ICML
International Conference
On Machine Learning

A Narrowing Geometry in Contaminated Reasoning

Jiakuan Xie, Pengfei Cao, Kang Liu, Jun Zhao

Background: Data Contamination

Motivation

- **Benchmark contamination is widespread.**

  Leaked data can induce *contaminated reasoning.*

- **Strong performance can be misleading.**

  Does high accuracy imply genuine reasoning?

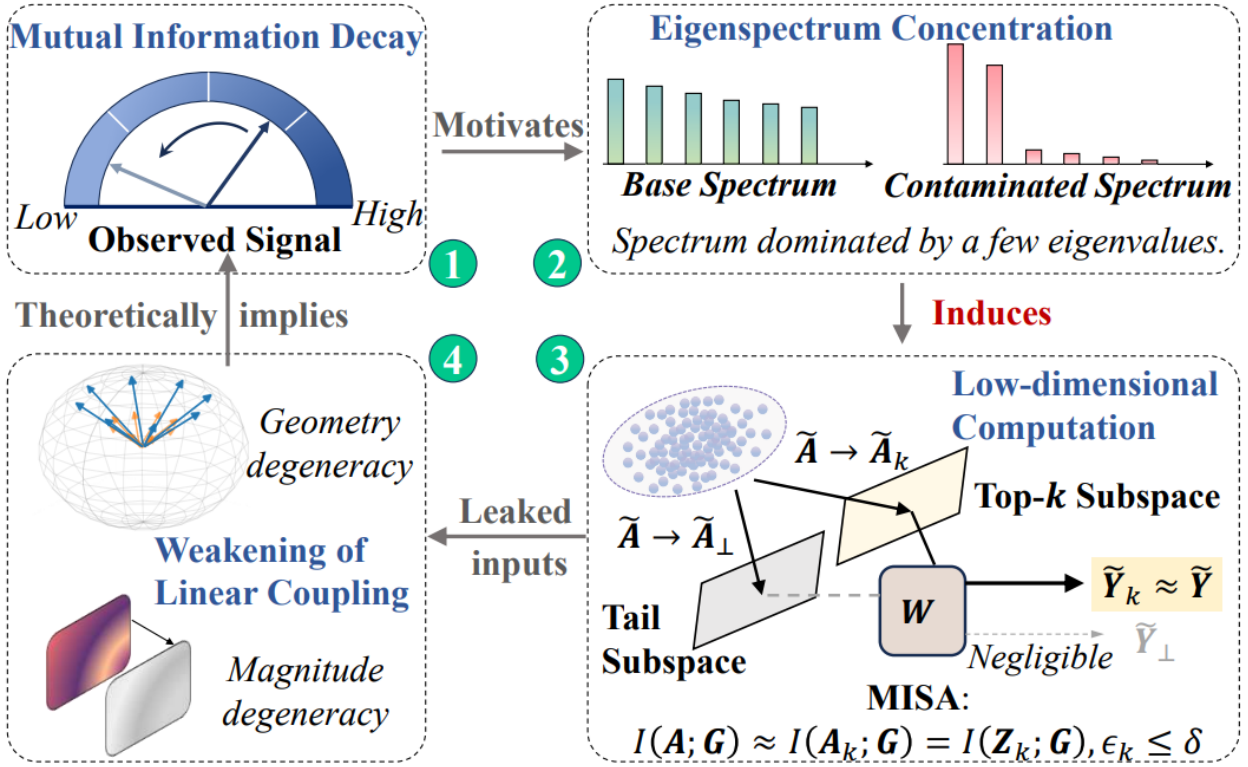
- **Robust evaluation remains difficult.**

  It is hard to design reliable evaluation framework.

