independent of the evaluation subsets and retrainings.

The observations for each subset $(S_m, f_{m,1}, \dots, f_{m,R})$, $m = 1, \dots, M$, are i.i.d., with $S_m \sim \pi_S$. Conditional on each subset, the R readouts are i.i.d. with mean $F(S)$ and residual ε as in (4). Write $\sigma_G^2 = \text{Var}_S(G) > 0$, $\sigma_F^2 = \text{Var}_S(F) > 0$, $\sigma_\varepsilon^2 = \mathbb{E}[\varepsilon^2]$, $\lambda = \sigma_\varepsilon^2 / \sigma_F^2$, and $N = MR$. Assume finite fourth moments, and define

$$\kappa := \max \left\{ \frac{\mathbb{E}[(G - \mathbb{E}G)^4]}{\sigma_G^4}, \frac{\mathbb{E}[(F - \mathbb{E}F)^4]}{\sigma_F^4}, \frac{\mathbb{E}[\varepsilon^4]}{\sigma_\varepsilon^4} \right\} < \infty,$$

where we define the last ratio as zero when $\sigma_\varepsilon^2 = 0$.

We set the sample correlation to zero whenever either sample variance vanishes.

A.1 Proof of Proposition 1

Proof. Centre and standardise the variables as

$$X = \frac{G - \mathbb{E}G}{\sigma_G}, \quad Z = \frac{F - \mathbb{E}F}{\sigma_F}, \quad \eta_r = \frac{\varepsilon(S, \xi_r)}{\sigma_F}, \quad Y_R = Z + \bar{\eta}_R, \quad \bar{\eta}_R = \frac{1}{R} \sum_{r=1}^R \eta_r.$$

Then $\text{Var}(X) = \text{Var}(Z) = 1$, $\mathbb{E}[\eta_r^2] = \lambda$, and $\text{Cov}(X, Y_R) = \rho^p$. Conditional independence and zero conditional means give

$$\mathbb{E}[\bar{\eta}_R^4] = \frac{\mathbb{E}[\eta_1^4]}{R^3} + \frac{3(R-1)}{R^3} \mathbb{E}[\mathbb{E}[\eta_1^2 | S]^2] \leq \frac{3\kappa\lambda^2}{R^2}.$$

Writing $t = \lambda/R$, it follows that $\text{Var}(Y_R) = 1 + t$ and

$$\mathbb{E}[Y_R^4] \leq 8\mathbb{E}[Z^4] + 8\mathbb{E}[\bar{\eta}_R^4] \leq 24\kappa(1+t)^2.$$

Thus X and $V_R = Y_R/\sqrt{1+t}$ both have unit variance and fourth moments at most 24κ , uniformly in R .

For M i.i.d. copies of centred variables U, V with covariance c , the sample covariance $\hat{c}$ satisfies

$$\mathbb{E}[(\hat{c} - c)^2] = \frac{\mathbb{E}[U^2V^2] - c^2}{M} + \frac{\text{Var}(U)\text{Var}(V) + c^2}{M(M-1)} \leq \frac{3\sqrt{\mathbb{E}[U^4]\mathbb{E}[V^4]}}{M}. \quad (13)$$

Setting $U = V$ gives $\mathbb{E}[(\widehat{\text{Var}}_M(U) - \text{Var}(U))^2] \leq 3\mathbb{E}[U^4]/M$ for the mean squared error of the sample variance. Applying these bounds to (X, V_R) shows that the sum of the mean squared errors of its two sample variances and sample covariance is at most $C_0\kappa/M$, for a universal constant C_0 .

On the event that both sample variances are at least $1/2$, the correlation error is bounded by a universal constant times the sum of these three moment errors. To see this, use the Lipschitz continuity of $(a, b) \mapsto (ab)^{-1/2}$ on $a, b \geq 1/2$ and split off the covariance error. By Chebyshev's inequality, the complementary event has probability at most $C_1\kappa/M$; there the error is at most 2, including the specified zero convention. Hence

$$\mathbb{E}[(\hat{\rho}_{R,M}^p - \rho_R^p)^2]^{1/2} \leq C\sqrt{\kappa/M}.$$

The triangle inequality in L^2 , (7), and $1 - (1+x)^{-1/2} \leq x/2$ for $x \geq 0$ therefore give

$$\mathbb{E}[(\hat{\rho}_{R,M}^p - \rho^p)^2]^{1/2} \leq C\sqrt{\kappa}\sqrt{\frac{R}{N}} + \frac{|\rho^p|\lambda}{2R}, \quad (14)$$

where C is universal and $N = MR$. For fixed κ and $\lambda > 0$, taking $R \asymp \lambda^{2/3}N^{1/3}$ gives RMSE $\mathcal{O}(\lambda^{1/3}N^{-1/3})$, as stated in the proposition. $\square$

A.2 Proof of Proposition 2

Proof. Use the normalised variables X, Y_R, η_r and $t = \lambda/R$ from the preceding proof. Let

$$\hat{a} = \widehat{\text{Var}}_M(X), \quad \hat{c} = \widehat{\text{Cov}}_M(X, Y_R), \quad \hat{w} = \frac{1}{M} \sum_{m=1}^M s_{\eta,m}^2, \quad \hat{d} = \widehat{\text{Var}}_M(Y_R) - \hat{w}/R,$$

where $s_{\eta,m}^2$ is the within subset sample variance of the R noise replicates,

$$s_{\eta,m}^2 := \frac{1}{R-1} \sum_{r=1}^R (\eta_{m,r} - \bar{\eta}_m)^2, \quad \bar{\eta}_m := \frac{1}{R} \sum_{r=1}^R \eta_{m,r},$$

with $\eta_{m,r} = \varepsilon(S_m, \xi_{m,r})/\sigma_F$. The normalised readouts $f_{m,r}/\sigma_F$ have the same within subset sample variance, since their shared signal cancels. Thus $\mathbb{E}[\hat{a}] = \mathbb{E}[\hat{d}] = 1$, $\mathbb{E}[\hat{c}] = \rho^{\mathbb{P}}$, and $\mathbb{E}[\hat{w}] = \lambda$.

