上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

SJTU

Diff-eRank: A Novel Rank-Based Metric for Evaluating Large Language Models

Lai Wei, *Zhiquan Tan, Chenghai Li, Jindong Wang, Weiran Huang*

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

MIFA LAB
机器智能基础与应用实验室

NEURAL INFORMATION PROCESSING SYSTEMS

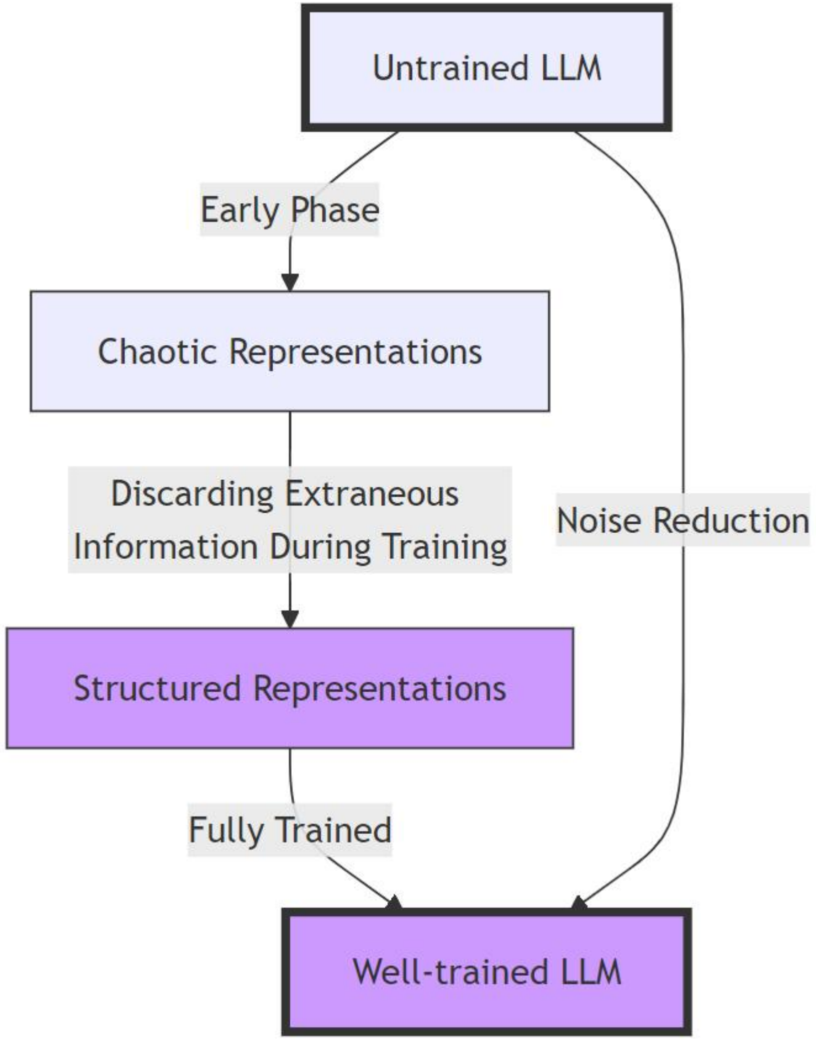
饮 水 思 源 • 爱 国 荣 校

Motivation

- Ilya Sutskever's talk: larger language models find more shared hidden structures in data samples by eliminating redundant information.

- Defining and quantifying this process remains a challenge.

- We hypothesize that an ideal metric should reflect the geometric characteristics of the data, such as the dimensionality of its representations, and should also be grounded in information theory. We choose to study the rank of the data representations.



Intuition

Why rank?

- It measures the extent of linear independence among these representations (i.e., the geometric structure).
- It is also related to the amount of information contained in the representation, while a lower rank indicates that the information has been structured or compressed.

We introduce Diff-eRank as an information-theoretic evaluation metric that meets the previous two requirements to quantify the degree of “noise reduction”.

