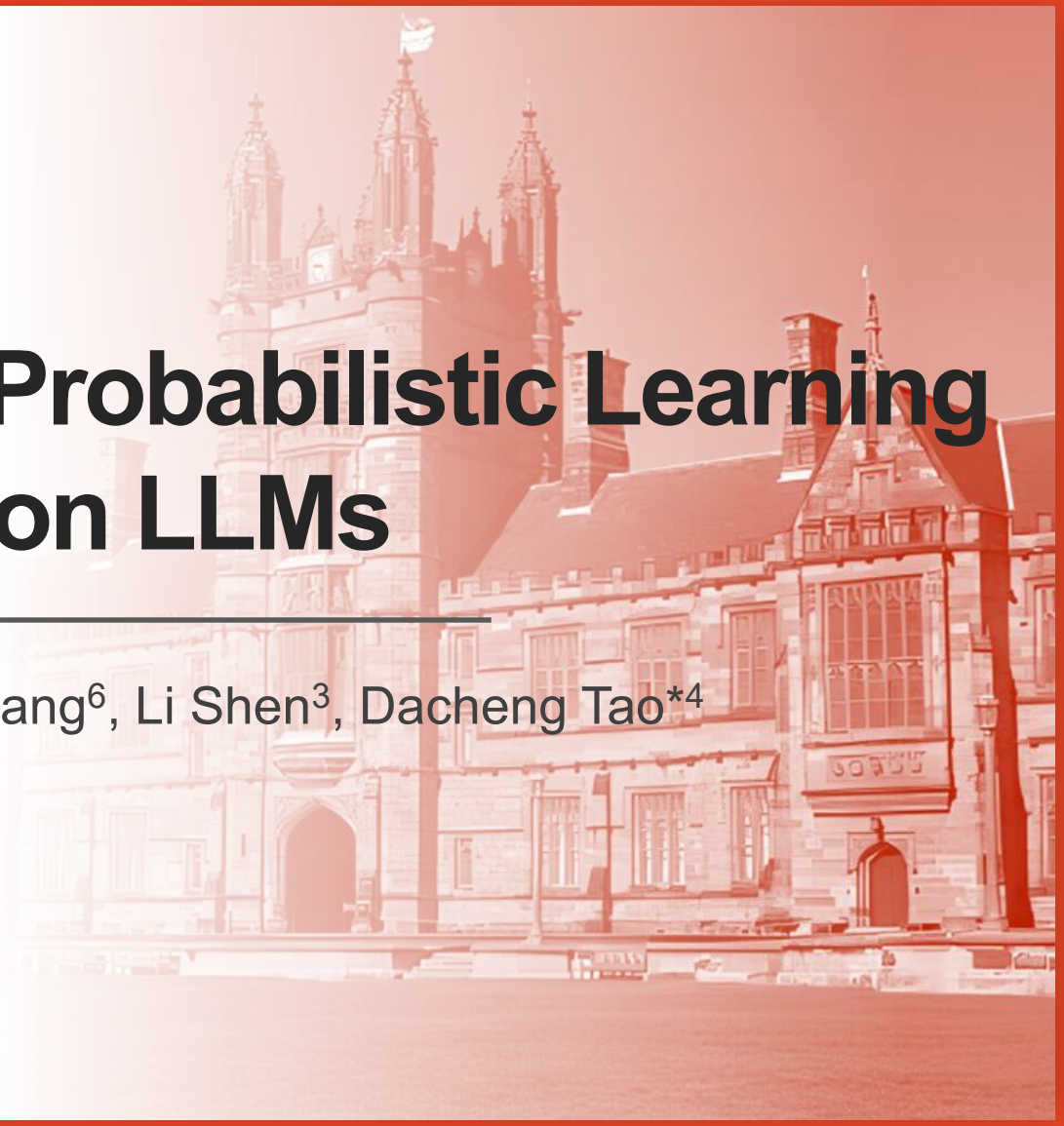
ICLR 2026

MaskPro: Linear-Space Probabilistic Learning for Strict (N:M)-Sparsity on LLMs

Yan Sun[12], Qixin Zhang[14], Zhiyuan Yu[5], Xikun Zhang[6], Li Shen[3], Dacheng Tao*[4]

