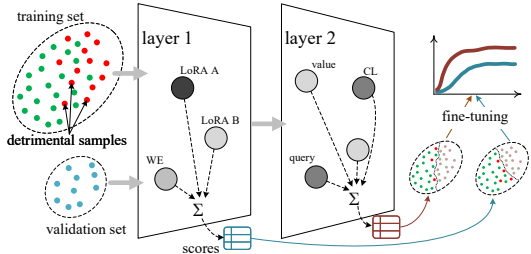
UNIVERSITY of
SOUTH FLORIDA
Bellini College of Artificial Intelligence,
Cybersecurity and Computing

First is Not Really Better Than Last: Evaluating Layer Choice and Aggregation Strategies in Language Model Data Influence Estimation

Dmytro Vitel, Anshuman Chhabra
{dvitel,anshumanc}@usf.edu

March 30, 2026

Research Questions



1. Is the gradient **cancellation effect**[1] **a reliable predictor** of layer contribution in influence estimation?
2. Which model **layers** yield the **most effective** influence scores for detecting detrimental samples?
3. Can **alternative aggregation strategies** improve influence estimation performance compared to traditional averaging?
4. How reliably can the **detrimental sample distribution measures** predict the downstream performance of influence-based data filtering methods?

Yeh et al. "First is better than last for language data influence", NeurIPS 2025

[RQ1] Issues with Cancellation Effect

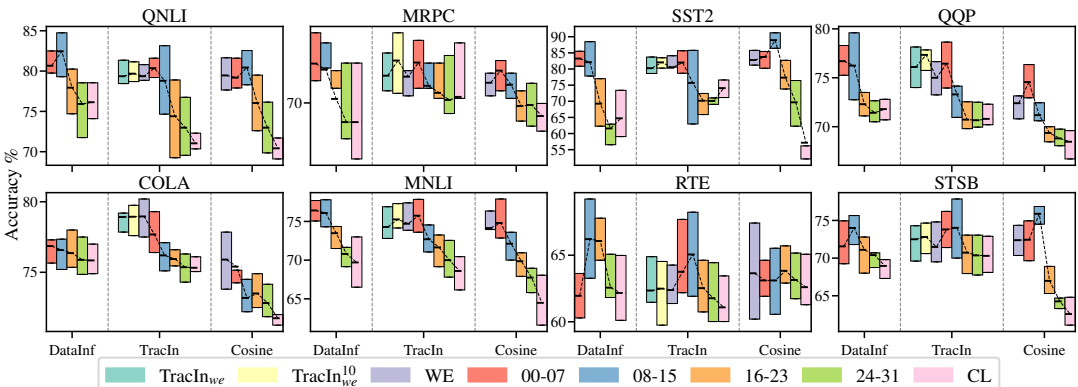
Training gradient cancellations: $1 \rightarrow$ no cancellation, $\infty \rightarrow$ extreme cancellation.

- Layer cancellation "smoothes" the gradient compensation over a large number of parameters, hiding weights where individual cancellation is high.
- Cancellation effect has a weak-to-no correlation to downstream performance after filtering of the least influential samples. (Spearman correlation ρ)
- High Cancellation Can Improve Sample Separability in Influence Score Range (Theorem 5.1)

	Layer	Mean $\pm$ Std	Median	Min	Max	ρ
Roberta-Large	WE	2.2 $\pm$ 0.3	1	1	∞	-0.3
	00-05	9.4 $\pm$ 3.9	11.8	1	10^6	0.1
	06-11	10.5 $\pm$ 4.6	14.3	1.7	10^6	0.1
	12-17	9.4 $\pm$ 5.1	12.5	1.6	∞	0.1
	18-23	8.5 $\pm$ 4.4	11.1	1.5	10^6	0.2
	CL	8.5 $\pm$ 6.1	11.1	1.6	∞	0.1
Llama-3.2 1B	WE	2.9 $\pm$ 0.3	1	1	∞	0.3
	00-03	8.4 $\pm$ 2.9	11.8	1.7	∞	-0.1
	04-07	5.8 $\pm$ 2.3	7.7	1.2	10^6	-0.2
	08-11	4.4 $\pm$ 1.6	5.8	1	10^5	-0.1
	12-15	4.0 $\pm$ 1.7	5.3	1	10^6	-0.1
	CL	3.1 $\pm$ 2.3	2.5	1	10^4	-0.1
Mistral 7B	WE	3.5 $\pm$ 0.3	1.1	1	∞	0.1
	00-07	17.7 $\pm$ 3.5	17	1.6	∞	0.0
	08-15	16.4 $\pm$ 6.4	18.6	2.2	∞	0.1
	16-23	15.6 $\pm$ 8.6	16	1.9	∞	0.0
	24-31	15.7 $\pm$ 10.2	15.5	1.8	∞	0.0
	CL	20.5 $\pm$ 19.5	11.7	3.8	10^5	0.1