Method

Construction of eRank

$$\Sigma_S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbf{z}_i - \bar{\mathbf{z}}}{\|\mathbf{z}_i - \bar{\mathbf{z}}\|}\right)\left(\frac{\mathbf{z}_i - \bar{\mathbf{z}}}{\|\mathbf{z}_i - \bar{\mathbf{z}}\|}\right)^{\top}, \quad \text{eRank}(\mathbf{A}) = \exp\left(-\sum_{i=1}^{Q}\frac{\sigma_i}{\sum_{i=1}^{Q}\sigma_i}\log\frac{\sigma_i}{\sum_{i=1}^{Q}\sigma_i}\right)$$

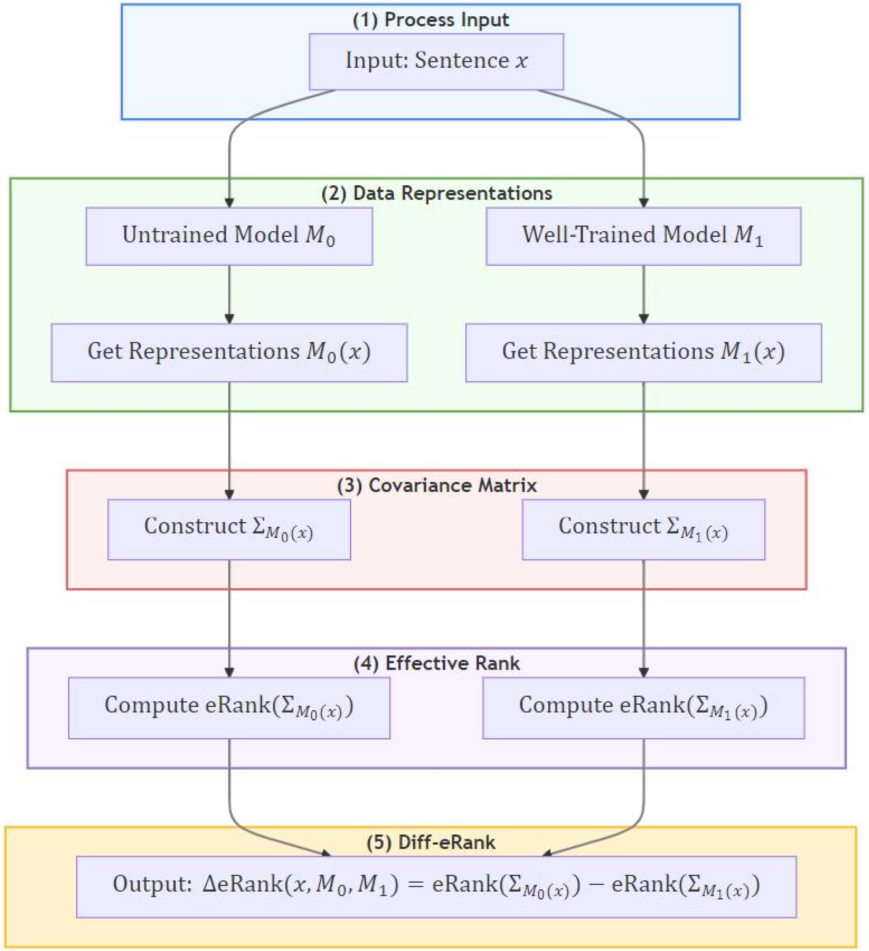
Relationship with Matrix Entropy

For a matrix $\mathrm{K} \in R^{d\times d}$ (positive semi-definite, $tr(\mathrm{K}) = 1$), $\mathrm{H}(\mathrm{K}) = -tr(\mathrm{K}\, log\, \mathrm{K})$, i.e., $\mathrm{H}(\mathrm{K}) = -\sum_{i=1}^{d}\lambda_i \log \lambda_i$. eRank($\Sigma_S$) is exactly the same as exp(H(Σ_S)). Σ_S is actually a density matrix. exp(H(Σ_S)) can be seen as a measure of randomness.

