LaRA: Layer-wise Representation Analysis for Detecting Data Contamination in RL Post-Training

Minju Gwak, Minseo Kwak, Dongseok Lee, Guijin Son, Alan Ritter, Jaehyung Kim

0. TL;DR

How do model internals change in RL-trained models?

For detecting contamination in RL, instead of looking at outputs, look at how its internal representations change across layers.

1. Introduction

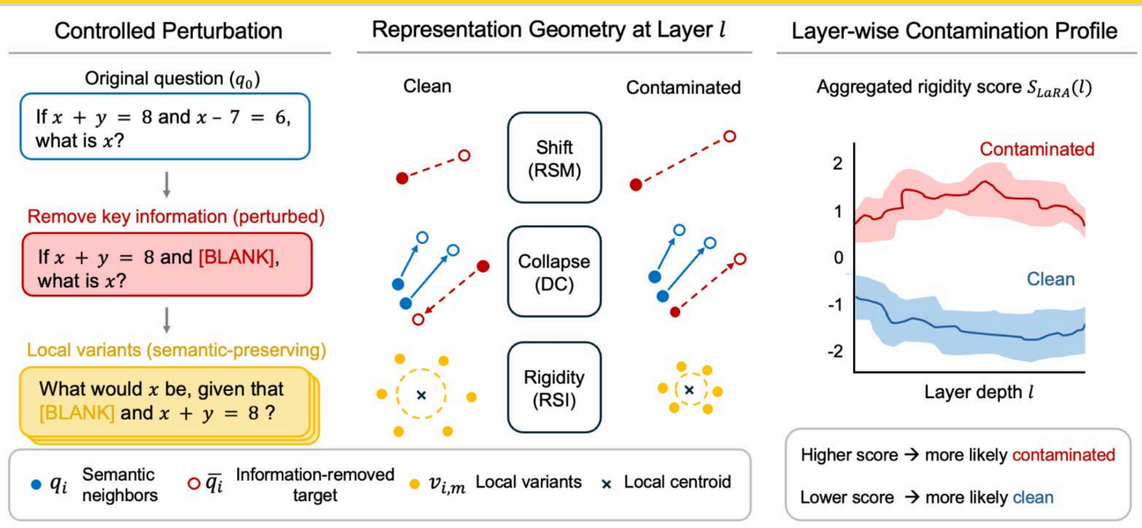
Limitations of Output-level Detection

1. Examine output statistics (likelihood, ent.)
2. RL optimizes trajectory-level reward, ***not token likelihood***.

Key Hypothesis:
*Memorization leaves **geometric fingerprints** inside the hidden representation*

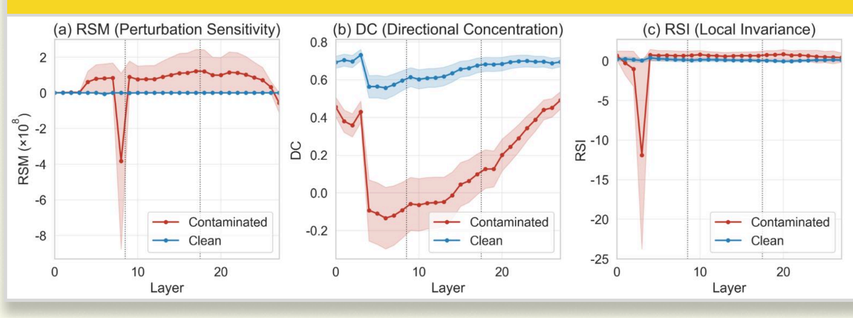
*Use this for **detecting RL contamination!***

2. Overview of LaRA Framework



1. **R**epresentation **S**hift **M**agnitude:
   - *How much does the representation move?*
2. **D**irectional **C**ollapse
   - *Where does it move?*
3. **R**epresentation **S**tability **I**ndex
   - *How stable is the representation?*

3. Representation Geometry of RL Model



Deviations btw. clean v. cont

4. Contamination Detection Protocol

1. Clean-reference Robust Standardization

$$\mu^{\text{clean}}_{m,\ell} = \text{median}\Big(\tilde{m}_\ell(x) : x \in \mathcal{D}^{\text{clean}}\Big),$$

$$\sigma^{\text{clean}}_{m,\ell} = 1.4826 \cdot \text{MAD}\Big(\tilde{m}_\ell(x) : x \in \mathcal{D}^{\text{clean}}\Big)$$