[RQ2] Most Effective Layers (1)

Results for Mistral 7B on GLUE benchmark



[RQ2] Most Effective Layers (2)

Layers	Roberta-Large			Llama-3.2 1B			Qwen-2.5 1.5B			Mistral 7B		
	DataInf	TracIn	Cosine	DataInf	TracIn	Cosine	DataInf	TracIn	Cosine	DataInf	TracIn	Cosine
TI_{we}^{10}	-	1 (.49)	-	-	**1 (.56)**	-	-	2 (.49)	-	-	1 (.58)	-
TI_{we}	-	2 (.48)	-	-	2 (.52)	-	-	1 (.51)	-	-	2 (.50)	-
WE	-	**1 (.54)**	2 (.41)	-	3 (.44)	3 (.42)	-	3 (.40)	3 (.19)	-	2 (.50)	1 (.58)
1	4 (.29)	3 (.50)	3 (.35)	1 (.53)	2 (.54)	**1 (.53)**	**1 (.49)**	2 (.42)	2 (.45)	2 (.53)	**1 (.61)**	**1 (.60)**
2	4 (.30)	5 (.40)	2 (.40)	**1 (.58)**	1 (.54)	**1 (.53)**	1 (.48)	**1 (.57)**	1 (.60)	**1 (.60)**	2 (.48)	2 (.55)
3	2 (.40)	4 (.44)	**1 (.56)**	2 (.42)	4 (.38)	2 (.45)	**1 (.49)**	2 (.48)	**1 (.61)**	2 (.39)	3 (.29)	3 (.37)
4	**1 (.55)**	6 (.36)	1 (.54)	3 (.23)	5 (.22)	4 (.28)	2 (.31)	4 (.30)	2 (.40)	3 (.20)	4 (.20)	4 (.23)
CL	3 (.36)	7 (.17)	4 (.15)	4 (.18)	6 (.17)	5 (.15)	3 (.16)	5 (.22)	3 (.18)	3 (.20)	4 (.20)	5 (.09)

[RQ3] Better Influence Aggregation Strategies (1)

We consider better strategies that address $\mathcal{A}_{X',L}$, the score compensation, and magnitude domination between validation samples and NN modules.

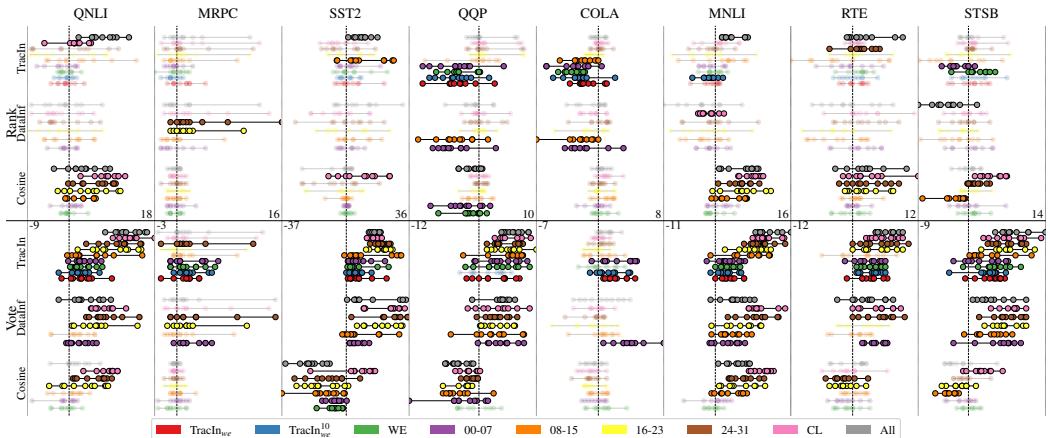
$$I(\bar{x}, X', \Theta_L) = \mathcal{A}_{X',L}(I'(\cdot, \bar{x}', \Theta_I))(\bar{x})$$

$$\text{Rank}(I') = \sum_{\bar{x}' \in X'', I \in L} \sum_{\bar{y} \in X} \mathbb{I}(I'(\bar{y}, \bar{x}', \Theta_I) < I'(\cdot, \bar{x}', \Theta_I))$$

$$\text{Vote}_k(I') = - \sum_{\bar{x}' \in X'', I \in L} \max(k - \sum_{\bar{y} \in X} \mathbb{I}(I'(\bar{y}, \bar{x}', \Theta_I) < I'(\cdot, \bar{x}', \Theta_I)), 0)$$