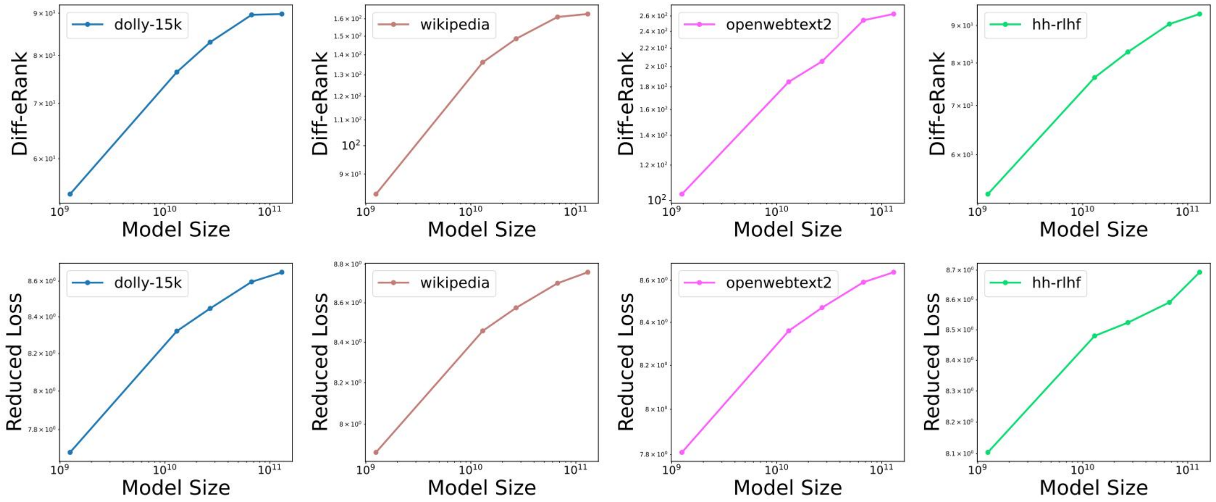
Diff-eRank

$$\Delta\,\text{eRank}(x, M_0, M_1) = \text{eRank}\left(\mathbf{\Sigma}_{M_0(x)}\right) - \text{eRank}\left(\mathbf{\Sigma}_{M_1(x)}\right)$$

$$\Delta\,\text{eRank}(\mathcal{D}, M_0, M_1) = \exp\left(\frac{\sum_{i=1}^{n}\mathrm{H}\left(\mathbf{\Sigma}_{M_0(x_i)}\right)}{n}\right) - \exp\left(\frac{\sum_{i=1}^{n}\mathrm{H}\left(\mathbf{\Sigma}_{M_1(x_i)}\right)}{n}\right).$$



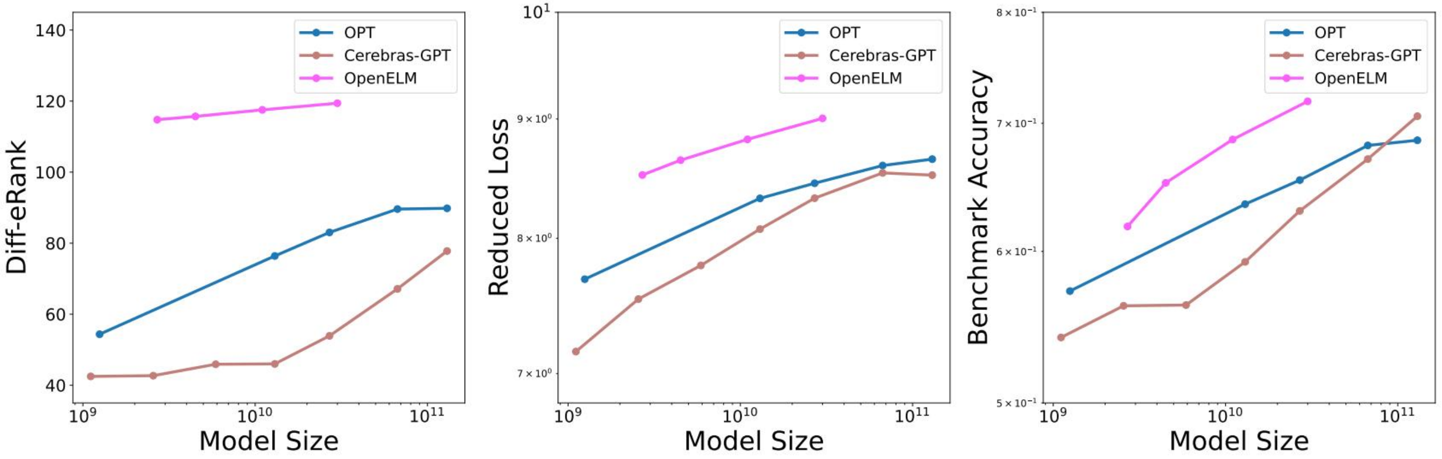
Evaluations of Large Language Models



- We define the reduced (cross-entropy) loss as: $\Delta L(U, M_0, M_1) = L(U, M_0) - L(U, M_1)$.
- Besides, we also include benchmark accuracy for comparison.