[1] Equal contributions
[2] The University of Sydney
[3] Shenzhen Campus of Sun Yat-sen University
[4] Nanyang Technological University
[5] University of Science and Technology of China
[6] Royal Melbourne Institute of Technology University
* Corresponding author



MaskPro Framework

A general linear-space probabilistic learning for Strict (N:M)-Sparsity on LLMs

01

Semi-Sparsity

Differential Privacy in Federated Learning

02

Problem Setups

Target of the Protection Mechanism

03

Refined PGE

Hypothesis-Testing-Based Privacy Metric

04

MaskPro

Convergent bound for Local Model Training

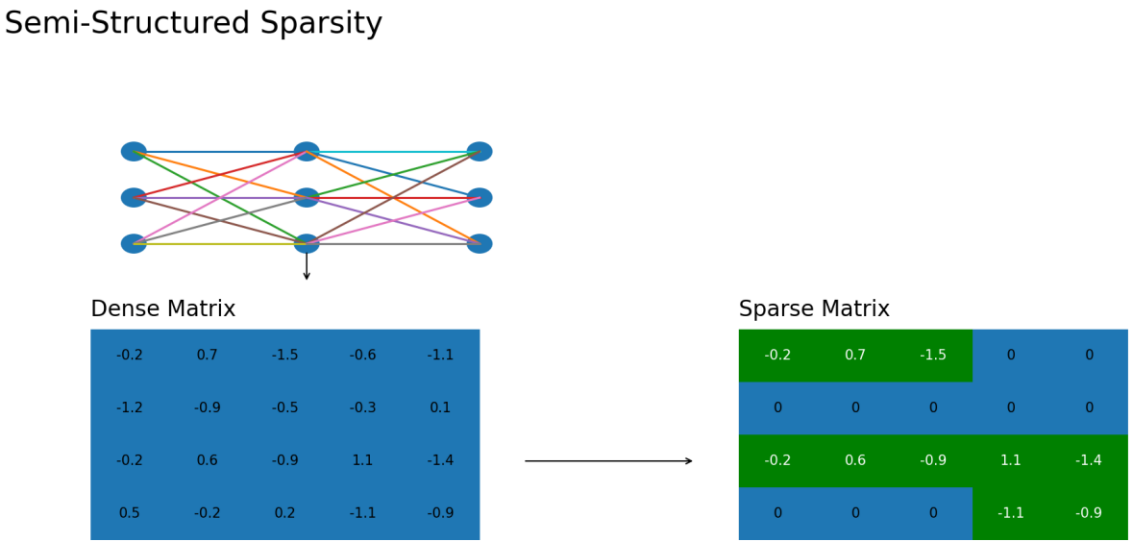
1. Semi-Sparsity

Semi-structured Sparsity in a 2D Weights

01

Semi-Sparsity

A semi-structured sparsity means a specific pattern with fixed activated weights in each weight group



2. Problem Setups

The core idea of semi-structured sparsity aims to divide the entire weights $\mathbf{w} \in \mathbb{R}^d$ into groups of M consecutive elements and then retain N effective weights for each group. More specifically, we can formulate the semi-structured sparsity as the following combinatorial optimization problem:

$$\mathbf{m}^\star = \underset{\mathbf{m}=\{\mathbf{m}_i | \mathbf{m}_i \in \mathcal{S}^{N:M}\}}{\arg\min} \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(\mathbf{m} \odot \mathbf{w}, \xi) \right], \tag{1}$$

where $f(\cdot)$ denotes the corresponding loss function, the symbol $\odot$ stands for the element-wise multiplication, $\xi \sim \mathcal{D}$ represents the minibatch sampled from the underlying distribution $\mathcal{D}$ and $\mathcal{S}^{N:M} = \left\{ \mathbf{m}_i \in \mathbb{B}^{1\times M} : \|\mathbf{m}_i\|_1 = N \right\}$ ($\mathbb{B}$ is the Boolean set and $\|\cdot\|_1$ denotes l_1 norm).