Expanding the centred sum of squares gives $s_{\eta,m}^2 = (R-1)^{-1}(\sum_r \eta_{m,r}^2 - R\bar{\eta}_m^2) \leq (R-1)^{-1} \sum_r \eta_{m,r}^2$. By Cauchy–Schwarz, $(\sum_r \eta_{m,r}^2)^2 \leq R \sum_r \eta_{m,r}^4$. Thus, for $R \geq 2$, the second moment of the sample variance satisfies $\mathbb{E}[(s_{\eta,m}^2)^2] \leq (R/(R-1))^2 \mathbb{E}[\eta_1^4] \leq 4\kappa\lambda^2$. Independence across subsets therefore implies

$$\mathbb{E}[(\hat{w}/R - t)^2] \leq \frac{4\kappa t^2}{M}.$$

Combining this with the sample-moment bound in (13), including its $U = V$ case, yields

$$\mathbb{E}[(\hat{a} - 1)^2] + \mathbb{E}[(\hat{c} - \rho^{\mathbb{P}})^2] + \mathbb{E}[(\hat{d} - 1)^2] \leq \frac{C_2\kappa(1+t)^2}{M}.$$

This step uses the inequality $(u - v)^2 \leq 2u^2 + 2v^2$ and does not require independence between $\hat{w}$ and the across-subset sample variance.

On $\{\hat{a} \geq 1/2, \hat{d} \geq 1/2\}$, the same ratio bound as above controls the error of $\hat{c}/\sqrt{\hat{a}\hat{d}}$. Projection onto $[-1, 1]$ cannot increase its distance to $\rho^{\mathbb{P}}$. On the complementary event, the returned estimate is still in $[-1, 1]$, so its squared error is at most 4. Chebyshev’s inequality consequently gives

$$\mathbb{E}[(\hat{\rho}_{R,M}^{\text{corr}} - \rho^{\mathbb{P}})^2]^{1/2} \leq C\sqrt{\kappa} \left(1 + \frac{\lambda}{R}\right) \sqrt{\frac{R}{N}}, \quad (15)$$

where we substituted $t = \lambda/R$ and $M = N/R$. For fixed $R \geq 2$ and a fixed data-generating distribution, this is $\mathcal{O}(N^{-1/2})$, proving (8). $\square$

B Derivation of the finite-subset Spearman correction

In this section we provide a derivation of the classical unbiased Spearman estimator [Hoeffding, 1948] used in Section 3.2. We then adapt it to the finite retraining-bank reference used in Fig. 1c.

Fix a query, and let (X, Y) denote the attribution prediction and model readout for a randomly drawn training subset. Consider $m \geq 3$ independent observations from this joint distribution. We assume X, Y have continuous marginals, so that ties occur with probability zero. Writing F_X and F_Y for the marginal distribution functions, the population Spearman and Kendall correlations are

$$\rho^{\text{s}} = 12 \mathbb{E}[F_X(X)F_Y(Y)] - 3 \quad \text{and} \quad \tau = \mathbb{E}[\text{sign}(X_1 - X_2) \text{sign}(Y_1 - Y_2)],$$

where the two pairs are independent copies of (X, Y) .

When given m i.i.d. observations $\{X_i, Y_i\}_{i=1}^m$, let $R_i := 1 + \sum_{j \neq i} 1(X_j < X_i)$ denote the rank of X_i (smallest first), and similarly $S_i := 1 + \sum_{j \neq i} 1(Y_j < Y_i)$.

Finite-sample bias. The finite-sample Spearman and Kendall coefficients are given by

$$\hat{\rho}^{\text{s}} = \frac{12}{m(m^2 - 1)} \sum_{i=1}^m R_i S_i - \frac{3(m+1)}{m-1}$$

and

$$\hat{\tau} = \frac{1}{\binom{m}{2}} \sum_{j < i} \text{sign}(X_i - X_j) \text{sign}(Y_i - Y_j)$$

respectively. To compute the expectation of sample Spearman, we expand $R_i S_i$. The constant and single-indicator terms contribute m in expectation. Among the remaining terms, $m - 1$ compare both coordinates against the same observation, while $(m - 1)(m - 2)$ use two distinct comparison observations:

$$\begin{aligned} \mathbb{E}[R_i S_i] &= m + (m - 1)\mathbb{P}[X_1 > X_2, Y_1 > Y_2] + (m - 1)(m - 2)\mathbb{E}[F_X(X_1)F_Y(Y_1)] \\ &= m + (m - 1) \cdot \frac{\tau + 1}{4} + (m - 1)(m - 2) \cdot \frac{\rho^s + 3}{12}, \end{aligned}$$

where the same observation terms have probability $(1 + \tau)/4$. For distinct comparison observations, conditioning on (X_i, Y_i) gives the product $F_X(X_i)F_Y(Y_i)$, whose expectation is $(\rho^s + 3)/12$.

By plugging this back into the equation for $\hat{\rho}^s$, we obtain

$$\mathbb{E}[\hat{\rho}^s] = \frac{(m - 2)\rho^s + 3\tau}{m + 1}. \quad (16)$$

This implies that the finite-sample $\hat{\rho}^s$ has a bias of $3(\tau - \rho^s)/(m + 1)$.

An unbiased estimator. Each pairwise term in sample Kendall's tau has expectation τ , so $\mathbb{E}[\hat{\tau}] = \tau$. Consequently, the estimator

$$\hat{\rho}^{s, \text{unb}} := \frac{(m + 1)\hat{\rho}^s - 3\hat{\tau}}{m - 2} \quad (17)$$

is unbiased for $m \geq 3$, i.e. $\mathbb{E}[\hat{\rho}^{s, \text{unb}}] = \rho^s$. This is the estimator in (12). By linearity of expectation, averaging the corrected estimates over queries gives an unbiased estimate of the query-averaged population Spearman correlation.