Benchmarks	Indicators	OPT Models Size 125M	1.3B	2.7B	6.7B	13B
OpenbookQA	Acc	0.276	0.332	0.370	0.360	**0.366**
	ΔL	5.734	6.138	6.204	**6.258**	6.236
	Diff-eRank	1.410	2.140	2.338	2.280	**3.032**
PIQA	Acc	0.619	0.714	0.733	0.756	**0.767**
	ΔL	6.472	6.928	6.999	**7.077**	7.068
	Diff-eRank	4.647	6.294	6.774	6.950	**7.267**

Ablation Study



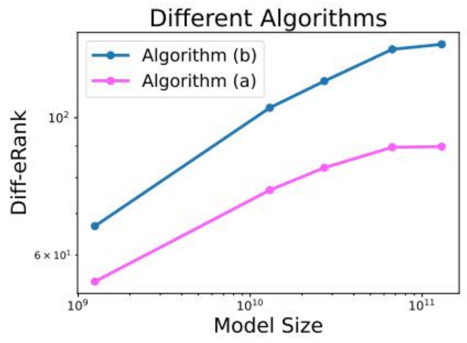


Figure 4: Different designs for Diff-eRank.

Table 4: Diff-eRank on different layers of OPT models. Only the Diff-eRank on the last layer indicates an increasing trend.

OPT Models	125M	1.3B	2.7B	6.7B	13B
First Layer	73.07	73.03	66.93	49.24	41.83
Middle Layer	87.75	51.98	56.16	66.63	73.88
Last Layer (↑)	54.35	76.39	83.02	89.60	89.81

- Comparing Diff-eRank with reduced loss and benchmark accuracy across different model families, including OPT, Cerebras-GPT, and OpenELM.

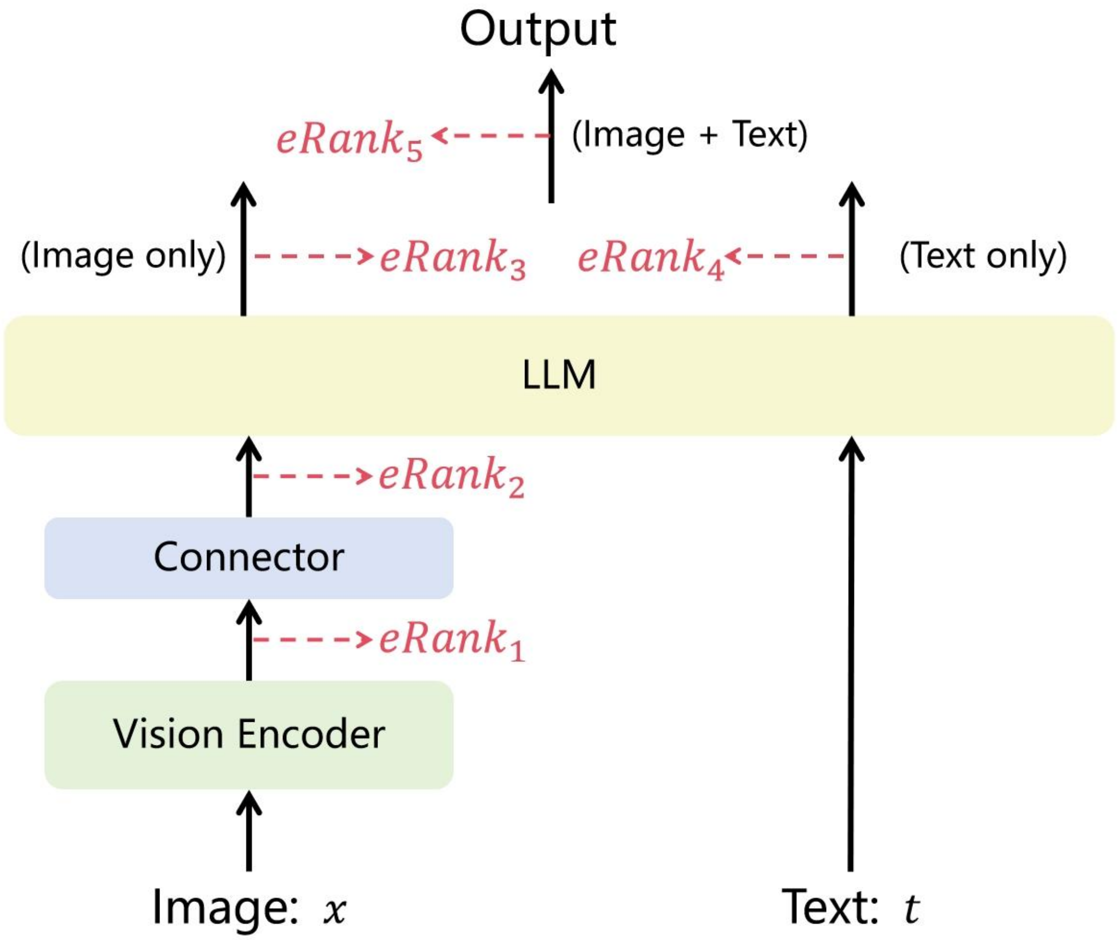
- We consider "Algorithm (b)" for Diff-eRank in Figure 4 defined below.

