Universal Properties of Activation Sparsity in Modern Large Language Models

Filip Szatkowski[1,2] Patryk Będkowski[1] Alessio Devoto[3] Jan Dubiński[1,4]
Pasquale Minervini[5,6] Mikołaj Piórczyński[1,2] Simone Scardapane[3] Bartosz Wójcik[7]

[1]Warsaw University of Technology [2]IDEAS Research Institute [3]Sapienza University of Rome
[4]NASK [5]University of Edinburgh [6]Miniml.AI [7]Jagiellonian University

ICLR International Conference On Learning Representations

How to approach LLM activation sparsity?

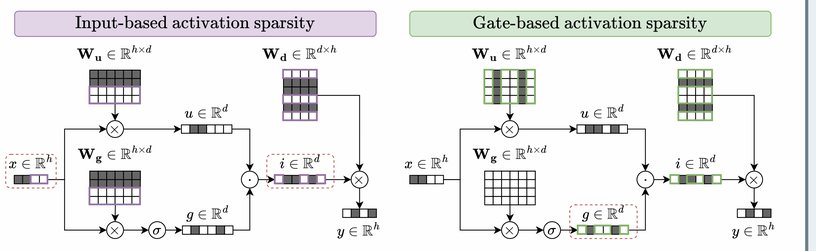
- Activations of DNNs contain many zero or near-zero values
- This leads to a lot of **wasted computation**
- Activation sparsity is trivial for ReLU-based models, but **LLMs use activations that do not induce exact sparsity**
- Prior work investigated several solutions that remain highly fragmented, so we devised **a more principled approach**
- We introduce a **unified, training-free framework to investigate functional activation sparsity** in modern LLMs

LLM MLP structure

Modern LLMs use **SiLU/GELU** activations in their FFN layers.
GLU FFN processes input $x \in \mathbb{R}^h$ as:

$$\mathcal{FFN}(x) = \mathbf{W_d}\big((\mathbf{W_u}x) \odot \sigma(\mathbf{W_g}x)\big)$$

Sparse FFN inference strategies



During matmul, activation sparsity methods skip columns or rows based on the activation vector values. We color-code these vectors as: x – *input*, $u{=}\mathbf{W_u}x$ – *up-proj.*, $g{=}\sigma(\mathbf{W_g}x)$ – *gate*, $i{=}u \odot g$ – *intermediate*, and also consider using *all FFN inputs*.

Top-p Framework & Critical Sparsity

Which of these strategies is the most suitable for modern LLMs? To study this, we propose a **simple, training-free** top-p sparsification rule: for vector $v \in \mathbb{R}^n$, we retain the largest-magnitude entries summing to fraction p of the L1 norm:

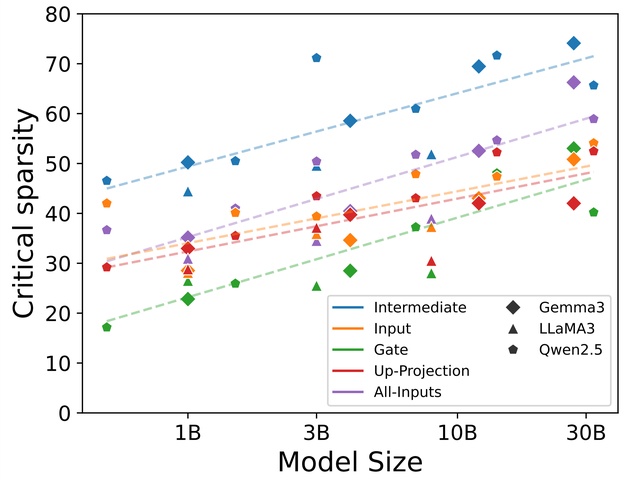
$$\text{top-p}(v) = m_p \odot v$$

$$m_p = \arg\min_m \|m\|_0 \;\text{ s.t. }\; \|m \odot v\|_1 \geq p \cdot \|v\|_1,\; m \in \{0,1\}^n$$

We define the **critical sparsity** as the *highest* sparsity at which the model retains ≥99% of its original performance.