Adjustment for a finite retraining bank. For Fig. 1c, we sample m prediction/readout pairs uniformly without replacement from a fixed collection of $P = 300$ pairs, each corresponding to a training subset. There are no ties in either marginal for any query. The target is the ordinary Spearman correlation $\hat{\rho}_P^s$ computed on all P pairs. The corresponding correction is

$$\hat{\rho}_{m, P}^{s, \text{bank}} := \frac{(P - 2)\hat{\rho}_m^{s, \text{unb}} + 3\hat{\tau}_m}{P + 1}, \quad 3 \leq m \leq P. \quad (18)$$

Here, expectations are taken over the subsample draw, conditional on the fixed bank. Since $\hat{\rho}_m^{s, \text{unb}}$ and $\hat{\tau}_m$ are U-statistics of orders three and two, uniform sampling without replacement gives $\mathbb{E}[\hat{\rho}_m^{s, \text{unb}}] = \hat{\rho}_P^{s, \text{unb}}$ and $\mathbb{E}[\hat{\tau}_m] = \hat{\tau}_P$. Substituting the statistics computed on all P pairs into (18) gives $\mathbb{E}[\hat{\rho}_{m, P}^{s, \text{bank}}] = \hat{\rho}_P^s$. The same identities hold after averaging over queries.

C Estimator details

C.1 Reconstructing a finite-seed LDS

Let $\bar{f}_k$ average $k \geq 2$ retrainings per subset, and estimate $\hat{\sigma}_\varepsilon^2$ as in Algorithm 1. The observed readout variance contains σ_ε^2/k of training noise, whereas the target R -seed protocol contains σ_ε^2/R . Subtracting their estimated difference gives

$$\hat{\rho}_R^{\text{p, rec}} = \frac{\widehat{\text{Cov}}_M(G, \bar{f}_k)}{\sqrt{\widehat{\text{Var}}_M(G) \left(\widehat{\text{Var}}_M(\bar{f}_k) - (1/k - 1/R)\hat{\sigma}_\varepsilon^2 \right)}}.$$

When the estimated signal and predictor variances are positive, this is (9) with $\hat{\lambda} = \hat{\sigma}_\varepsilon^2/\hat{\sigma}_F^2$, before variance safeguards and final clipping. At $R = k$ it reduces to the ordinary sample Pearson correlation.

C.2 Attribution-ensemble correction

Let $g(S, \zeta) = G(S) + \gamma(S, \zeta)$, with $\mathbb{E}_\zeta[\gamma(S, \zeta) \mid S] = 0$, and let the attribution members be i.i.d. and independent of the evaluation subsets and retrainings. Define the noise variance across subsets and its signal ratio by

$$\sigma_\gamma^2 := \mathbb{E}_\zeta[\text{Var}_{S \sim \pi_s}(\gamma(S, \zeta))], \quad c := \frac{\sigma_\gamma^2}{\text{Var}_S(G)}.$$

The variance across subsets removes member-specific offsets, which do not affect LDS. Independence of members gives the moment identities

$$\mathbb{E}_{\zeta_{1:E}}[\text{Var}_S(\bar{g}_E)] = \text{Var}_S(G) + \frac{\sigma_\gamma^2}{E}, \quad \mathbb{E}_{\zeta_{1:E}}[\text{Cov}_S(\bar{g}_E, F)] = \text{Cov}_S(G, F).$$

The ratio formed from these expected moments is therefore

$$\tilde{\rho}_E^p = \frac{\text{Corr}_S(G, F)}{\sqrt{1 + c/E}}.$$

Comparing ensemble sizes K and E gives the main-text forecast. This approximates the correlation of a realised ensemble: increasing the number of subsets alone does not remove uncertainty from the shared ensemble members.

For $K \geq 2$ observed members, paired differences cancel the shared signal and yield

$$\hat{\sigma}_\gamma^2 := \binom{K}{2}^{-1} \sum_{i < j} \frac{\widehat{\text{Var}}_M(g(S, \zeta_i) - g(S, \zeta_j))}{2}, \quad \hat{\sigma}_G^2 := \widehat{\text{Var}}_M(\bar{g}_K) - \hat{\sigma}_\gamma^2/K.$$

When $\hat{\sigma}_G^2 > 0$, we use $\hat{c} = \hat{\sigma}_\gamma^2/\hat{\sigma}_G^2$ in the forecast.

C.3 Combining noise and finite-subset corrections

For Fig. 1a, the ensemble experiments and Figs. 6 and 7, let $\hat{s}_U$ be the query average of (12) applied to the observed prediction/readout pairs, and let $\hat{p}_{\text{corr}}$ and $\hat{p}_{\text{obs}}$ be the corresponding query-averaged corrected and ordinary Pearson correlations. With $h(p) = \frac{6}{\pi} \arcsin(p/2)$ and Π denoting projection onto $[-1, 1]$, the combined estimator is

$$\hat{\rho}^{\text{s,comb}} = h\left(\Pi\left[\frac{\hat{p}_{\text{corr}}}{\hat{p}_{\text{obs}}} h^{-1}(\Pi[\hat{s}_U])\right]\right). \quad (19)$$

In the experiments, the corrected readout variance is floored at $\widehat{\text{Var}}_M(\bar{f}_R)/4$. This safeguard is not part of Proposition 2 and can retain attenuation when $\lambda/R > 3$. The ratio factor is set to 1 when $|\hat{p}_{\text{obs}}| \leq 10^{-12}$. Averaging over queries precedes the transformations; the nonlinear combination is not guaranteed to be unbiased.