$$\mathrm{eRank}^{(b)}(\mathcal{D}, M) = \frac{\sum_{x\in\mathcal{D}} \exp(\mathrm{H}(\mathbf{\Sigma}_{M(x)}))}{|\mathcal{D}|} = \frac{\sum_{x\in\mathcal{D}} \mathrm{eRank}(\mathbf{\Sigma}_{M(x)})}{|\mathcal{D}|}.$$

$$\Delta\,\mathrm{eRank}^{(b)}(\mathcal{D}, M_0, M_1) = \mathrm{eRank}^{(b)}(\mathcal{D}, M_0) - \mathrm{eRank}^{(b)}(\mathcal{D}, M_1).$$

- We also extend our experiments to encompass additional layers within the models in Table 4.

Modality Alignment



- We define new metrics for Multi-modal LLMs to evaluate the modality alignment by analyzing the eRanks of different parts of representation .

$$\text{Image Reduction Ratio} = \frac{\text{eRank}_1 - \text{eRank}_2}{\text{eRank}_1},$$

$$\text{Image-Text Alignment} = \frac{\text{avg}(\text{eRank}_3, \text{eRank}_4, \text{eRank}_5)}{\max(\text{eRank}_3, \text{eRank}_4, \text{eRank}_5)}.$$

Evaluations of Multi-Modal Large Language Models

- Both LLaVA-1.5 and MiniGPT-v2 align well as they all have a relatively high alignment score.
- LLaVA-1.5 outperforms MiniGPT-v2 in "Image-Text Alignment", which is also consistent with their performance, as LLaVA-1.5 surpasses MiniGPT-v2 in most of benchmarks.

Effective Rank	LLaVA-1.5		MiniGPT-v2	
	DETAIL_23K	CC_SBU_ALIGN	DETAIL_23K	CC_SBU_ALIGN
eRank_1	18.34	9.00	90.59	74.79
eRank_2	11.28	5.20	55.70	46.15
eRank_3	45.62	28.47	58.50	48.68
eRank_4	74.21	59.00	63.63	52.68
eRank_5	76.34	47.63	108.53	93.29
Image Reduction Ratio (↑)	0.3850	0.4222	0.3851	0.3829
Image-Text Alignment (↑)	0.8566	0.7618	0.7084	0.6955

- We also calculate the eRank after rotating the images clockwise, which indicates that subtle changes in the vision encoder's understanding of images can be effectively conveyed to the LLM part and affect the MLLM's modality alignment.

Effective Rank	LLaVA-1.5 on DETAIL_23K	
	BASE	ROTATE IMAGE CLOCKWISE
eRank_1	18.34	19.20 (↑)
eRank_2	11.28	12.31 (↑)
eRank_3	45.62	46.54 (↑)
eRank_4	74.21	74.21 (-)
eRank_5	76.34	77.69 (↑)
Image Reduction Ratio	0.3850	0.3588 (↓)
Image-Text Alignment	0.8566	0.8514 (↓)

Conclusion and Discussion

- We introduce Diff-eRank, a new metric that can measure the "noise reduction" ability of LLM based on data representation. Our method reveals the geometric characteristics of the data and is grounded in information theory.
- The empirical investigations show that the Diff-eRank increases when the model scales and correlates with the trend of loss and downstream task accuracy.
- Moreover, we use this metric to define the alignment metrics for multi-modal LLMs and find contemporary models align very well.
- Some useful techniques like pruning, quantization, and distillation may benefit from such metrics that reveal internal redundancies. The Diff-eRank metric may aid in identifying which parts of the model can be compressed without significant loss of information.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Thank You

Paper Github Wechat

饮 水 思 源 爱 国 荣 校