

Efficient and Accurate Causal Validation of Training Data Attribution

Simon Vary

MATS Research

joint work with Alireza Mousavi-Hosseini and Roger Grosse

October 30, 2026

Which training data mattered, and how can we check?

Slides: simonvary.com/apollo26.pdf

Paper: simonvary.com/corrected_lds.pdf

Plan for today

- 1 What is training data attribution (TDA)?
- 2 How do people evaluate how well TDA methods work?
- Our contribution
3 **How can we evaluate TDA using less compute?**

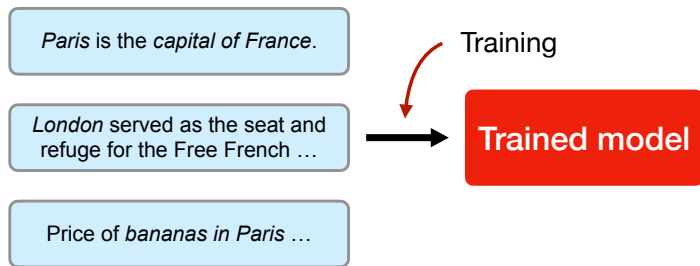
Training Data Attribution (TDA)

Paris is the capital of France.

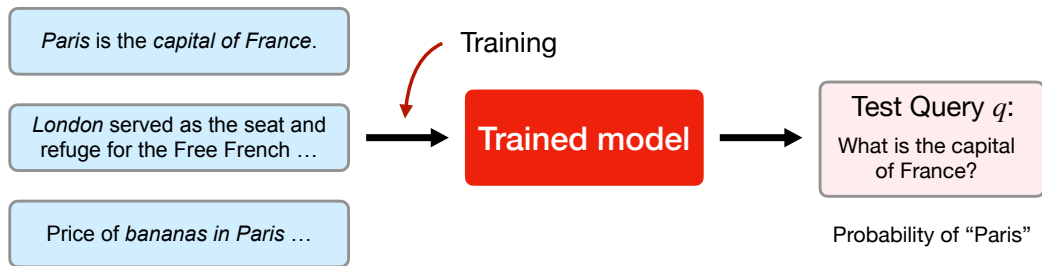
*London served as the seat and
refuge for the Free French ...*

Price of bananas in Paris ...

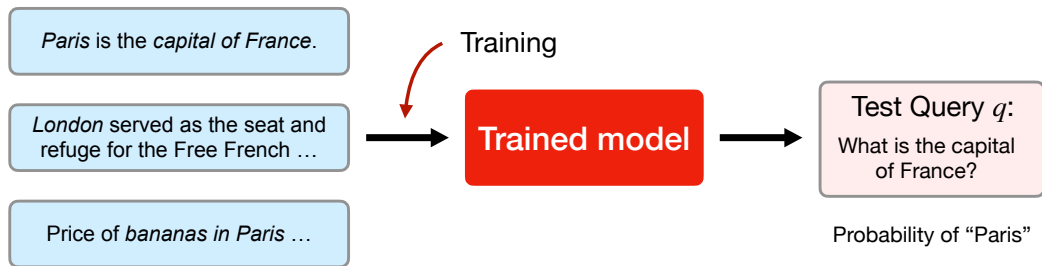
Training Data Attribution (TDA)



Training Data Attribution (TDA)

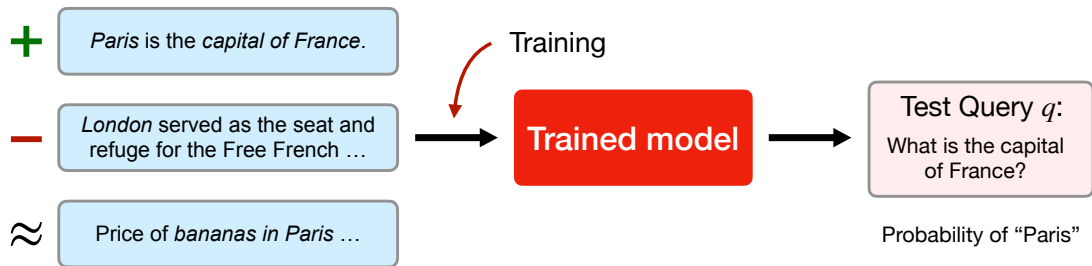


Training Data Attribution (TDA)