Recent advance provides a learning method to address Problem 1, named MaskLLM (Fang et al., 2024). Specifically, for each group of M consecutive weights, MaskLLM defines a categorical distribution with class probability $\left[p_1, p_2, \cdots, p_{|\mathcal{S}^{N:M}|}\right]$ where $\sum_i p_i = 1$, and each p_i represents the probability of the corresponding element in $\mathcal{S}^{N:M}$. By random sampling, if a certain mask performs better, it is reasonable to increase the probability of the sampled mask. Otherwise, the sampling probability should be decreased. Thus, Problem 1 can be transformed as,

$$\{p^\star(\mathbf{m}_i)\} = \underset{\{p(\mathbf{m}_i)\}}{\arg\min} \mathbb{E}_{\xi \sim \mathcal{D}, \mathbf{m}=\{\mathbf{m}_i | \mathbf{m}_i \sim p(\mathbf{m}_i)\}} \left[f(\mathbf{m} \odot \mathbf{w}, \xi) \right], \tag{2}$$

where $p(\mathbf{m}_i)$ is the categorical distribution of the i-th mask $\mathbf{m}_i$ over $\mathcal{S}^{N:M}$.

2. Problem Setups

Probabilistic Sum

Theorem 1 (Representation of N:M Sparsity)

$$\mathcal{S}^{N:M} = \left\{ \bigoplus_{i=1}^{N} \mathbf{a}_i : \ \mathbf{a}_i \in \{\mathbf{e}_1, \ldots, \mathbf{e}_M\}, \forall i \in [N] \ \textit{and} \ \mathbf{a}_1 \neq \mathbf{a}_2 \neq \ldots \neq \mathbf{a}_N \right\}, \quad (4)$$

where each $\mathbf{e}_j$ *denotes the* j*-th basis vector of the space* $\mathbb{R}^{1\times M}$.

$$\min_{\|\mathbf{p}_i\|_1=1, \forall i \in [\frac{d}{M}]} \Phi(\mathbf{p}) \coloneqq \mathbb{E}_{\{\mathbf{a}_{i,j}\}_{j=1}^{N} \sim \mathbf{p}_i, \xi \sim \mathcal{D}} \left[f\left(\bigoplus_{j=1}^{N} \mathbf{a}_{i,j} \odot \mathbf{w}_i, \xi \right) \right], \quad (6)$$

where $\{\mathbf{a}_{i,j}\}_{j=1}^{N} \sim \mathbf{p}_i$ represents the N-step sampling-without-replacement process guided by the categorical distribution $\mathbf{p}_i$. Note that representing all $\frac{d}{M}$ different categorical distributions $\{\mathbf{p}_i\}_{i=1}^{\frac{d}{M}}$ typically requires $\frac{d}{M} * M = d$ unknown parameters. Thus, by introducing randomness, the parameter scale of problem 6 can be further reduced from the previous Nd of problem 5 to a linear d.

2. Refined PGE

Ambiguity on Mask $\mathbf{m}_t$ and Minibatch ξ. The policy gradient updates logits based on the loss metric, aiming to encourage the logits to select masks that result in lower loss values. However, when the loss variation caused by mask sampling is significantly smaller than the loss variation caused by changing the minibatch, the loss metric alone cannot effectively distinguish whether the current mask is beneficial or detrimental. For example, we denote ξ_{low} as the minibatch whose loss is inherently low and ξ_{high} as the minibatch with high loss. Then we sample two masks and denote one that achieves lower loss by $\mathbf{m}_{\text{good}}$ and the other by $\mathbf{m}_{\text{bad}}$. There are typically two scenarios during training.

