# $\mu$-MoE: Test-Time Pruning as Micro-Grained Mixture-of-Experts

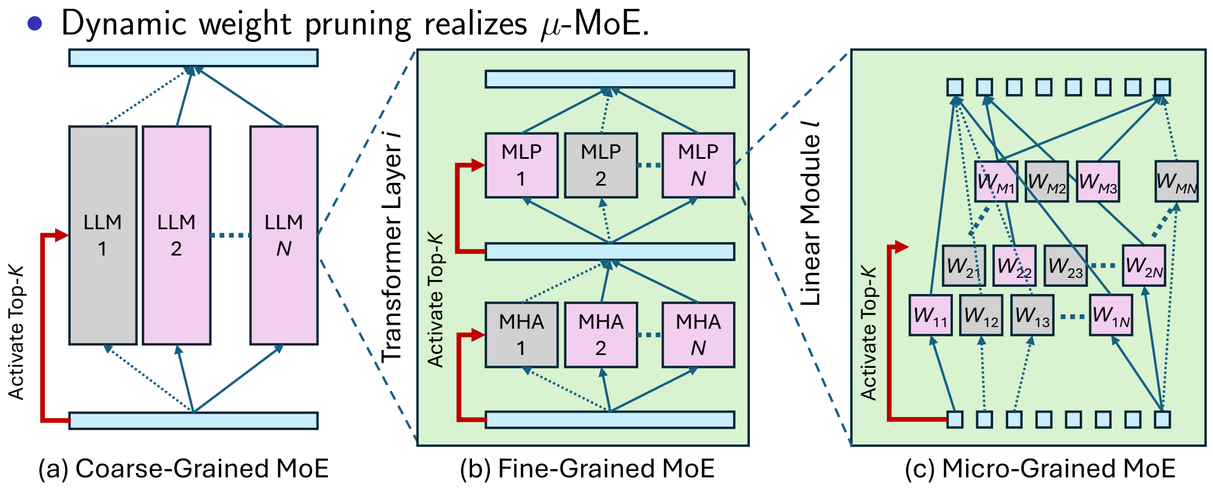
**Toshiaki Koike-Akino, Jing Liu, Ye Wang**
**Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA**

## Highlights

- We propose a mixture of micro-experts concept — **$\mu$-MoE** — to realize the finest-grained adaptation of large foundation models.
- We adopt low-complexity activation-aware pruning to realize **test-time LLM compression** as $\mu$-MoE.
- We tackle the **domain shift issue** caused by offline calibration required for baseline static pruning.
- We demonstrate the benefit of $\mu$-MoE over state-of-the-art methods for several LLM/VLM benchmarks.

## Coarse to Micro-Grained MoE

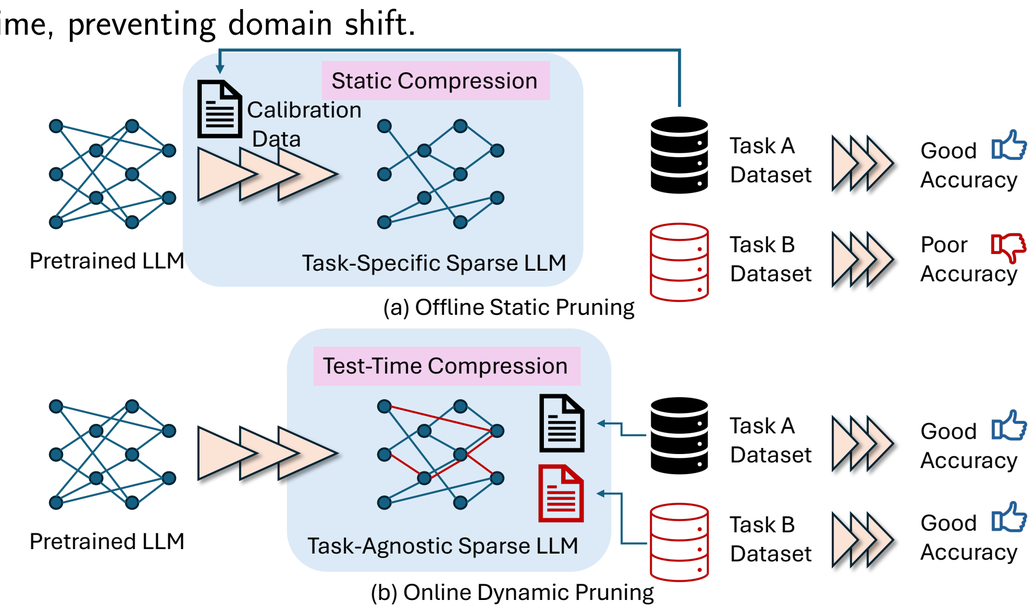
- Finest-grained experts would be a single-parameter weight multiplier within the linear modules of LLMs.
- Dynamic weight pruning realizes $\mu$-MoE.