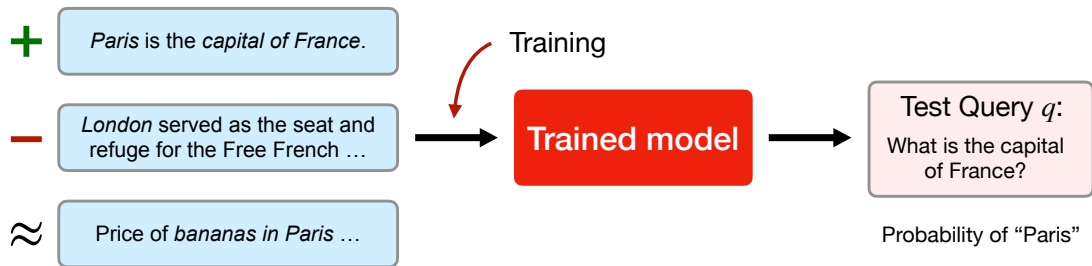
Which training examples made "Paris" more likely?

Training Data Attribution (TDA)



Which training examples made "Paris" more likely?

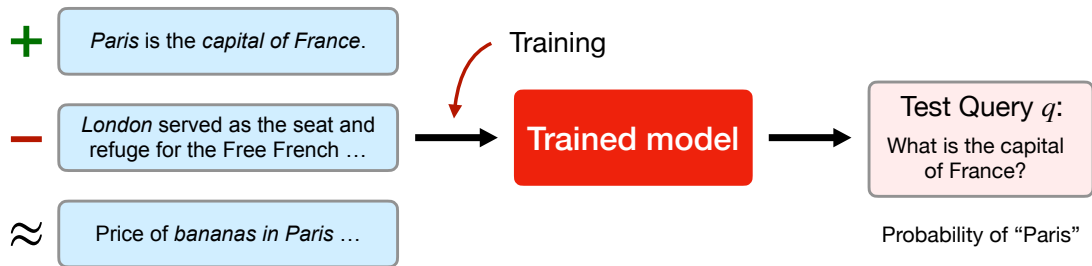
Training Data Attribution (TDA)



Which training examples made “Paris” more likely?

Say $F_q(D \setminus \text{Paris is the capital of France.})$ is the probability of “Paris” after training without Paris is the capital of France.

Training Data Attribution (TDA)



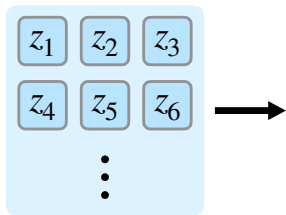
Which training examples made "Paris" more likely?

Say $F_q(D \setminus \text{Paris is the capital of France.})$ is the probability of "Paris" after training without Paris is the capital of France.

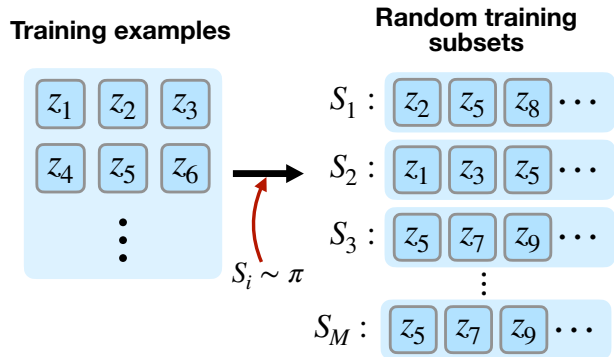
Attribution G_q estimates $F_q(D \setminus \text{Paris is the capital of France.}) \approx G_q(D \setminus \text{Paris is the capital of France.})$.