D Experimental Details

D.1 Model training and retraining banks.

We use precomputed retraining banks summarised below, which we release publicly alongside the code that produces them. Each correlation is computed per query before averaging over queries.

Dataset / model	Subsets	Seeds per subset	Queries
CIFAR-10 / ResNet-9	300	40	2,000
QNLI / BERT	200	10	2,000
WikiText-2 / GPT-2	100	10	481

CIFAR-10/ResNet-9. We follow the training setup of Park et al. [2023], using ResNet-9 and the full CIFAR-10 training set. Each retraining subset contains 25,000 images ($\alpha = 0.5$). Models are trained from scratch for 24 epochs using SGD with momentum 0.9, batch size 512, and weight decay 5×10^{-4} . The learning rate increases linearly from zero to 0.5 over the first five epochs, then decreases linearly to zero. Training uses random horizontal flips, translations, and cutout augmentation. Each subset–replica pair uses an independent training seed. For each query image, the readout is the correct-class log-odds margin.

QNLI/BERT. Following Park et al. [2023], we restrict the QNLI training pool to 50,000 examples and fine-tune pretrained bert-base-cased on subsets of 25,000 examples ($\alpha = 0.5$). We remove the pooler’s tanh activation and fine-tune for 20 epochs using SGD with batch size 8. The learning rate starts at 10^{-3} and decreases linearly to zero without warmup. Each subset–replica pair uses an independent training seed. For each query example, the readout is the correct-class log-odds margin.

WikiText-2/GPT-2 small. We follow the setup of Bae et al. [2024], as implemented in Bergson [Quirke et al., 2026]. Each retraining subset contains a random half of the 4,656 training chunks ($\alpha = 0.5$). For each subset and seed, we fine-tune GPT-2 small (124M parameters) from the base checkpoint using AdamW with a constant learning rate of 3×10^{-5} , batch size 8, weight decay 0.01, and three epochs in fp32. Dropout is enabled during training and disabled during evaluation. Each query is one of the 481 validation documents, and its readout is the negative cross-entropy loss. On this bank, our EK-FAC (ten models) attains a five-seed Spearman LDS of 0.464, close to the 0.468 ± 0.015 reported for this setup.

Following the Bergson implementation, the WikiText-2 experiment shares each retraining seed across all subsets, whereas CIFAR-10 and QNLI use independent seeds for each subset–replica pair. The pretrained GPT-2 checkpoint is fixed, while the seed controls data ordering and dropout. For a given query, the effect of a seed can include both an offset shared across subsets and a component that varies with the subset. The shared offset cancels in correlations across subsets. We therefore centre each seed’s readouts across subsets before estimating the seed-noise variance. Shared seeds can still induce dependence across subsets, whereas the bounds in Section A assume i.i.d. subset observations.

D.2 Attribution configurations

TRAK. We use TRAK [Park et al., 2023] with models trained on independently sampled halves of the training data. We featurize all model parameters, with projection dimensions 20,480 for CIFAR-10, 4,096 for QNLI, and 2,048 for MNIST and WikiText-2. The model output used for featurization is the correct-class log-odds margin for classification and the sum of per-token log-odds for WikiText-2. For TRAK- k , each attribution member combines k models by separately averaging the kernel scores and output-to-loss weights before taking their product. We then average the scores of E such members, so the total number of models is kE . The ensemble experiments use $k \in \{1, 10, 20\}$ on CIFAR-10 and $k \in \{1, 5\}$ on QNLI and WikiText-2.

EK-FAC. We compute influence function attributions using EK-FAC [Grosse et al., 2023], implemented with the `kronfluence` package⁶, at the final checkpoints of models trained on the full training set. In the case of TDA ensembles, each attribution member corresponds to one model, and we average their attribution scores. We use damping 10^{-8} and estimate the Fisher factors using labels sampled from each model’s predictive distribution, with an independent factor seed for each member. The attribution banks contain 30 members on CIFAR-10 and MNIST, 20 on QNLI, and 10 on WikiText-2. The computation includes the supported convolutional and linear layers; for GPT-2, these are the 48 transformer linear layers, excluding embeddings and LayerNorms.

SOURCE. We use SOURCE [Bae et al., 2024] with $L = 3$ segments and two checkpoints per segment on CIFAR-10 and MNIST. For CIFAR-10, the six checkpoints are saved every four epochs of a 24-epoch full-data training run. For MNIST, we use checkpoints at epochs 3, 6, 10, 13, 16, 20. We compute one attribution estimate per training trajectory and average over up to 30 trajectories. Thus, L counts the segments within one member, and E counts the trajectories being averaged.

⁶<https://github.com/pomonam/kronfluence>

TracIn. We use TracInCP [Pruthi et al., 2020], combining checkpoint contributions within each full-data training trajectory and then averaging the resulting attribution scores over trajectories. The QNLI attribution bank contains 20 trajectories with checkpoints at epochs 4, 8, 12, 16, 20. We project gradients over all parameters to dimension 1,024 and weight checkpoint contributions according to the learning-rate schedule over the corresponding intervals. On WikiText-2, we use the same ten training runs and 48 transformer linear layers as EK-FAC, with six half-epoch checkpoints per run and a fresh 16,384-dimensional projection at each checkpoint.

D.3 Evaluation methodology

This section describes the estimators and reference scores for noise-corrected LDS, finite-seed LDS, and ordinary Spearman LDS on a fixed bank. For the noise-correction and finite-seed reconstruction benchmarks, we partition the retraining seeds into disjoint reference and estimation sets. Reference scores use all training subsets and only the reference seeds; the evaluated estimators use only the estimation seeds. The subset bank, queries, and attribution ensemble are shared.