(a) Coarse-Grained MoE (b) Fine-Grained MoE (c) Micro-Grained MoE

## Test-Time Pruning

- Test-time pruning can reduce the total floating-point operations (FLOPs) for inference computation.
- Online dynamic pruning can find prompt-dependent sparse structure at test time, preventing domain shift.




(a) Offline Static Pruning

(b) Online Dynamic Pruning

## Activation-Aware Instant Pruning

- Magnitude-based pruning: $\min_M \|W - W \odot M\|^2$ for $M \in \{0,1\}^{d' \times d}$ being a mask.
- Activation-aware pruning: $\min_{W'} \|(W - \hat{W})X\|^2$ s.t. $\|\hat{W}\|_0 \le \rho d d'$ for $X \in \mathbb{R}^{d \times T}$ being an input activation of token length $T$.
- SparseGPT uses a score: $S_{i,j} = |W_{i,j}|^2 / [\mathrm{Chol}[(XX^\top + \lambda I)^{-1}]]^2_{j,j}$, requiring at least cubic complexity for Cholesky decomposition and matrix inversion.
- Fast pruning — Wanda — uses a score: $S'_{i,j} = |W_{i,j}| \cdot \|X_{j,:}\|^2$, requiring quadratic complexity, and thus suited for instant pruning for every prompt.
- Test-time pruning can reduce the test-time complexity nearly proportional to $\rho$: $\frac{3dd'+dT+\rho dd'T}{dd'T} = \rho + \frac{3}{T} + \frac{1}{d'} \simeq \rho$.
- Original `sort` operation in Wanda can be simplified with `topk` or `kthvalue` with a marginal speedup at larger sizes $d$.

```
# W:(d',d), X:(d,Tc), kc=int((1-rho) * d)
S = W.abs() * X.norm(p=2, dim=-1)
val, _ = torch.kthvalue(S, dim=-1, k=kc)
W = torch.where(S > val[:,None], W, 0)
```

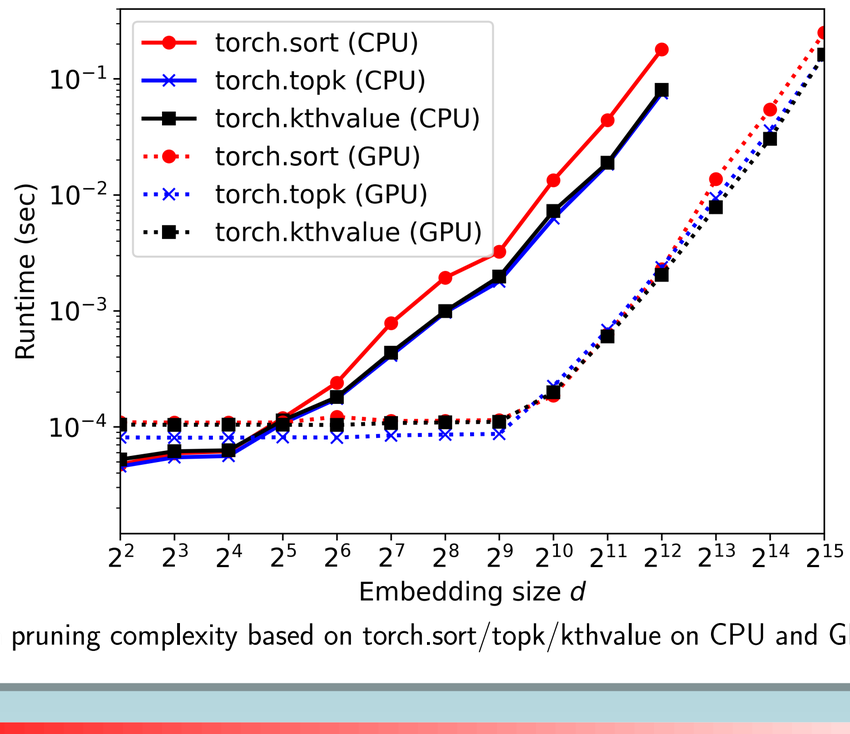



Figure: Wanda pruning complexity based on torch.sort/topk/kthvalue on CPU and GPU for $\rho = 0.5$.

## Experiments on LLM/VLM Benchmarks

- $\mu$-MoE shows best average performance on LLM/VLM benchmarks
- Offline calibration often suffers from domain shift issue.