Checking the validity of attribution

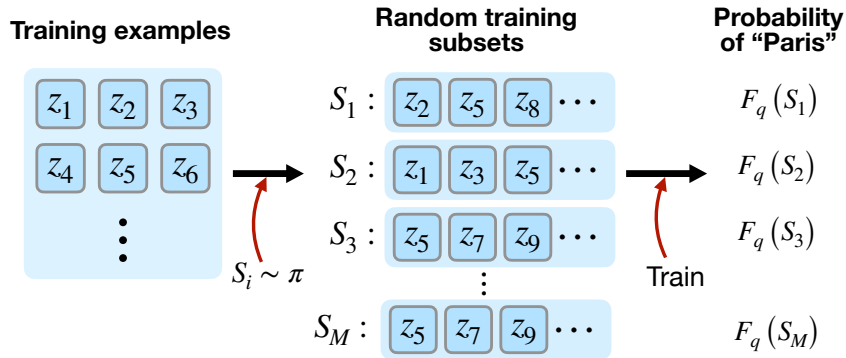
Training examples



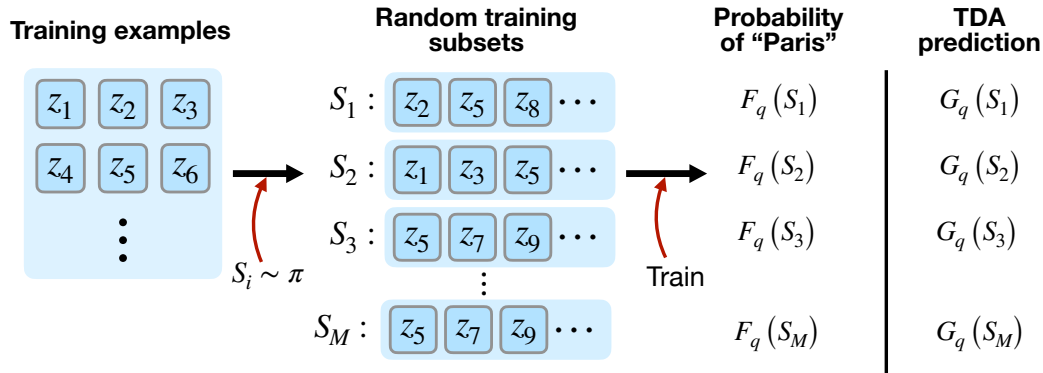
Checking the validity of attribution



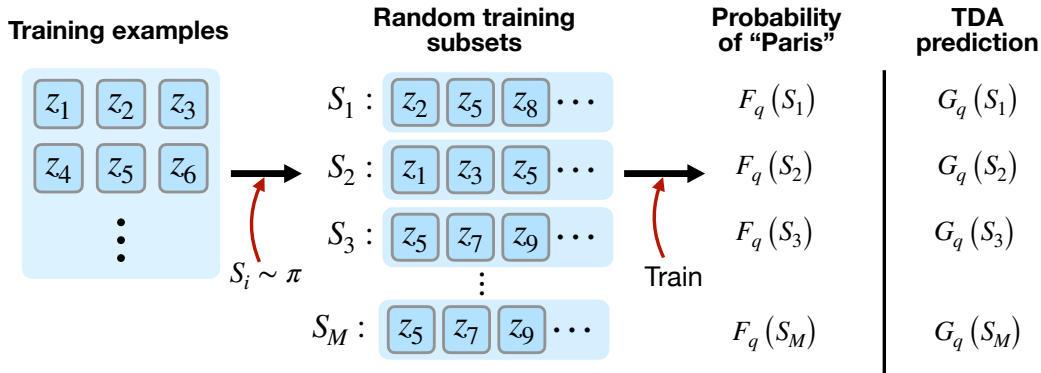
Checking the validity of attribution



Checking the validity of attribution

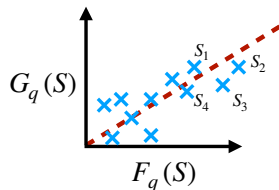


Checking the validity of attribution



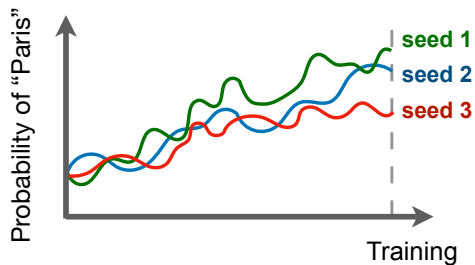
Linear Datamodeling Score

$$\text{LDS} = \frac{1}{Q} \sum_q \text{Corr}_{S \sim \pi} \left(G_q(S), F_q(S) \right)$$



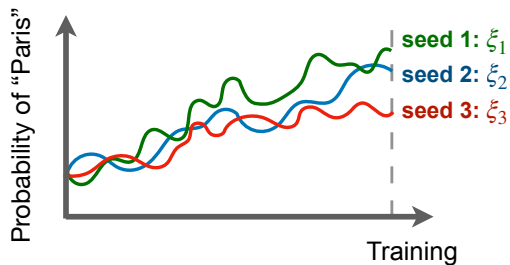
... but, the same data can produce different predictions!

Even for the same training set S :



... but, the same data can produce different predictions!