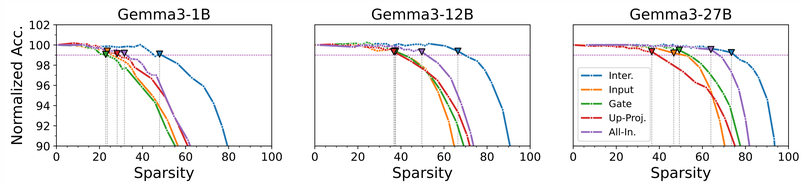
Larger Models Are More Robust to Sparsity

Critical sparsity **increases with model size**, which suggests that models will become **increasingly sparse** as scaling continues. We report the average for the base models on the suite of 9 tasks.



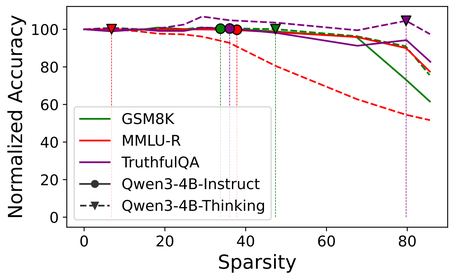
Sparsity Across FFN Components

Gate sparsity offers no clear advantage. **Input-based sparsity appears to be the most practical** predictor-free option, as it reaches high sparsity levels and is applicable to all matmuls in FFN.



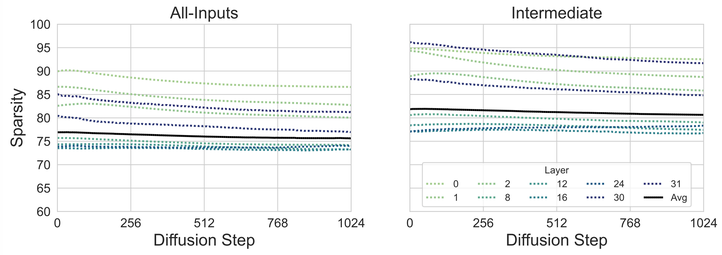
Thinking Models

Qwen3-4B Thinking shows robust sparsity on math and factuality. However, the thinking models can sometimes struggle with high sparsity, as it leads to longer reasoning which hits the max generation limit.



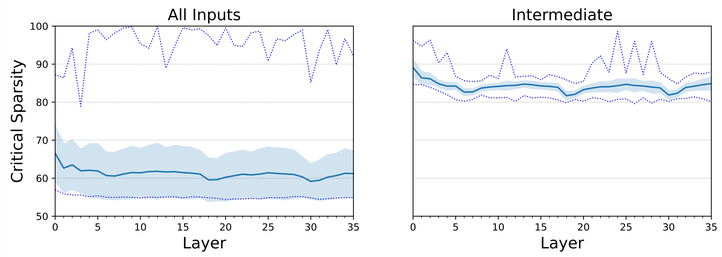
Activation Sparsity in Diffusion LLMs

We conduct the **first analysis of activation sparsity in diffusion-based LLMs**, using LLaDA-8B. LLaDA achieves **higher critical sparsity than autoregressive LLaMA**. The sparsity levels are stable across diffusion steps and similar across layers.



Expert-wise Sparsity in MoE Models

We investigate expert-wise sparsity in **Qwen3-30B-A3B** (8/128 experts active/token), plotting min, mean, and max across layers. **MoEs exhibit high sparsity robustness** like the dense models.



Key Takeaways

1. **Functional sparsity is universal** in modern LLMs.
2. **Larger models are increasingly sparse**.
3. Sparsification should be **data-free** to avoid overfitting.
4. **Input sparsification** is the most practical.
5. **Sparsity extends beyond dense LLMs**, and is highly prevalent in MoEs, reasoning models, and diffusion LLMs.

Stay in Touch

Scan the QR code for the full paper.
Write to us at `fmszatkowski@gmail.com`
Code link published in the paper.