[RQ3] Better Influence Aggregation Strategies (2)

Results for Mistral 7B on GLUE benchmark

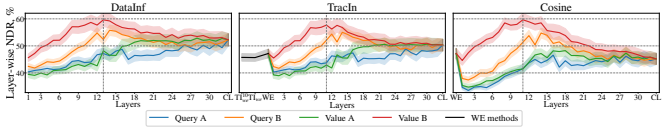


[RQ4] AUC NDR Reliability as Proxy Measure

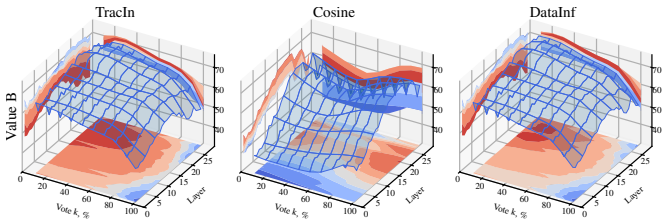
Medium-to-strong correlation of NDR to downstream performance,

	Infl.func	Roberta-Large			Llama-3.2 1B			Mistral 7B		
		C	NDR	AUC	C	NDR	AUC	C	NDR	AUC
Mean	DataInf	0.2	0.7	0.5	0.0*	0.6	0.5	0.1	0.5	0.5
	$\text{TracIn}_{\text{we}}$	-0.3	0.6	0.4	0.2*	0.6	0.5	0.1*	0.4	0.3
	$\text{TracIn}_{\text{we}}^{10}$	-0.3	0.6	0.4	0.0*	0.5	0.5	0.0*	0.5	0.3
	TracIn	0.0*	0.4	0.3	0.1*	0.6	0.5	0.0*	0.6	0.5
	Cosine	0.2	0.7	0.6	-0.0*	0.5	0.5	-0.1	0.6	0.5
Rank	DataInf	0.3	0.7	0.7	-0.1*	0.6	0.6	0.0*	0.6	0.7
	$\text{TracIn}_{\text{we}}$	-0.3	0.6	0.5	-0.1*	0.2	0.2*	0.0*	0.3	0.2
	$\text{TracIn}_{\text{we}}^{10}$	-0.3	0.5	0.5	-0.2*	0.1*	0.0*	-0.1*	0.2*	0.1*
	TracIn	0.2	0.5	0.5	-0.1*	0.5	0.5	-0.1*	0.4	0.5
	Cosine	0.2	0.7	0.7	-0.1*	0.6	0.6	0.1*	0.6	0.6
Vote	DataInf	0.3	**0.8**	**0.8**	-0.1	0.6	0.6	0.1	**0.9**	**0.8**
	$\text{TracIn}_{\text{we}}$	-0.4	**0.8**	**0.8**	-0.1*	0.5	0.5	-0.1*	**0.8**	**0.8**
	$\text{TracIn}_{\text{we}}^{10}$	-0.4	0.7	**0.8**	-0.1*	0.5	0.5	0.0*	**0.8**	**0.8**
	TracIn	0.2	**0.8**	0.7	-0.1	0.6	0.6	0.1	**0.8**	**0.8**
	Cosine	0.2	0.7	0.6	-0.1	0.2	0.2	0.0*	0.4	0.4

Mistral NDR across layers and LoRA modules.



Positional Voting NDR Dependency on k.



Conclusions

1. Cancellation effect is not a reliable predictor of downstream performance after detrimental set filtering.
2. Middle attention layers achieve better influence-scoring performance than the WE and CL heads.
3. Alternative aggregations of individual influence scores may drastically improve the filtering capability (Rank, Vote-k methods).
4. NDR is a better proxy measure that may be used for studying alternative aggregations and layer/module set selection.