Even for the same training set S :

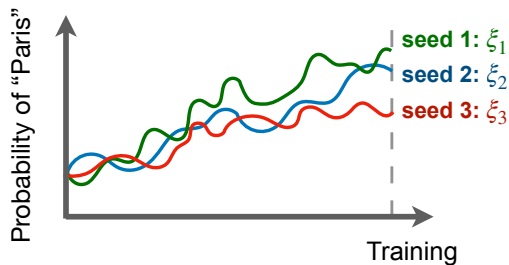


Instead of $F(S)$ we observe

$$f(S, \xi) = F(S) + \varepsilon(S, \xi)$$

... but, the same data can produce different predictions!

Even for the same training set S :



Instead of $F(S)$ we observe

$$f(S, \xi) = F(S) + \varepsilon(S, \xi)$$

Expected
data effect

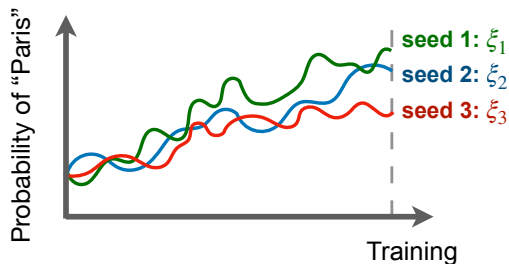
Training noise

Source of
randomness

(data ordering, initialisation, ...)

...but, the same data can produce different predictions!

Even for the same training set S :



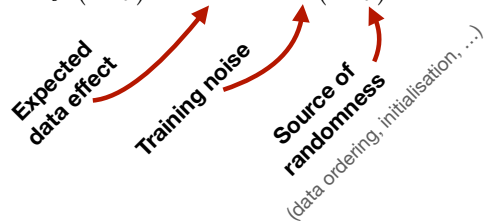
Standard methodology (Park et al., 2023) :

Average out R -seeds!

$$F^{(R)}(S) = \frac{1}{R} \sum_{i=1}^R f(S, \xi_i)$$

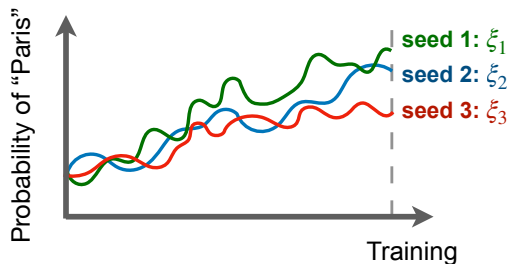
Instead of $F(S)$ we observe

$$f(S, \xi) = F(S) + \varepsilon(S, \xi)$$



...but, the same data can produce different predictions!

Even for the same training set S :



Standard methodology (Park et al., 2023) :

Average out R -seeds!

$$F^{(R)}(S) = \frac{1}{R} \sum_{i=1}^R f(S, \xi_i)$$

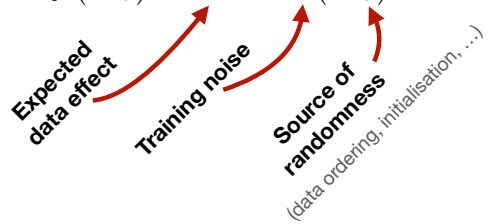
+

M i.i.d. subset samples

$$S_1, S_2, \dots, S_M \sim \pi$$

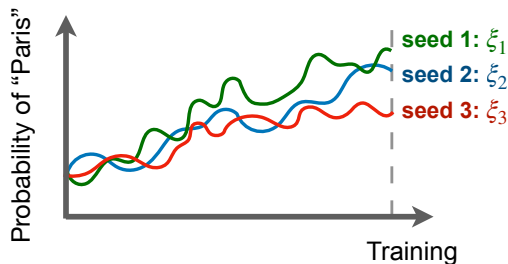
Instead of $F(S)$ we observe

$$f(S, \xi) = F(S) + \varepsilon(S, \xi)$$



...but, the same data can produce different predictions!

Even for the same training set S :



Instead of $F(S)$ we observe

$$f(S, \xi) = F(S) + \varepsilon(S, \xi)$$

Expected data effect

Training noise

Source of randomness
(data ordering, initialisation, ...)

