# Attribution-based Sparse Activation in Large Language Models

Jifeng Song*, Xiangyu Yin*, Boyuan Yang, Kai Huang, Weichen Liu, Wei Gao

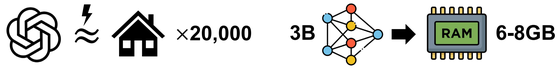
Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh

[Paper] [Code]

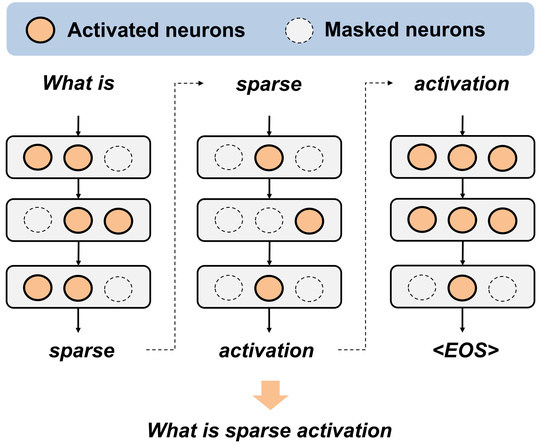
**MLSys 2026**

## Problem Statement

- **LLM inference is expensive**, especially on mobile devices



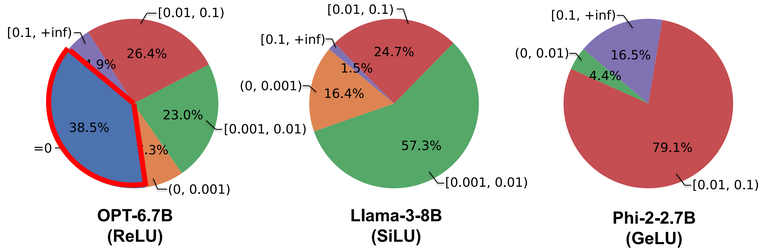

- Existing approaches like pruning, quantization, and distillation are **fixed**, **offline**, and **not runtime adaptable**
- **Sparse activation:** Per-input neuron selection that deactivates unimportant neurons at runtime

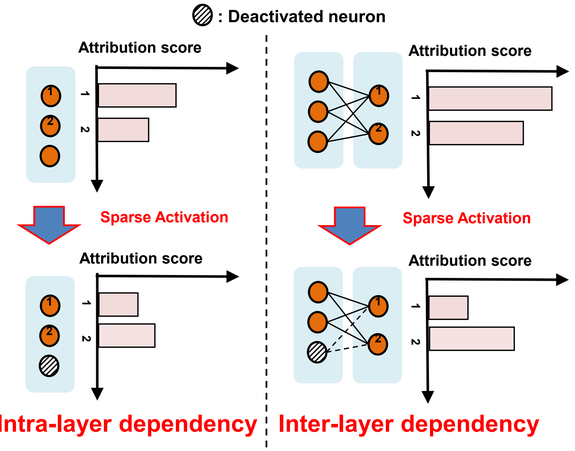



- **Key question:** which neurons to deactivate?

## Method

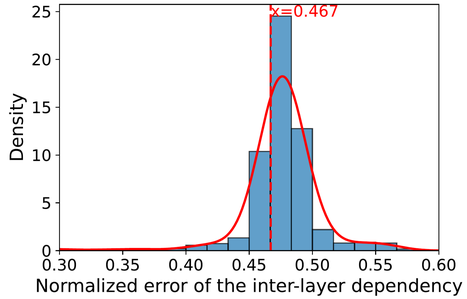
- **Traditional method:** Deactivate **Zero-Output** neurons, but only work for old LLMs w/ ReLU activations
- **Our approach:** **Attribution-based** sparse activation that deactivates neurons by contribution, not magnitude.
- **Attribution metrics:** Integrated Gradients (IG), Gradient × Output (GxO), etc.
- **Challenge:** Attribution scores are Interdependent




- **Our Solution: Corrected GxO** that analytically bound the inter-layer error
  - Tight upper bound from neuron magnitudes & gradients
  - Expectation under truncated-normal distribution

$$\text{Corrected GxO} = \text{GxO} + \frac{1}{2}|x_i|\sqrt{\sum_{k=1}^{N_1}\left(\frac{\partial F}{\partial x_k}\right)^2}$$




## Evaluations

**Llama-3-8B:** 60% sparsity, <5% BLEU loss
**Phi-2 (2.7B):** 60% sparsity, <5% loss
**Gemma-2B, MobiLlama-0.5B:** 70% sparsity, <5% loss