2. Metric-specific Anomaly Alignment

$$\hat{z}_{m,\ell}(x) = \begin{cases} z_{m,\ell}(x), & m = \text{RSM}, \\ z_{m,\ell}(x), & m = \text{DC}, \\ -z_{m,\ell}(x), & m = \text{RSI}. \end{cases}$$

3. Layer-wise Aggregation

$$S_{\text{LaRA}}(x) = \frac{1}{|\mathcal{M}|\,|\mathcal{L}|} \sum_{m \in \mathcal{M}} \sum_{\ell \in \mathcal{L}} \hat{z}_{m,\ell}(x)$$

5. Results & Analysis

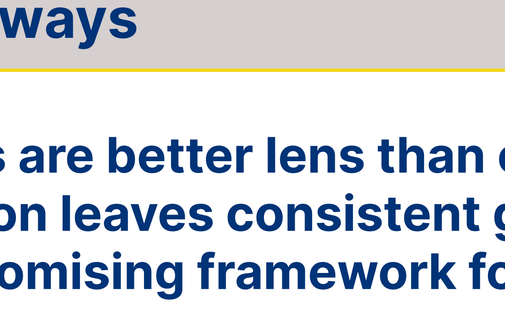
Method	Eurus		LIMR		OLMO	
	AUC	TPR	AUC	TPR	AUC	TPR
Recall	0.59	0.07	0.54	0.07	0.42	0.03
CDD	0.38	0.00	0.64	0.00	0.50	0.00
Min-K%	0.32	0.07	0.26	0.03	0.50	0.03
Min-K%++	0.29	0.03	0.60	0.07	0.46	0.07
PPL	0.68	0.23	0.73	0.13	0.49	0.07
SC	0.70	0.27	0.44	0.00	0.42	0.03
S_{LaRA}	0.63	0.19	**0.80**	**0.46**	**0.54**	0.04
SC + S_{LaRA}	**0.73**	**0.31**	0.47	0.04	0.50	**0.08**

(a) Initial checkpoint results across models.

	Eurus						LIMR					
	Init		E1		E2		Init		E1		E2	
Method	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
Recall	0.59	0.07	0.56	0.13	0.52	0.00	0.54	0.07	0.47	0.03	0.47	0.00
CDD	0.38	0.00	0.38	0.00	0.38	0.00	0.64	0.00	0.61	0.00	0.59	0.00
Min-K%	0.32	0.07	0.31	0.07	0.33	0.07	0.26	0.03	0.26	0.00	0.26	0.00
Min-K%++	0.29	0.03	0.26	0.03	0.20	0.03	0.60	0.07	0.58	0.07	0.63	0.00
PPL	0.68	0.23	**0.69**	0.23	0.67	0.17	0.73	0.13	**0.74**	0.13	0.74	0.17
SC	0.70	0.27	0.58	0.07	0.78	0.30	0.44	0.00	0.40	0.00	0.46	0.00
S_{LaRA}	0.63	0.19	0.68	0.19	0.70	0.15	**0.80**	**0.46**	0.72	**0.23**	**0.81**	**0.20**
SC + S_{LaRA}	**0.73**	**0.31**	0.65	**0.35**	**0.79**	**0.38**	0.47	0.04	0.46	0.04	0.54	0.04

(b) Results across Eurus and LIMR training checkpoints.

- Outperforms Baseline,
- Orthogonal to output-level metrics





- Robust to Perturbation Types & #

6. Ablations

RSM	DC	RSI	Init		E1		E2	
			AUC	TPR	AUC	TPR	AUC	TPR
✗	✗	✓	0.60	0.19	0.51	0.15	0.56	0.16
✗	✓	✗	0.71	0.42	0.66	0.08	0.76	0.24
✓	✗	✗	0.75	0.12	0.69	0.08	0.71	0.04
✗	✓	✓	0.76	0.46	0.69	0.15	0.81	0.32
✓	✗	✓	0.74	0.27	0.68	0.08	0.71	0.12
✓	✓	✗	0.78	0.31	0.71	0.19	0.79	0.12
✓	✓	✓	0.80	0.46	0.72	0.23	0.81	0.20

Table 2: **Metric ablations across RL training epochs.**

- DC provides the strongest standalone detection,
- but all three is required

Check Us Out!

ML3 Page **MJ's Homepage**

7. Key Takeaways

- **Hidden reps are better lens than outputs**
- **Memorization leaves consistent geometric signatures**
- **LaRA is a promising framework for studying post-training**