Standard methodology (Park et al., 2023) :

Average out R -seeds!

$$F^{(R)}(S) = \frac{1}{R} \sum_{i=1}^R f(S, \xi_i)$$

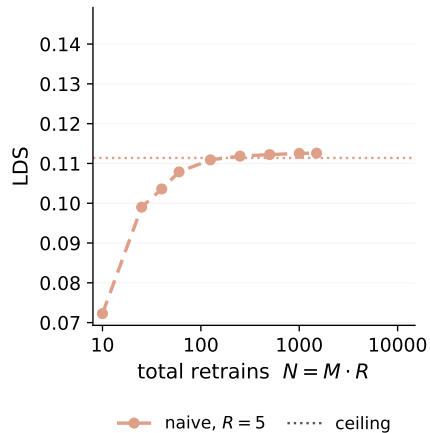
M i.i.d. subset samples

$$S_1, S_2, \dots, S_M \sim \pi$$

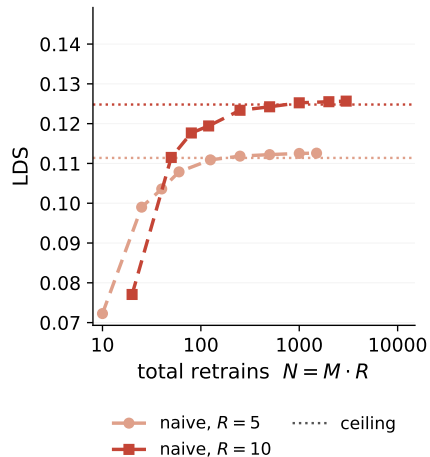
$$\Rightarrow \widehat{\text{LDS}}_{R,M} = \text{Corr}_{S \sim \hat{\pi}_M} (G(S), F^{(R)}(S))$$

Practice: $M = 100$ and $R = 5, 10, 20$

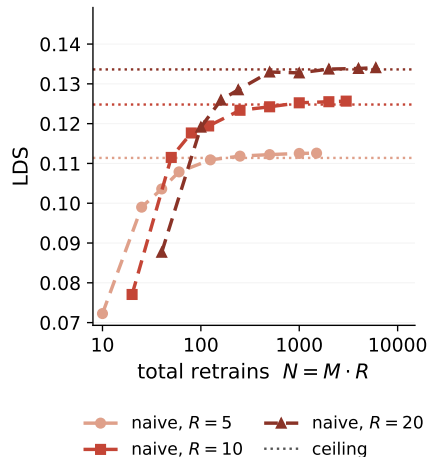
Standard LDS estimation for a fixed R retrains per subset



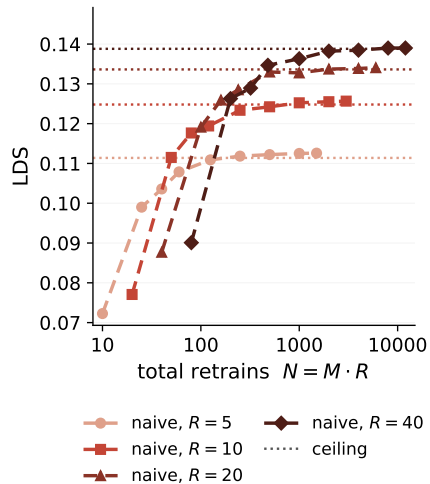
Standard LDS estimation for a fixed R retrains per subset



Standard LDS estimation for a fixed R retrains per subset

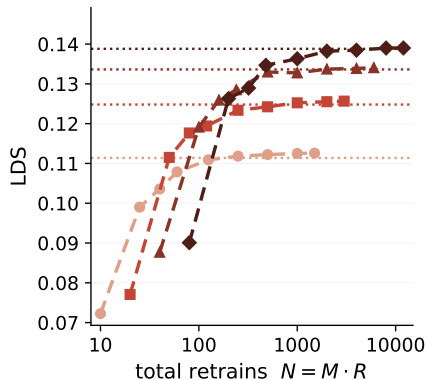


Standard LDS estimation for a fixed R retrains per subset



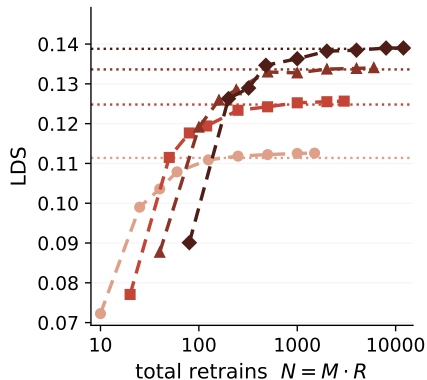
Standard LDS estimation for a fixed R retrains per subset

$$\widehat{\text{LDS}}_{R,M} = \frac{\text{Cov}_S(G, F^{(R)})}{\sqrt{\text{Var}_S(G) \text{Var}_S(F^{(R)})}}$$