Noise-corrected LDS estimation. Figure 2(a, d, g) compares naive and corrected Pearson LDS as the number of retraining seeds varies, using a single attribution member ($E = 1$) and all subsets. We average over 50 random seed splits, with 20 reference and 20 estimation seeds on CIFAR-10, and 2 reference and 8 estimation seeds on QNLI and WikiText-2. From the estimation set, we sample $R \in \{1, 2, 3, 5, 10, 20\}$ seeds on CIFAR-10 and $R \in \{1, 2, 3, 5, 8\}$ on QNLI and WikiText-2, without replacement. The noise correction uses these selected seeds for $R \geq 2$. For each split, we compute the attenuation curve from (7), using the corrected Pearson correlation and noise-to-signal ratio estimated from the reference seeds for each query. Both the estimates and the curves are averaged over queries and seed splits.

In Figs. 1a, 6 and 7, we evaluate the error in estimating the noise-corrected Spearman LDS. We compare ordinary sample Spearman with the combined correction (19). The reference applies the combined estimator to all subsets and the reference seeds, using the same fixed attribution ensemble as the evaluated estimates. This ensemble consists of all 30 CIFAR-10 EK-FAC members in Fig. 1a, and the first 10 members of each method in Figs. 6 and 7.

For each of 200 repetitions, we split the seeds into equal reference and estimation sets: 20/20 on CIFAR-10 and 5/5 on QNLI and WikiText-2. At each budget $N = mR$, we sample m of the bank’s subsets and R seeds from the estimation set, uniformly without replacement. The base grid is $m \in \{5, 8, 12, 25, 50, 100, 200, 300\}$, capped at the bank size, with additional points at $m = 2$ for naive estimates and $m = 3$ for corrected estimates, retaining budgets $mR \geq 9$. Naive estimates use $R = 5, 10, 20$ on CIFAR-10 and $R = 5$ on QNLI and WikiText-2; corrected estimates use $R = 2, 5$ throughout. In the appendix experiments, all attribution methods within a dataset use the same splits and draws. Figure 6 reports RMSE against each repetition’s reference, and Fig. 7 shows the first repetition with its corresponding reference.

Writing $\hat{\rho}_{\text{ref}}^s$ for a split’s reference score, $\hat{\lambda}$ for the median per-query noise-to-signal ratio estimated from its reference seeds, and $h(r) = (6/\pi) \arcsin(r/2)$, the predicted finite-seed plateau is

$$h\left(\frac{h^{-1}(\hat{\rho}_{\text{ref}}^s)}{\sqrt{1 + \hat{\lambda}/R}}\right).$$

The predicted RMSE floor is the absolute difference between this plateau and the reference score. We average these floors over seed splits for the RMSE plots; the single-draw plot uses its own split’s reference and plateaus. Variability in the reference estimates contributes to the reported RMSE.

Finite-seed LDS reconstruction. In Fig. 2(b, e, h), the target is the standard five-seed Spearman LDS, which we estimate using either naive five-seed averaging or reconstruction from $k = 2$ seeds. These experiments use a single attribution member ($E = 1$). The reconstructed estimate applies (9) to each query, converts the result to Spearman using (11), and averages over queries.

We use 30 random seed splits into equal reference and estimation sets (20/20 on CIFAR-10 and 5/5 on QNLI and WikiText-2), cycling through these splits over 300 repetitions. For each split, the reference is evaluated on all subsets: on CIFAR-10, it averages Spearman LDS over 40 random

groups of five reference seeds; on QNLI and WikiText-2, it uses the single group of five reference seeds. Each repetition samples subsets and estimation seeds without replacement and is scored against its split’s reference. The subset counts are $m \in \{5, 10, 20, 40, 60, 100\}$ for CIFAR-10 and QNLI, and $m \in \{5, 10, 20, 30, 50\}$ for WikiText-2. The budgets are $N = 5m$ for naive averaging and $N = 2m$ for reconstruction. We compute the reported budget ratios from log–log fits at RMSE 0.01, except for QNLI/TracIn, where we use 0.0134, the lowest RMSE reached by both curves.

Ensemble forecasting. In Fig. 1b and Fig. 2(c, f, i), we compare forecasts based on a small attribution ensemble with measured noise-corrected LDS as the ensemble size increases. We use all subsets and retraining seeds in each retraining bank. The forecast uses the first K attribution members. We map their combined Spearman estimate (19) to the Pearson scale using h^{-1} , apply (10), and convert the forecast back using h . We estimate the attribution noise-to-signal ratio from differences between members as described in Section C.2, and take the median of the finite per-query estimates of c .

We consider $K \in \{2, 3, 4, 5, 8, 15\}$ and select the smallest available value for which at most 15% of queries have a nonpositive estimated attribution-signal variance. This gives $K = 2$ except for CIFAR-10/TRAK-1, where the criterion selects $K = 15$. For this method, we show the forecast from $K = 5$ to illustrate the error when the initial ensemble is too small. The measured Spearman scores use (19). In both figures, the measured curve uses the first E members of the same sequence as the forecast, following one attribution ensemble as it grows.

Finite-subset debiasing. In Fig. 1c, we fix the predictions from 100 TRAK-1 members and average the readouts over all 40 retraining seeds. We then sample subsets from the 300-subset CIFAR-10 bank and compare ordinary Spearman with the finite-bank correction (18). The reference is ordinary Spearman computed on all 300 subsets, averaged over the 2,000 queries. We use $m \in \{4, 5, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128\}$, with 2,500 uniform draws without replacement at each subset count. Both estimators use the same draws. Relative RMSE is obtained by dividing RMSE by the absolute reference LDS.

Figure 4 decomposes the same squared error into squared bias and variance. In Fig. 8, we use the same subset draws and vary the number of queries, taking $Q = 1, 100, 1000$ from nested prefixes of one random permutation of the queries. Each reference score is computed over the corresponding query set.