- $f(\mathbf{m}_{\text{good}} \odot \mathbf{w}, \xi_{\text{low}}) \leq f(\mathbf{m}_{\text{bad}} \odot \mathbf{w}, \xi_{\text{low}})$ and $f(\mathbf{m}_{\text{good}} \odot \mathbf{w}, \xi_{\text{high}}) \leq f(\mathbf{m}_{\text{bad}} \odot \mathbf{w}, \xi_{\text{high}})$.
- A bad case: $f(\mathbf{m}_{\text{bad}} \odot \mathbf{w}, \xi_{\text{low}}) \leq f(\mathbf{m}_{\text{good}} \odot \mathbf{w}, \xi_{\text{high}})$.

The first case is likely to hold in most cases, as a good mask can generally reduce the loss on most minibatches. But when the bad case occurs, Eq.(10) interprets that the lower-loss sample as the better one, yielding more erroneous learning on m_{bad}. To better illustrate this phenomenon, we randomly select two minibatches during the training of LLaMA-2-7B and extract the logits at the 500-th iteration. We then sample 1000 masks and plot their *loss distributions*, as shown in Figure 2. It is clearly observed that $f(\mathbf{m}_{\text{bad}} \odot \mathbf{w}, \xi_1) \leq f(\mathbf{m}_{\text{good}} \odot \mathbf{w}, \xi_2)$. Such disparities between minibatches are quite common, causing Eq.(10) to frequently encounter conflicting information when learning solely based on loss value $f(\mathbf{m} \odot \mathbf{w}, \xi)$.

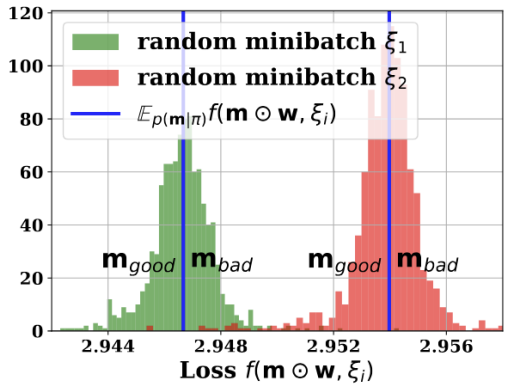


Figure 2: Loss-related misconceptions.

$$\pi_{t+1} = \pi_t - \eta \left(f(\mathbf{m}_t \odot \mathbf{w}, \xi) - f(\mathbf{m}_0 \odot \mathbf{w}, \xi) - \delta \right) \nabla \log \left(p(\mathbf{m}_t | \pi_t) \right),$$
$$\delta = \alpha \delta + (1 - \alpha) \left(f(\mathbf{m}_t \odot \mathbf{w}, \xi) - f(\mathbf{m}_0 \odot \mathbf{w}, \xi) \right).$$

2. MaskPro

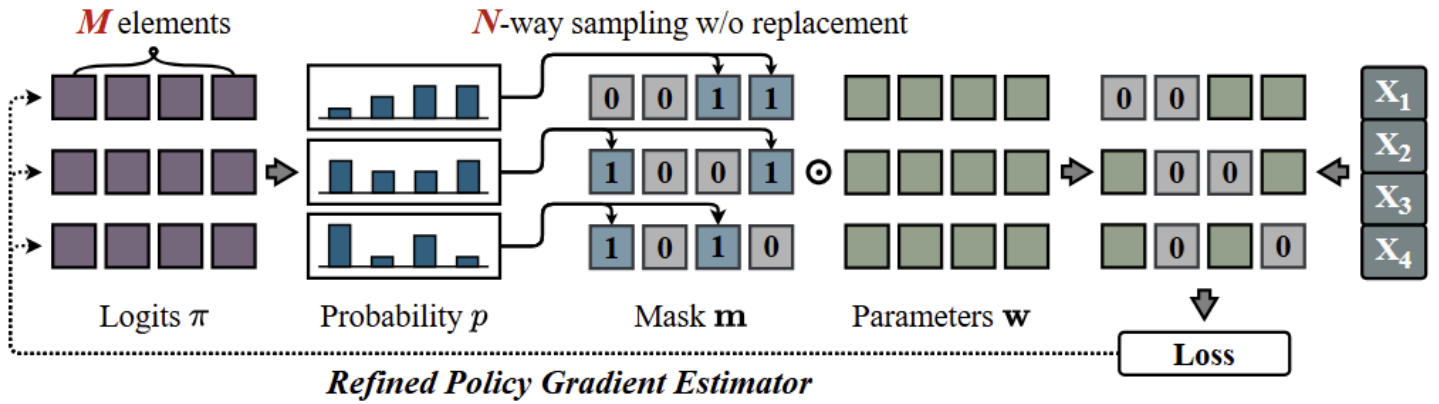


Figure 1: Implementation of our proposed MaskPro for learning (2:4)-sparse masks.