- naive, $R=5$
- naive, $R=10$
- naive, $R=20$
- naive, $R=40$
- ceiling

Standard LDS estimation for a fixed R retrains per subset

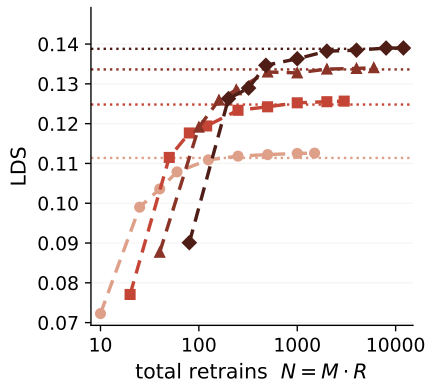


- naive, $R = 5$
- naive, $R = 10$
- ▲— naive, $R = 20$
- ◆— naive, $R = 40$
- ceiling

$$\begin{aligned}\widehat{\text{LDS}}_{R,M} &= \frac{\text{Cov}_S(G, F^{(R)})}{\sqrt{\text{Var}_S(G) \text{Var}_S(F^{(R)})}} \\ &= \frac{\text{Cov}_S(G, F)}{\sqrt{\text{Var}_S(G) \left(\underbrace{\text{Var}(F)}_{\text{data signal}} + \underbrace{\frac{\text{Var}(\epsilon)}{R}}_{\text{training noise}} \right) }}\end{aligned}$$

- $\text{Cov}(G, \epsilon) = 0 \Rightarrow \text{Cov}_S(G, F^{(R)}) = \text{Cov}_S(G, F)$.

Standard LDS estimation for a fixed R retrains per subset

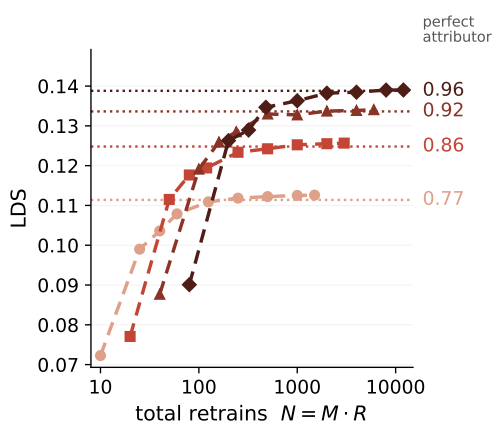


- naive, $R = 5$
- naive, $R = 10$
- ▲— naive, $R = 20$
- ◆— naive, $R = 40$
- ceiling

$$\begin{aligned}
 \widehat{\text{LDS}}_{R,M} &= \frac{\text{Cov}_S(G, F^{(R)})}{\sqrt{\text{Var}_S(G) \text{Var}_S(F^{(R)})}} \\
 &= \frac{\text{Cov}_S(G, F)}{\sqrt{\text{Var}_S(G) \left(\underbrace{\text{Var}(F)}_{\text{data signal}} + \underbrace{\frac{\text{Var}(\epsilon)}{R}}_{\text{training noise}} \right)}} \\
 &= \underbrace{\text{Corr}_S(G, F)}_{\text{true LDS}} \cdot \underbrace{\frac{1}{\sqrt{1 + \lambda/R}}}_{\text{noise factor}}.
 \end{aligned}$$

- $\text{Cov}(G, \epsilon) = 0 \Rightarrow \text{Cov}_S(G, F^{(R)}) = \text{Cov}_S(G, F)$.
- $\lambda = \text{Var}(\epsilon)/\text{Var}(F)$: noise-to-signal ratio.

Standard LDS estimation for a fixed R retrains per subset

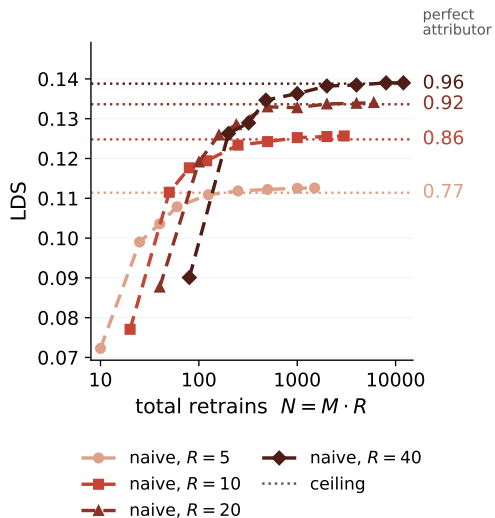


- naive, $R = 5$
- naive, $R = 10$
- ▲ naive, $R = 20$
- ◆ naive, $R = 40$
- ceiling