Figure 9 extends the comparison to TRAK-1 with $E = 100$, EK-FAC with $E = 30$, and SOURCE with $L = 3$ and $E = 30$, on CIFAR-10 and MNIST. The MNIST bank contains 300 subsets, 40 retraining seeds per subset, and 2,000 queries. It uses a training pool of 6,144 examples, half-sized retraining subsets, and an MLP with three hidden layers of widths 512, 256, 128. The models are trained for 20 epochs using SGD with momentum 0.9, learning rate 0.03, batch size 64, and weight decay 0.001. For both datasets, we fix the first five retraining seeds and use 600 subset draws at each value of m in the same grid as Fig. 1c. We compare ordinary Spearman with (18), using ordinary Spearman on all 300 subsets with these fixed five seeds as the reference.

E Analysis of the Pearson-Spearman conversion

We assess the accuracy of the Pearson–Spearman conversion (11) in a separate reconstruction diagnostic on CIFAR-10 and QNLI, with $E = 1$. Five of the six estimated reconstruction biases are smaller in magnitude than one single-estimate standard error.

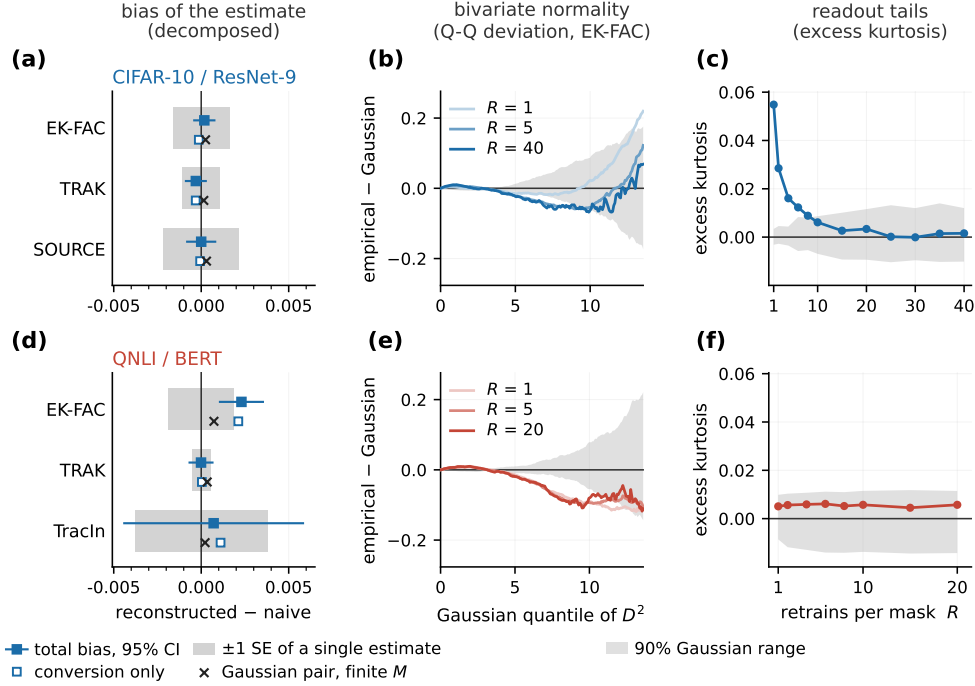


Figure 5: **The arcsin bridge in detail.** Top: CIFAR-10 / ResNet-9; bottom: QNLI / BERT; single attribution members ($E = 1$). (a, d) Bias of the reconstructed estimate (filled squares, 95% CI), its conversion component (open squares), and the corresponding finite-mask Gaussian baseline (crosses); grey: ± 1 SE of a single estimate. (b, e) Q-Q deviation of D^2 for EK-FAC, with the readout averaged over R retrains per mask; 0 denotes the Gaussian reference. (c, f) Excess kurtosis of the readout relative to its finite-mask Gaussian expectation, against R . Grey in (b, c, e, f): 90% Gaussian reference range.

Reconstruction and conversion error (a, d). We estimate reconstruction bias as the difference between the two-seed reconstruction and the naive five-seed Spearman score, evaluated on the full subset bank. To isolate conversion error, we compare the transformed five-seed sample Pearson correlation with sample Spearman, averaging over queries. These statistics differ even under bivariate normality at finite M . The largest observed magnitude of the additional discrepancy relative to this Gaussian baseline is 0.0012. QNLI EK-FAC is the exception, with reconstruction bias +0.0026 (1.3 standard errors); most of its conversion discrepancy is also present in the Gaussian baseline.

Distributional diagnostics (b, c, e, f). On CIFAR-10, the heavier upper tail in the Q-Q diagnostic and the excess readout kurtosis diminish as retraining seeds are averaged. QNLI retains Q-Q departures, while its readout kurtosis remains within the Gaussian reference range. The distance diagnostic assesses only the radial distribution; neither diagnostic establishes full bivariate normality or Gaussianity of the noise-free pair.

Diagnostic protocol. We average the reconstruction over 200 random two-seed draws and the naive score over 40 random five-seed subsets. Both sample from the full seed pool without reserving reference seeds. The TRAK diagnostic uses TRAK-1. The bias estimate is conditional on the subset bank; its 95% confidence interval uses a 20-block delete-a-block jackknife over subsets. The single-estimate standard error is measured by the run-to-run spread over two-seed draws. For the Q-Q diagnostic, D^2 is the squared Mahalanobis distance of each attribution/readout pair from its query’s sample mean, using sample covariance normalized by $M - 1$. Under independent bivariate-normal sampling,

$$\frac{MD^2}{(M-1)^2} \sim \text{Beta}\left(1, \frac{M-3}{2}\right).$$

We plot empirical minus Gaussian quantiles of D^2 , pooled over queries, for $R = 1, 5$, and all available seeds. The 90% Q-Q reference range uses 40 random orthogonal rotations of the five-seed

pairs within the centred subset subspace, preserving each query’s mean and covariance. Readout excess kurtosis is averaged over queries and centred by its finite- M Gaussian expectation.