Algorithm 1 Learning (N:M)-Sparsity via MaskPro

Input: frozen weights $\mathbf{w}$, initial logits π_0, initial mask $\mathbf{m}_0$, learning rate η, smoothing coefficient $\alpha = 0.99$, smoothing tracker $\delta = 0$.

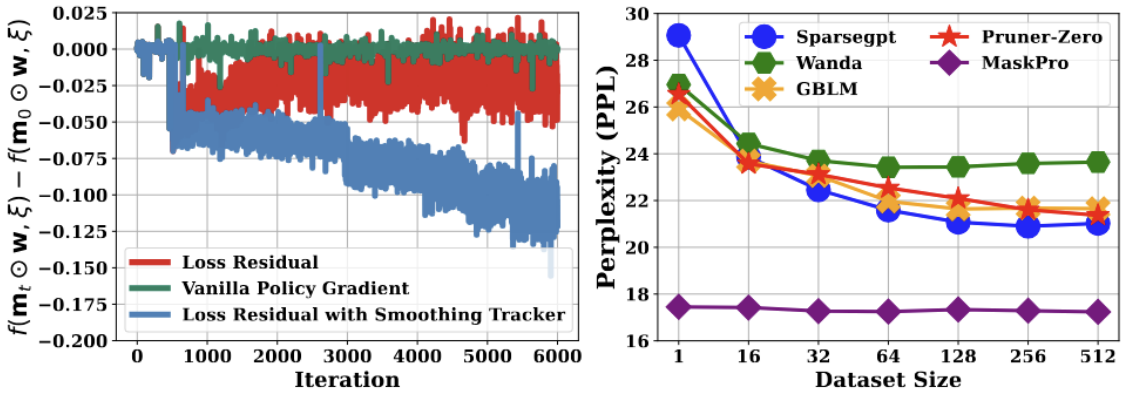
Output: learned logits π_T

for $t = 0, 1, 2, \cdots, T-1$ **do**
  sample a minibatch ξ for training
  reshape π_t into groups of M elements and calculate $p_t = \text{softmax}(\pi_t)$ for each group
  perform N-way sampling without replacement by p_t to generate the mask $\mathbf{m}_t$
  perform inference and calculate the loss residual $f(\mathbf{m}_t \odot \mathbf{w}, \xi) - f(\mathbf{m}_0 \odot \mathbf{w}, \xi)$
  update logits $\pi_{t+1} = \pi_t - \eta \left(f(\mathbf{m}_t \odot \mathbf{w}, \xi) - f(\mathbf{m}_0 \odot \mathbf{w}, \xi) - \delta\right) \nabla \log\left(p(\mathbf{m}_t | \pi_t)\right)$
  update the smoothing tracker $\delta = \alpha\delta + (1-\alpha)\left(f(\mathbf{m}_t \odot \mathbf{w}, \xi) - f(\mathbf{m}_0 \odot \mathbf{w}, \xi)\right)$
end for

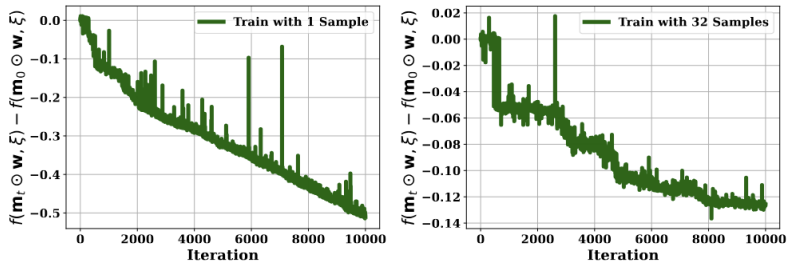
2. MaskPro

Table 1: Zero-shot evaluations of (2:4)-sparsity. In the test, we freeze weight updates and directly apply masks. The results corresponding to each model name reflects the evaluation of dense weights.