$$\begin{aligned}
 \widehat{\text{LDS}}_{R,M} &= \frac{\text{Cov}_S(G, F^{(R)})}{\sqrt{\text{Var}_S(G) \text{Var}_S(F^{(R)})}} \\
 &= \frac{\text{Cov}_S(G, F)}{\sqrt{\text{Var}_S(G) \left(\underbrace{\text{Var}(F)}_{\text{data signal}} + \underbrace{\frac{\text{Var}(\epsilon)}{R}}_{\text{training noise}} \right)}} \\
 &= \underbrace{\text{Corr}_S(G, F)}_{\text{true LDS}} \cdot \underbrace{\frac{1}{\sqrt{1 + \lambda/R}}}_{\text{noise factor}}.
 \end{aligned}$$

- $\text{Cov}(G, \epsilon) = 0 \Rightarrow \text{Cov}_S(G, F^{(R)}) = \text{Cov}_S(G, F)$.
- $\lambda = \text{Var}(\epsilon)/\text{Var}(F)$: noise-to-signal ratio.
- More subsets M do not lift this ceiling.

We correct LDS for training noise



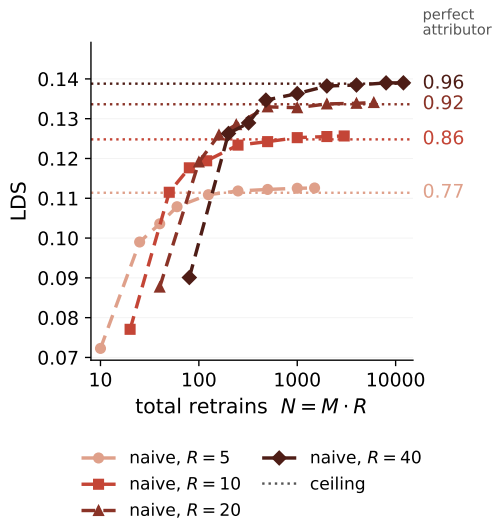
1. Estimate training noise

$$\widehat{\text{Var}}(\epsilon) = \frac{1}{M} \sum_{m=1}^M \widehat{\text{Var}}_{\xi}[f(S_m, \xi)]$$

Across seeds, then average across subsets.

Correction for attenuation (Spearman, 1904).

We correct LDS for training noise



1. Estimate training noise

$$\widehat{\text{Var}}(\epsilon) = \frac{1}{M} \sum_{m=1}^M \widehat{\text{Var}}_{\xi}[f(S_m, \xi)]$$

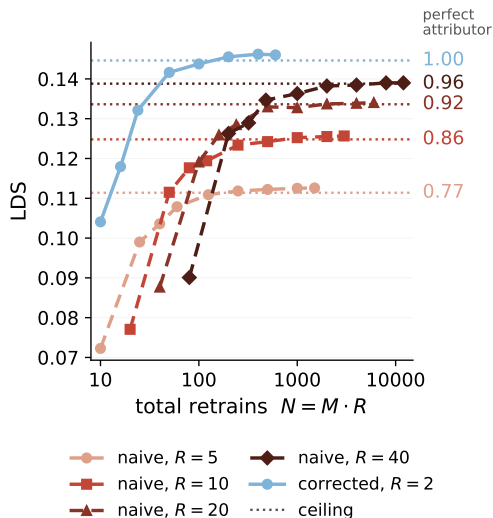
Across seeds, then average across subsets.

2. Estimate the data signal

$$\widehat{\text{Var}}(F) = \widehat{\text{Var}}_S(F^{(R)}) - \frac{\widehat{\text{Var}}(\epsilon)}{R}$$

Correction for attenuation (Spearman, 1904).

We correct LDS for training noise



1. Estimate training noise

$$\widehat{\text{Var}}(\varepsilon) = \frac{1}{M} \sum_{m=1}^M \widehat{\text{Var}}_{\xi}[f(S_m, \xi)]$$

Across seeds, then average across subsets.

2. Estimate the data signal

$$\widehat{\text{Var}}(F) = \widehat{\text{Var}}_S(F^{(R)}) - \frac{\widehat{\text{Var}}(\varepsilon)}{R}$$

3. Correct the LDS

$$\widehat{\text{LDS}}_{\text{corr}} = \widehat{\text{LDS}}_{R,M} \cdot \sqrt{1 + \hat{\lambda}/R}$$

$$\hat{\lambda} = \widehat{\text{Var}}(\varepsilon) / \widehat{\text{Var}}(F)$$

Just $R = 2$ runs per subset.