F Additional Results for Noise Correction

Figures 6 and 7 show the effect of combining training-noise and finite-subset corrections on CIFAR-10, QNLI and WikiText-2 with $E = 10$ attribution members. Here, we compare estimates against a noise-corrected Spearman reference computed from held-out retraining seeds, whereas Fig. 2(b, e, h) targets the biased five-seed score. Correction with $R = 2$ or $R = 5$ reduces RMSE as the retraining budget grows. The naive attenuation floors are largest on CIFAR-10; on QNLI and WikiText-2, the smaller floors make sampling variability more prominent. The reference and sampling procedure are described in Section D.

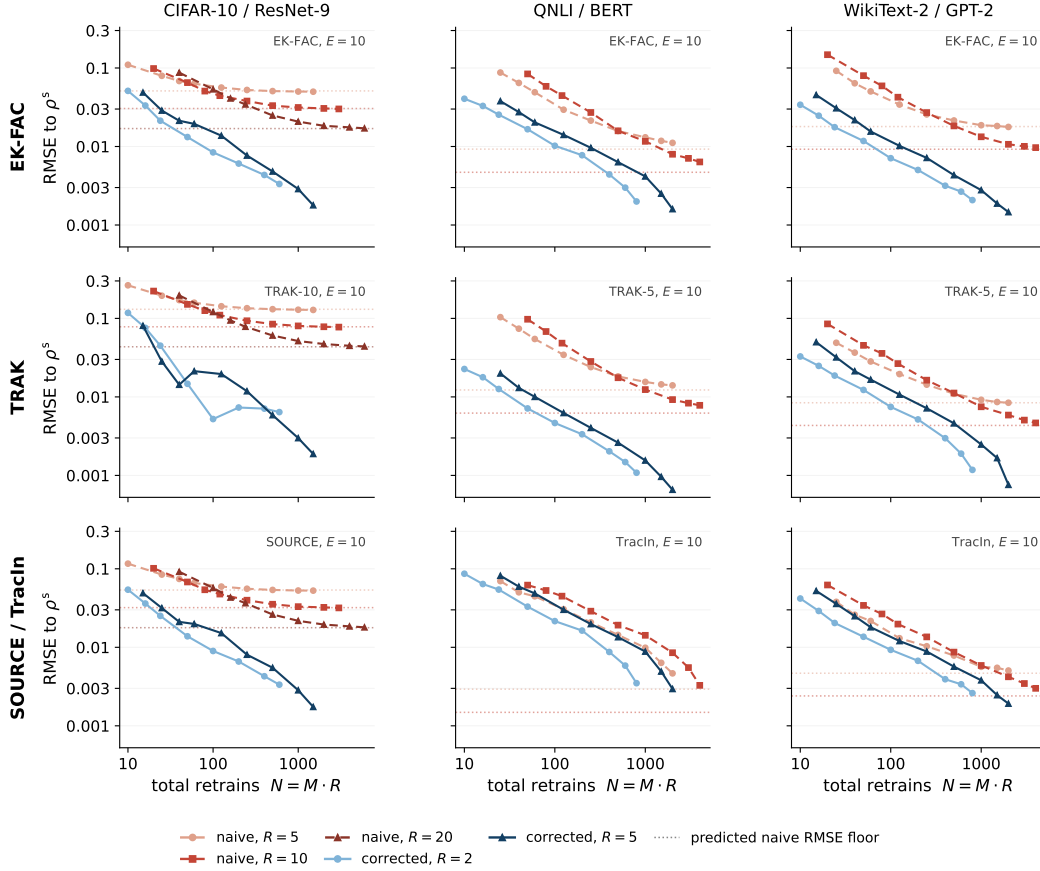


Figure 6: **RMSE of corrected Spearman LDS.** Error to references computed from held-out retraining seeds over 200 draws, with $E = 10$. Columns: CIFAR-10, QNLI, WikiText-2. Rows: EK-FAC, TRAK, SOURCE/TracIn. Dashed: naive Spearman; solid: combined training-noise and finite-subset correction; dotted: predicted naive RMSE floors.