	Wiki.	HellaS.	RACE	PIQA	WinoG.	ARC-E	ARC-C	OBQA	Memory
GEMMA-7B	112.39	60.54	40.19	79.71	73.09	81.65	49.91	32.80	—
- MASKLLM	—	25.42	20.10	51.52	49.49	25.21	21.59	18.40	467.14 G
- MAGNITUDE	—	25.23	21.24	51.85	50.75	26.43	21.84	12.40	16.32 G
- SPARSEGPT	—	26.07	22.39	55.11	50.36	30.64	18.43	14.80	34.94 G
- WANDA	—	26.80	22.78	56.47	48.86	32.66	17.75	13.60	29.63 G
- GBLM	—	26.81	22.49	54.52	51.07	32.38	17.66	14.00	39.38 G
- PRUNER-ZERO	—	25.27	21.63	53.21	50.75	24.58	**22.70**	15.20	39.38 G
- **MaskPro**	—	**26.97**	**23.26**	**57.88**	**52.82**	**32.92**	22.65	**16.40**	48.63 G
VICUNA-1.3-7B	11.86	56.32	41.91	77.37	69.46	74.28	42.41	34.60	—
- MASKLLM	14.91	49.07	39.13	75.24	65.35	65.57	33.57	25.60	331.16 G
- MAGNITUDE	389.92	40.19	28.61	67.03	57.62	54.59	28.75	19.40	12.82 G
- SPARSEGPT	24.93	44.87	37.81	70.62	63.30	62.92	32.42	**25.00**	22.20 G
- WANDA	25.24	44.28	37.89	70.57	61.56	61.70	32.17	23.00	21.25 G
- GBLM	24.60	44.29	38.37	70.51	61.80	62.84	31.40	24.00	26.87 G
- PRUNER-ZERO	24.02	44.77	37.42	71.22	62.75	62.33	32.76	24.00	26.87 G
- **MaskPro**	**21.10**	**46.81**	**38.76**	**71.60**	**64.25**	**64.23**	**33.19**	24.80	35.90 G
LLAMA-2-7B	8.71	57.15	39.62	78.07	68.90	76.35	43.34	31.40	—
- MASKLLM	12.55	51.17	38.56	74.70	65.04	69.57	35.67	26.80	331.16 G
- MAGNITUDE	307.39	45.43	31.48	70.08	60.93	61.87	30.20	21.80	12.82 G
- SPARSEGPT	21.07	43.20	36.56	70.89	64.56	64.52	31.48	24.60	22.20 G
- WANDA	23.44	41.32	35.89	70.46	62.12	62.79	30.20	24.20	21.25 G
- GBLM	21.64	41.79	34.61	70.57	62.75	63.17	29.86	23.20	26.87 G
- PRUNER-ZERO	22.09	41.17	34.64	70.18	62.35	61.32	27.05	22.80	26.87 G
- **MaskPro**	**17.17**	**46.18**	**37.13**	**73.07**	**65.82**	**66.12**	**32.85**	**26.20**	35.90 G
DEEPSEEK-7B	9.70	56.94	39.62	79.27	70.40	75.25	43.60	32.60	—
- MASKLLM	12.90	51.73	39.14	75.95	65.80	68.10	35.32	25.80	339.56 G
- MAGNITUDE	285.06	40.97	28.52	69.75	60.06	54.92	27.56	20.80	13.13 G
- SPARSEGPT	19.12	45.58	**37.80**	73.94	65.43	66.37	32.94	24.80	22.50 G
- WANDA	19.68	45.38	35.12	73.56	63.14	65.49	32.00	22.80	21.55 G
- GBLM	19.55	45.34	36.17	73.99	62.98	65.82	32.85	23.60	27.98 G
- PRUNER-ZERO	20.71	44.93	35.22	73.23	62.12	64.94	30.89	23.20	27.98 G
- **MaskPro**	**17.97**	**47.78**	37.75	**74.72**	**65.59**	**66.74**	**33.49**	**28.60**	36.82 G



(a) Training Effectiveness of Three PGE Updates. (b) Training Performance of Different Dataset Size.