Key Question

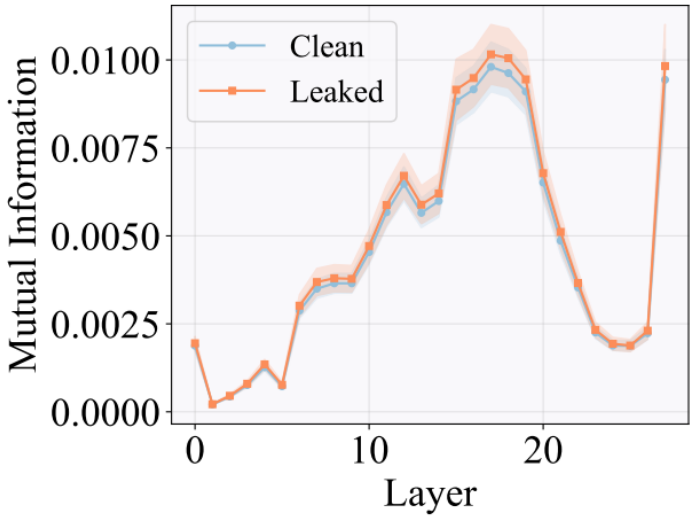
What changes inside the model when reasoning is contaminated?

Overview

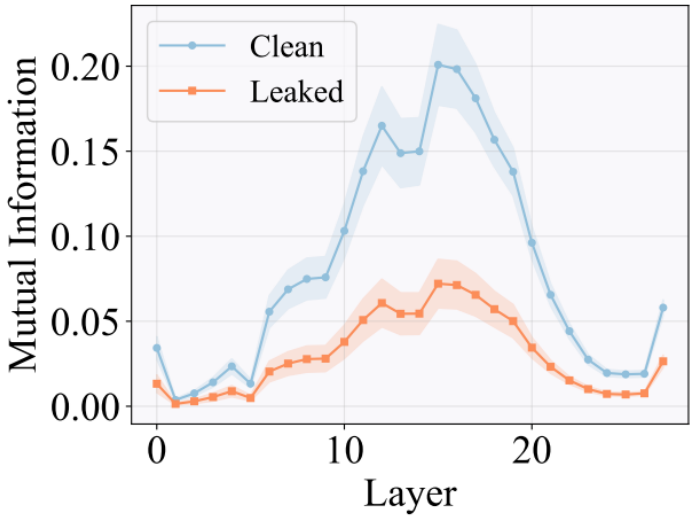


- **Signal**: Mutual Information Decay
- **Mechanism**: Eigenspectrum Concentration → Low-dimensional Computation
- **Explanation**: Why Mutual Information Decay happens
- **Application**: Reasoning Restoration

Mutual Information Decay



(a) Base Model

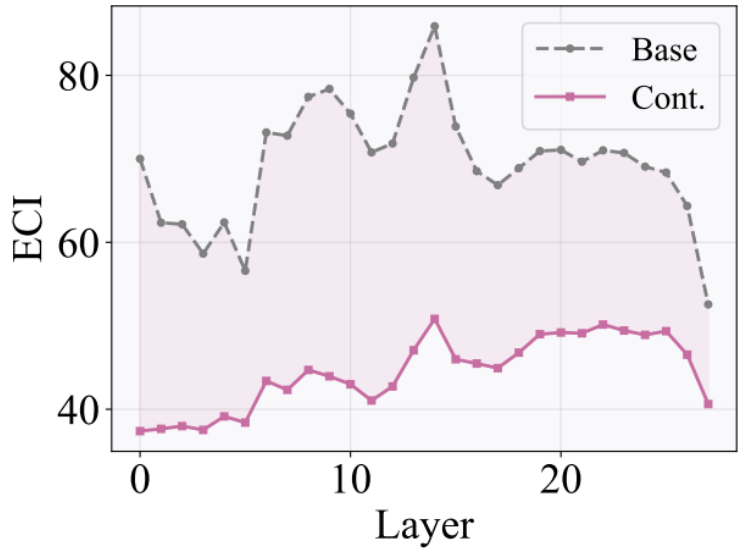


(b) Contaminated Model

- **Metric**: Gaussian Mutual Information $I(\boldsymbol{A};\boldsymbol{G}) = \frac{1}{2}\log\frac{\det(\boldsymbol{C_A})\det(\boldsymbol{C_G})}{\det(\boldsymbol{C})}$
- **Observation**
  - Base model: Leaked ≈ Clean
  - Contaminated model: Leaked < Clean
- **Takeaway**
  - Mutual information decay is a distinctive signal of contaminated reasoning.