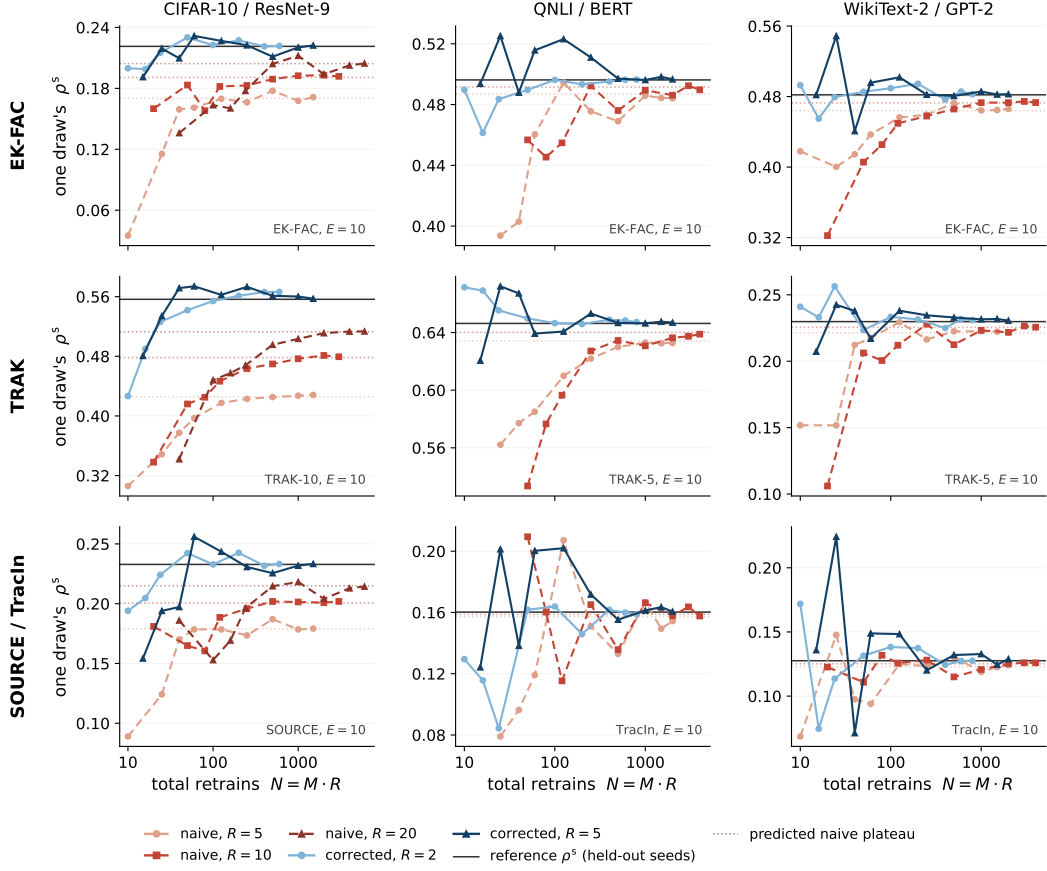


Figure 7: **Spearman LDS estimates from a single draw.** The first of the 200 draws in Fig. 6, with the same $E = 10$ ensembles and estimators. Black lines: reference computed from the first split’s held-out retraining seeds; dotted lines: predicted naive plateaus from that split.

F.1 Additional Results for Finite-Sample Debiasing

This section contains additional figures showing the bias and variance in estimating the LDS, and the effect of the unbiased U-statistic estimator.

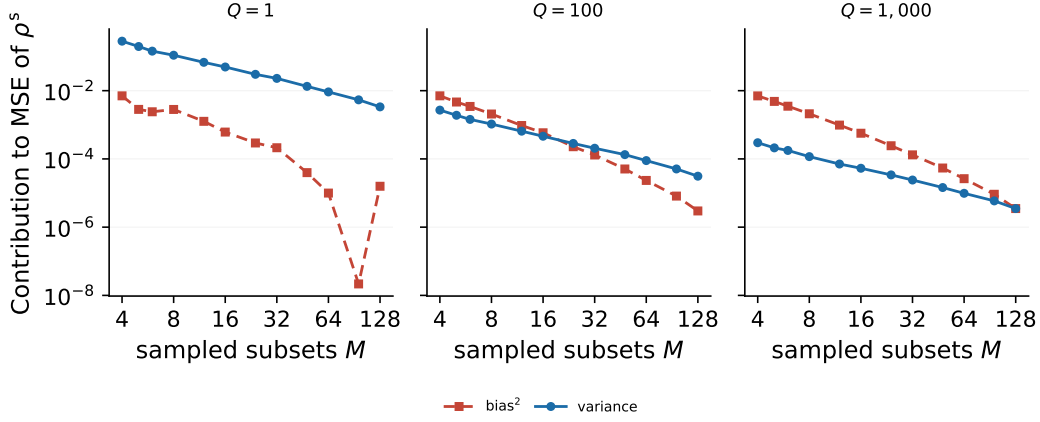


Figure 8: Comparing the bias (squared) and variance of estimating ρ^s using the finite-sample Spearman correlation estimator $\hat{\rho}^s$, across different numbers of query points Q .

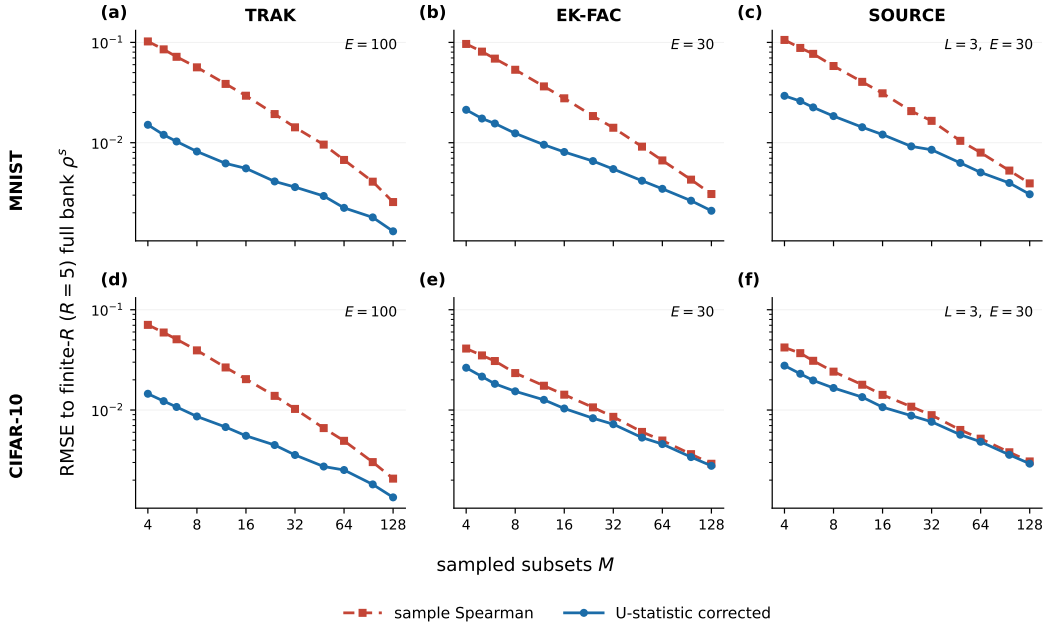


Figure 9: Comparing RMSE in estimating the full-bank LDS using finite-sample Spearman and the finite-bank correction (18). All experiments use $R = 5$ retraining seeds.