(a) Training set = 1. (b) Training set = 32.

(c) Training set = 128. (d) Training set = 320k.

2. MaskPro

Table 8: Zero-shot evaluations of (8:16)-sparsity on LLaMA2-7B.

	HellaS.	RACE	PIQA	WinoG.	ARC-E	ARC-C	OBQA	Avg.
LLaMA-7B	57.15	39.62	78.07	68.90	76.35	43.34	31.40	56.40
- MAGNITUDE	52.27	35.02	72.74	64.48	67.68	37.03	27.20	50.92
- SPARSEGPT	50.19	39.04	74.43	66.22	70.45	36.43	28.80	52.22
- WANDA	49.77	39.14	75.30	**66.61**	70.62	36.18	28.80	52.35
- GBLM	49.51	**39.90**	75.68	66.38	69.91	36.43	27.60	52.20
- PRUNER-ZERO	50.12	38.68	75.22	66.13	69.93	35.48	27.80	51.91
- **MaskPro**	**53.15**	39.23	**76.15**	66.56	**72.87**	**40.13**	**29.60**	**53.96**
LLaMA-13B	60.05	40.48	79.11	72.22	79.42	48.46	35.20	59.28
- MAGNITUDE	55.43	37.51	74.48	66.06	68.94	38.05	27.60	52.58
- SPARSEGPT	54.24	**40.38**	77.15	70.19	75.08	41.31	**31.00**	55.62
- WANDA	54.50	39.62	77.09	70.09	73.19	40.36	30.80	55.09
- GBLM	54.45	39.18	76.35	69.92	73.75	40.07	29.60	54.76
- PRUNER-ZERO	54.11	38.64	76.28	70.41	72.92	40.55	30.00	54.70
- **MaskPro**	**57.35**	39.92	**77.83**	**70.68**	**76.45**	**43.26**	30.60	**56.58**

Table 9: Zero-shot evaluations of (2:4)-sparsity on 13B/30B models.

	HellaS.	RACE	PIQA	WinoG.	ARC-E	ARC-C	OBQA	Avg.
LLaMA-13B	60.05	40.48	79.11	72.22	79.42	48.46	35.20	59.28
- MAGNITUDE	**50.10**	36.84	71.76	61.88	62.29	31.74	23.40	48.29
- SPARSEGPT	47.73	**38.95**	73.61	69.22	69.95	36.35	**27.40**	51.89
- WANDA	46.24	38.47	73.94	67.32	68.73	34.13	24.20	50.43
- GBLM	46.65	37.97	73.46	69.04	69.33	34.75	25.80	51.00
- PRUNER-ZERO	46.15	38.85	73.13	67.24	67.52	33.89	25.20	50.28
- **MaskPro**	49.24	38.91	**75.12**	**70.33**	**71.85**	**38.26**	**27.40**	**53.02**
LLaMA-30B	63.36	39.14	80.63	75.85	80.64	51.45	36.40	61.07
- MAGNITUDE	49.57	35.69	70.24	65.59	57.32	31.66	27.80	48.27
- SPARSEGPT	55.25	37.77	77.45	**73.68**	75.25	43.27	31.80	56.35
- WANDA	54.18	**40.00**	77.69	73.24	74.24	42.15	31.60	56.16
- GBLM	54.68	37.35	75.24	73.12	74.68	42.32	30.80	55.46
- PRUNER-ZERO	53.69	37.13	75.86	73.04	74.23	41.25	31.20	55.20
- **MaskPro**	**59.76**	37.28	**78.24**	73.32	**76.83**	**45.65**	**33.20**	**57.75**

Thank You!

Thanks for attention!

The University of Sydney