Mechanisms

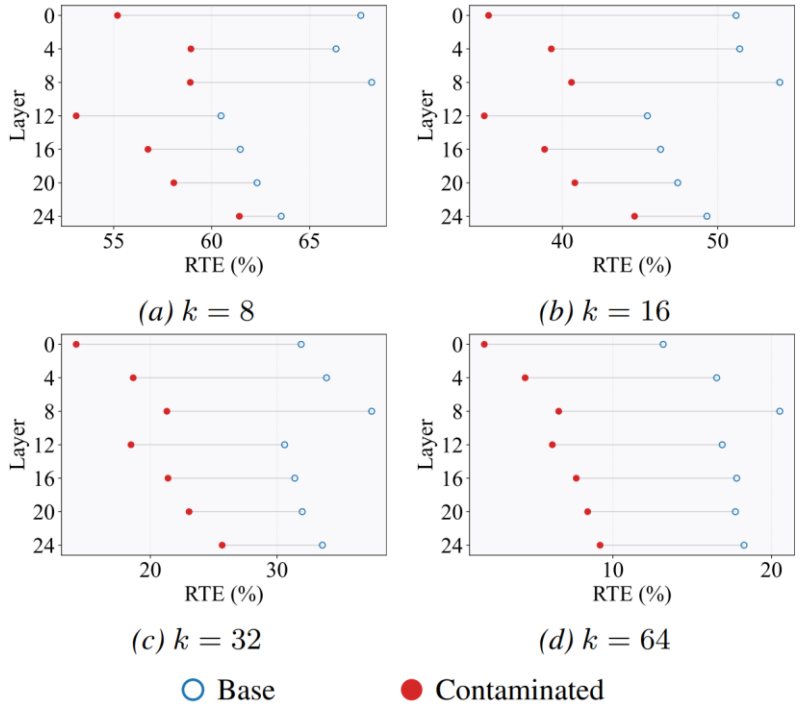
Eigenspectrum Concentration



Takeaways

- Lower ECI → variance dominated by a few eigenvalues.
- Contaminated activations exhibit stronger eigenspectrum concentration.

Low-Dimensional Computation



Takeaways

- Lower RTE → stronger low-dimensional computation.
- Module outputs are well captured by a small top-k subspace.

Causal Validation

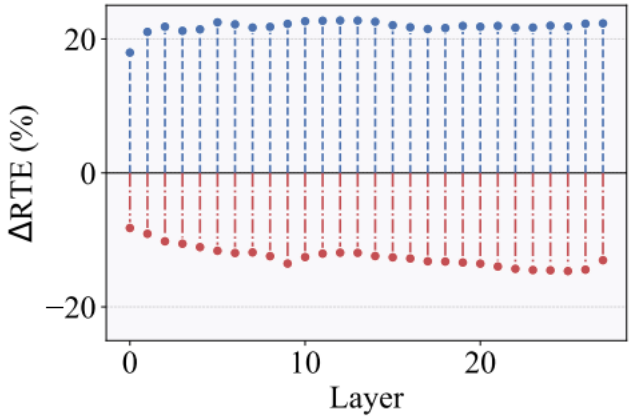
No Significant Tail Vector Amplification

Models	Metrics	k = 8	16	32	64
Base	R_{worst}	1.82	1.80	1.78	1.76
	R_{med}	0.93	0.94	0.95	0.98
Contaminated	R_{worst}	1.71	1.69	1.68	1.60
	R_{med}	0.95	0.95	0.96	0.98

Observation

- Transformed tail eigenvectors have comparable norms to principal eigenvectors.