Correction for attenuation (Spearman, 1904).

Conclusions

Main takeaways

- **Training noise** limits measured LDS, even for a perfect attributor.
- Sampling more subsets does not remove this bias.
- Correct LDS using only **two retrainings per subset**.

Next steps / questions

- How much of this translates to downstream tasks?
- Other metrics reflecting the magnitude of an intervention?

Paper: simonvay.com/corrected_lds.pdf

Code: github.com/simonvay/corrected-lds

Data: huggingface.co/datasets/corrected-lds/data

Backup 1: Estimating the biased $R = 5$ seed LDS with fewer retrainings

Estimate from two runs

$$\underbrace{\hat{\sigma}_F^2}_{\text{data signal}} \quad \underbrace{\hat{\sigma}_\epsilon^2}_{\text{training noise}}$$

Reconstruct five-seed variance

$$\widehat{\text{Var}}_S(F^{(5)}) = \hat{\sigma}_F^2 + \frac{\hat{\sigma}_\epsilon^2}{5}$$

Backup 1: Estimating the biased $R = 5$ seed LDS with fewer retrainings

Estimate from two runs

$\hat{\sigma}_F^2$
data signal

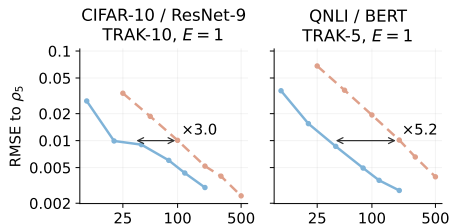
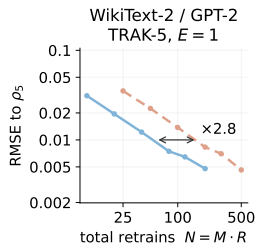
$\hat{\sigma}_\epsilon^2$
training noise

Reconstruct five-seed variance

$$\widehat{\text{Var}}_S(F^{(5)}) = \hat{\sigma}_F^2 + \frac{\hat{\sigma}_\epsilon^2}{5}$$

3–5× fewer retrainings

across CIFAR-10, QNLI and WikiText-2



— naive, $R = 5$ — reconstructed, $R = 2$

Backup 2: Forecasting larger attribution ensembles

$$g(S, \zeta) = \underbrace{G(S)}_{\text{mean}} + \underbrace{\gamma(S, \zeta)}_{\text{zero-mean noise}}$$

$$G^{(E)}(S) = \frac{1}{E} \sum_{i=1}^E g(S, \zeta_i)$$

E independent members.

$$\hat{c} = \underbrace{\hat{\sigma}_{\gamma}^2}_{\text{attribution noise}} / \underbrace{\hat{\sigma}_G^2}_{\text{data signal}}$$

Backup 2: Forecasting larger attribution ensembles

$$g(S, \zeta) = \underbrace{G(S)}_{\text{mean}} + \underbrace{\gamma(S, \zeta)}_{\text{zero-mean noise}}$$

$$G^{(E)}(S) = \frac{1}{E} \sum_{i=1}^E g(S, \zeta_i)$$

E independent members.

$$\hat{c} = \underbrace{\hat{\sigma}_{\gamma}^2}_{\text{attribution noise}} / \underbrace{\hat{\sigma}_G^2}_{\text{data signal}}$$

Forecast from two members

$$\hat{\rho}_E^{\text{forec}} = \widehat{\text{Corr}}(G^{(2)}, F) \sqrt{\frac{1 + \hat{c}/2}{1 + \hat{c}/E}}$$

$\widehat{\text{Corr}}$: corrected Pearson LDS.

Backup 2: Forecasting larger attribution ensembles

$$g(S, \zeta) = \underbrace{G(S)}_{\text{mean}} + \underbrace{\gamma(S, \zeta)}_{\text{zero-mean noise}}$$

$$G^{(E)}(S) = \frac{1}{E} \sum_{i=1}^E g(S, \zeta_i)$$

E independent members.

$$\hat{c} = \underbrace{\hat{\sigma}_{\gamma}^2}_{\text{attribution noise}} / \underbrace{\hat{\sigma}_G^2}_{\text{data signal}}$$

Forecast from two members

$$\hat{\rho}_E^{\text{forec}} = \widehat{\text{Corr}}(G^{(2)}, F) \sqrt{\frac{1 + \hat{c}/2}{1 + \hat{c}/E}}$$

$\widehat{\text{Corr}}$: corrected Pearson LDS.