**At AR = 30% on Phi-2:** 35% latency reduction & 43% GPU memory reduction

| Model & Metric | AR=10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| **Phi-2-2.7B** | | | | | | | | | | |
| Magnitude | 10.5 | 19.1 | 21.5 | 23.5 | 25.3 | 24.9 | 24.2 | 24.5 | 24.0 | 33.9 |
| Gradient | 3.8 | 4.3 | 5.1 | 5.0 | 5.3 | 5.4 | 5.5 | 7.5 | 8.9 | 33.9 |
| SNIP/Fisher | 10.8 | 13.3 | 18.5 | 20.7 | 23.7 | 24.1 | 23.1 | 24.3 | 24.8 | 33.9 |
| IG | 13 | 16.4 | 18.6 | 22.8 | 23.4 | 22.7 | 23.8 | 25.2 | 30.7 | 33.9 |
| GxO | 15.7 | 19.7 | 19.0 | 21.0 | 22.5 | 23.6 | 22.7 | 24.1 | 31.4 | 33.9 |
| Corrected GxO | **17.8** | **18.3** | **26.8** | **28.3** | **29.6** | **31.5** | **30.7** | **32.2** | **36.7** | 33.9 |
| **Gemma-2B** | | | | | | | | | | |
| Magnitude | 0 | 0.2 | 2.75 | 5.57 | 6.95 | 7.58 | 8.19 | 8.79 | 10.71 | 10.72 |
| Gradient | 0 | 0 | 0 | 0.37 | 0.61 | 3.04 | 6.6 | 8.14 | 10.03 | 10.72 |
| SNIP/Fisher | 0 | 0.11 | 1.86 | 2.47 | 3.06 | 4.41 | 6.77 | 8.07 | 10.32 | 10.72 |
| IG | 0 | 0 | 0.27 | 1.16 | 1.56 | 4.7 | 6.24 | 8.64 | 8.73 | 10.72 |
| GxO | 0 | 0.02 | 0.31 | 0.94 | 0.96 | 4.83 | 5.74 | 6.86 | 11.8 | 10.72 |
| Corrected GxO | 0 | **0.55** | **5.29** | **5.63** | **8.04** | **10.2** | **10.71** | **11.58** | **13.83** | 10.72 |
| **MobiLlama-0.5B** | | | | | | | | | | |
| Magnitude | 0.26 | 1.21 | 1.73 | 2.33 | 2.86 | 2.65 | 3.6 | 3.98 | 5.39 | 5.45 |
| Gradient | 0.45 | 0.52 | 0.68 | 0.76 | 0.76 | 0.61 | 0.93 | 1.53 | 2.31 | 5.45 |
| SNIP/Fisher | 0.45 | 1.12 | 1.66 | 2.33 | 2.55 | 3.23 | **4.82** | **5.13** | **5.5** | 5.45 |
| IG | 0.84 | 1.61 | 1.84 | 1.75 | 1.48 | 0.94 | 1.19 | 1.3 | 1.28 | 5.45 |
| GxO | 0.68 | 1.25 | 1.19 | 1.04 | 1.07 | 0.68 | 0.95 | 1.04 | 1.23 | 5.45 |
| Corrected GxO | **1.41** | **3.8** | **4.07** | **4.48** | **4.63** | **4.63** | 4.39 | 4.66 | 5.01 | 5.45 |
| **Llama-3-8B** | | | | | | | | | | |
| Magnitude | 1.59 | 2.76 | 6.89 | 11.97 | 13.58 | 15.8 | 18.7 | 16.97 | 14.95 | 26.52 |
| Gradient | 0.75 | 0.6 | 1.11 | 0.93 | 1.39 | 1.61 | 1.63 | 2.56 | 10.75 | 26.52 |
| SNIP/Fisher | 3.07 | 3.22 | 4.17 | 6.89 | 10.08 | 12.86 | 18.13 | 16.48 | 18.35 | 26.52 |
| GxO | 1.93 | 1.71 | 2.53 | 3.59 | 4.2 | 5 | 6.42 | 7.78 | 20.73 | 26.52 |
| Corrected GxO | **4.6** | **7.97** | **11.84** | **21.66** | **24.48** | **26.08** | **26.94** | **27.8** | **29.3** | 26.52 |

*Table 1.* Accuracy of sparsely activated LLMs with different activation ratios (AR) on TruthfulQA

## Acknowledgments

This work was supported in part by National Science Foundation (NSF) under grant number IIS-2205360 and CCF-2215042, and National Institutes of Health (NIH) under grant number R01HL170368 and R01HL184168.