Spectrum Reshaping



Flatten: RTE ↓

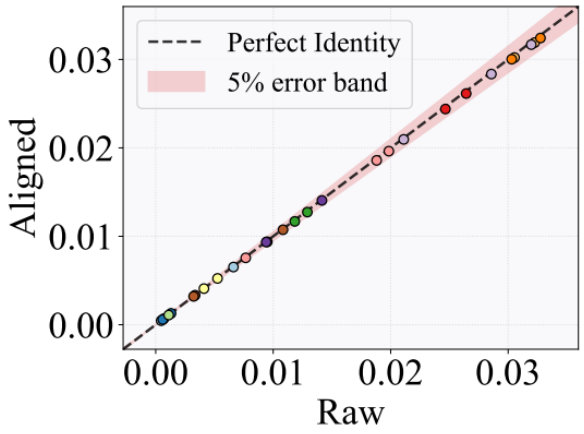
Sharpen: RTE ↑

Observation

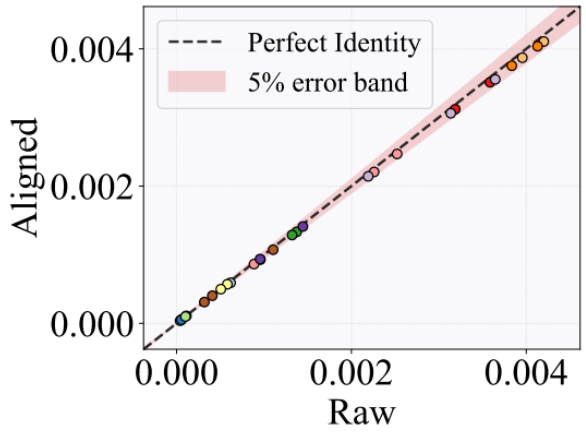
- Manipulating only the eigenspectrum changes the computation dimensionality.

Explanation of Mutual Information Decay

Mutual Information Subspace Alignment (MISA)



(a) Clean Data



(b) Leaked Data

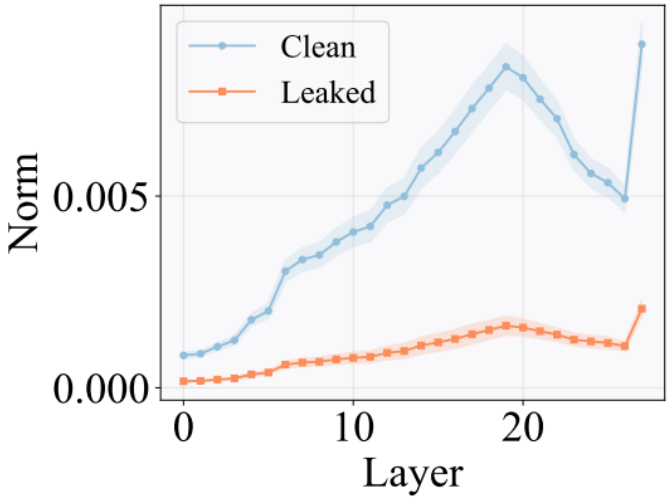
$$I(\boldsymbol{A};\boldsymbol{G}) \approx I(\boldsymbol{A}_k;\boldsymbol{G}) = I(\boldsymbol{Z}_k;\boldsymbol{G}) \quad (\epsilon_k \leq \delta)$$

Takeaway

- The mutual information is determined primarily within the active computation subspace.