Table: Perplexity (↓) of OPT models with different pruning methods at 60–40% active weights. Red-highlighted cells indicate that Wanda uses a matched calibration-test dataset. Bold-face letters indicate the best cases.

| Active Weights | 60% | | | | 50% | | | | 40% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test Dataset | WT2 | PTB | C4 | Avg | WT2 | PTB | C4 | Avg | WT2 | PTB | C4 | Avg |
| OPT-125M (WT2: 27.7, PTB: 39.0, C4: 26.6, Avg: 31.1) | | | | | | | | | | | | |
| Magnitude Prune | 43.9 | 71.9 | 39.8 | 51.9 | 90.9 | 168.6 | 71.9 | 110.4 | 533.2 | 906.3 | 349.9 | 596.5 |
| Wanda (WT2 Calib) | 30.8 | 44.2 | 30.3 | 35.1 | 37.1 | 56.3 | 37.5 | 43.6 | 65.4 | 106.5 | 37.5 | 81.6 |
| Wanda (PTB Calib) | 32.4 | 43.8 | 31.7 | 36.0 | 44.0 | 52.8 | 42.1 | 46.3 | 89.7 | **86.5** | 87.6 | 87.9 |
| Wanda (C4 Calib) | 30.7 | 44.4 | 29.3 | 34.8 | 39.1 | 57.1 | 34.8 | 43.7 | 75.1 | 104.3 | 60.4 | 80.0 |
| **$\mu$-MoE** | **30.3** | **43.3** | **28.6** | **34.1** | **35.8** | **51.8** | **32.5** | **40.1** | **61.0** | 87.5 | **52.3** | **66.9** |
| OPT-6.7B (WT2: 10.9, PTB: 15.8, C4: 12.7, Avg: 13.1) | | | | | | | | | | | | |
| Magnitude Prune | 16.3 | 23.9 | 17.0 | 19.1 | 532.2 | 281.6 | 257.4 | 357.1 | 9490.4 | 6743.4 | 6169.1 | 7467.6 |
| Wanda (WT2 Calib) | 11.0 | 17.2 | 14.2 | 14.2 | 12.0 | 19.0 | 16.3 | 15.8 | 15.1 | 25.0 | 22.8 | 21.0 |
| Wanda (PTB Calib) | 11.2 | 16.3 | 14.6 | 14.0 | 13.6 | 17.1 | 17.6 | 16.1 | 19.4 | 20.6 | 25.8 | 21.9 |
| Wanda (C4 Calib) | **10.9** | 16.4 | 13.3 | 13.5 | 11.9 | 17.9 | 14.3 | 14.7 | 15.3 | 23.6 | 18.2 | 19.0 |
| **$\mu$-MoE** | 11.1 | **16.1** | **13.0** | **13.4** | **11.7** | **16.7** | **13.5** | **14.0** | **13.7** | **19.7** | **15.7** | **16.4** |
| OPT-13B (WT2: 10.1, PTB: 14.5, C4: 12.1, Avg: 12.2) | | | | | | | | | | | | |
| Magnitude Prune | 59.8 | 78.4 | 44.5 | 60.9 | 2960.9 | 5406.3 | 3432.5 | 3933.2 | 112900.6 | 28381.4 | 13734.1 | 51672.0 |
| Wanda (WT2 Calib) | 10.7 | 15.8 | 13.6 | 13.3 | 12.0 | 18.7 | 15.7 | 15.5 | 15.5 | 25.3 | 20.7 | 20.5 |
| Wanda (PTB Calib) | 10.9 | 15.2 | 14.2 | 13.4 | 13.4 | 16.8 | 17.4 | 15.9 | 20.6 | 20.5 | 24.6 | 21.9 |
| Wanda (C4 Calib) | 10.9 | 15.2 | 14.2 | 13.4 | 13.4 | 16.8 | 17.4 | 15.9 | 20.6 | 20.5 | 24.6 | 21.9 |
| **$\mu$-MoE** | **10.6** | **15.0** | **12.3** | **12.7** | **11.5** | **16.4** | **12.9** | **13.6** | **14.3** | **20.2** | **14.6** | **16.4** |

Table: Complexity analysis with calflops for OPT-13B models with $\mu$-MoE.

| Active Weights | FLOPs | MACs |
|---|---|---|
| 100% | 3.29T | 1.64T |
| 80% | 3.21T | 1.33T |
| 60% | 2.55T | 999G |
| 40% | 1.90T | 671G |
| 20% | 1.24T | 342G |

Table: Accuracy in percent (↑) on ScienceQA dataset of LLaVA-7B model ( Full-weight accuracy is 70.03%). TextVQA is used for calibration.

| Active Weights | 60% | 50% | 40% |
|---|---|---|---|
| Magnitude Prune | 62.70 | 41.57 | 0.21 |
| SparseGPT | 63.83 | 53.20 | 39.05 |
| Wanda | 65.17 | 57.96 | 32.45 |
| **$\mu$-MoE** | **67.15** | **59.84** | **41.12** |

Table: Accuracy in percent (↑) on TextVQA dataset of LLaVA-7B model ( Full-weight accuracy is 61.32%). ScienceQA is used for calibration.

| Active Weights | 60% | 50% | 40% |
|---|---|---|---|
| Magnitude Prune | 54.12 | 45.56 | 24.62 |
| SparseGPT | 53.37 | 47.42 | 28.27 |
| Wanda | 55.80 | 52.36 | 39.27 |
| **$\mu$-MoE** | **57.16** | **54.65** | **46.97** |

**ICML2025 ES FoMo III Workshop; Paper No. 48;** **Email: koike@merl.com; jiliu@merl.com; yewang@merl.com**