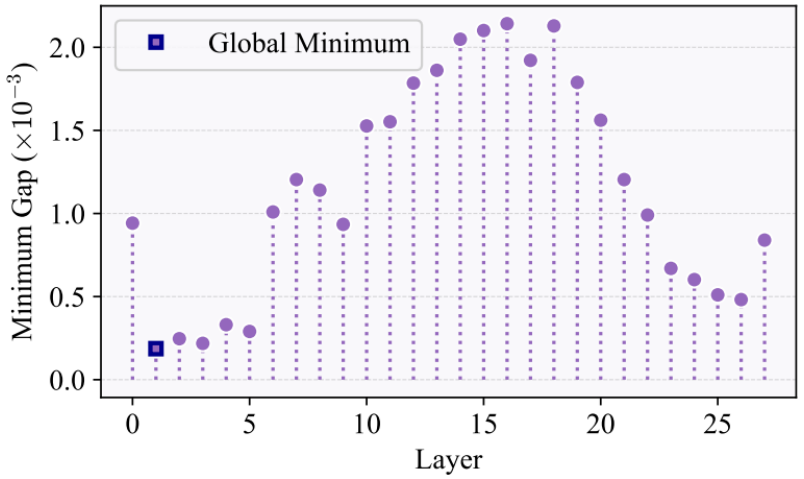
Explanation of Mutual Information Decay

Weakening of Linear Coupling



$$\left\| C^{(l)}_{Z_k G} \right\|_F < \left\| C^{(c)}_{Z_k G} \right\|_F$$

Singular Value Partial Order



$$\boldsymbol{R} = \boldsymbol{C}_{\boldsymbol{Z}_k}^{-\frac{1}{2}} \boldsymbol{C}_{\boldsymbol{Z}_k \boldsymbol{G}} \boldsymbol{C}_{\boldsymbol{G}}^{-\frac{1}{2}} \qquad \rho_i^{(l)} \leq \rho_i^{(c)} \quad \forall i, \text{ and } \exists j : \rho_j^{(l)} < \rho_j^{(c)}$$

Observation

- Leaked data shows weaker representation-gradient coupling in the active subspace.

Observation

- The decay holds direction-wise in the whitened canonical space.

Application: Reasoning Restoration

Method: Top-Subspace Suppression (TSS)

$$a^{\text{new}} = \mu + \left(I - \gamma^{\alpha} U_k U_k^{\top}\right)(a - \mu), \quad \gamma = \frac{\left\| U_k U_k^{\top}(a - \mu) \right\|}{\|a - \mu\|}, \quad \alpha \in [0, 1].$$

Quantitative Results

Models	Strategy	Restor. (↑)	Retain (↑)	Avg. (↑)
Qwen-1.5B	Cont.	0.00	100.00	50.00
	Sampling	23.08	89.02	56.05
	TED	27.47	86.83	57.15
	LNE-B.	39.56	89.96	64.76
	TSS (Ours)	**51.65**	**94.37**	**73.01**
Qwen-7B	Cont.	0.00	100.00	50.00
	Sampling	9.09	91.57	50.33
	TED	22.73	89.33	56.03
	LNE-B.	31.82	88.76	60.29
	TSS (Ours)	**45.45**	**97.19**	**71.32**

- TSS achieves the best average consistency.

Qualitative Case

Question

Ground Truth: 50000 × 0.01 × 0.05 = 25

Marcy makes homemade candles that she markets as 99% guaranteed not to explode. 5% of the more dangerous candles also have a defect that makes them smell like wet dog. If she makes 50000 candles, how many of them will both smell like wet dog and explode?

Base Model (✕)

… Marcy makes 50,000 candles… 99% of these...
0.99×50,000 = 49,500 candles…
5% of these 49,500 candles have defects.
0.05×49,500 = 2,475 defective candles…
Thus, the number of candles that both smell like wet dog and explode is: 2475

TSS + Contaminated Model (✕)

… find the number of candles that explode: .99 * 50000 = 49500 candles…
Then multiply that number of candles with the other defect to find the total number of them: 49500 * .05 = 2475 candles… 2475

- TSS makes the contaminated model follow the base model's reasoning trajectory.

Mechanistic understanding enables reasoning restoration.

Summary

- **Signal**
  - Mutual Information Decay
- **Mechanism**
  - Narrowing Geometry
- **Application**
  - Reasoning Restoration

ICML
International Conference
On Machine Learning

THANK YOU