TAPIOCA

Why Task-Aware Pruning Improves OOD Model Capability

Krish Sharma[1,*], Omar Naim[1,*], Soumadeep Saha[1], Vinija Jain[2], Aman Chadha[3], Nicholas Asher[4]

[1]Université de Toulouse [2]Meta [3]Apple [4]CNRS * Equal contribution

Meta Apple

NeurIPS 2026

Task-Aware Layer Elimination

TALE is a greedy task-aware layer pruning approach [1].

The geometric picture

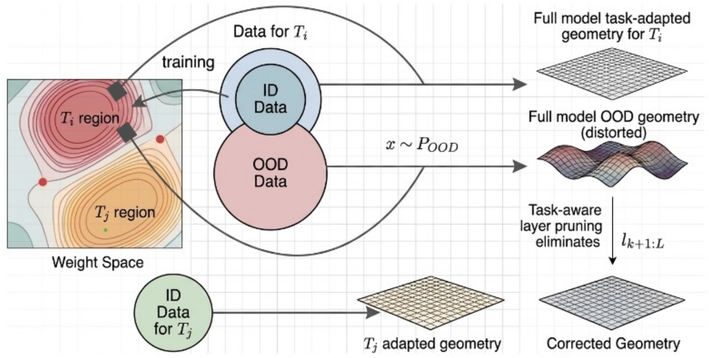


Figure 1. ID inputs induce a task-adapted geometry. OOD inputs distort it; task-aware pruning can move representations back toward that geometry.

1 When does pruning help?

Train a 12-layer transformer for in-context regression, $f(x) = ax + b$, with $a, b \sim \mathcal{U}(-1, 1)$. At test time, widen the coefficient range to $\mathcal{U}(-\sigma, \sigma)$; keep $x \sim \mathcal{U}(-1, 1)$.

Evaluation	Full model	Pruned
ID: $\sigma = 1$	0.000008	0.011770
OOD: $\sigma = 2$	0.04054	**0.03440**

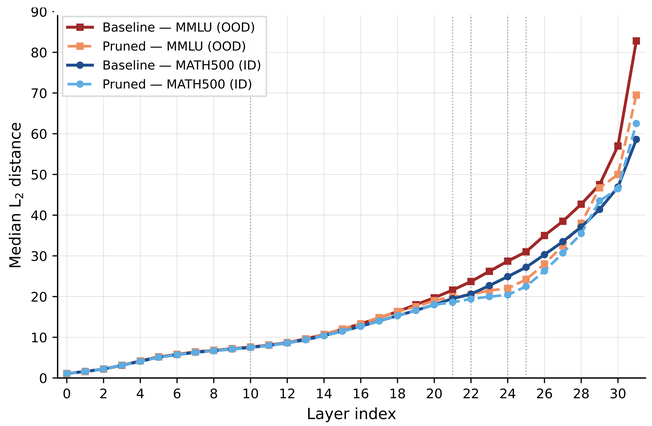
Mean squared error (lower is better). The same model, pruned for $\sigma = 2$, is evaluated on both distributions.

Lower OOD error, but much worse ID error. On ID validation data, TALE selects no deletions.

OOD gains transfer. Selecting layers at $\sigma = 2$ or 3 also improves performance at $\sigma = 4, \ldots, 10$, with MSE ratios of 0.82–0.93 relative to the baseline.

2 Pruning realigns distances

OOD token distances diverge from the adapted-task profile in later layers. Removing TALE-selected layers reduces this mismatch.



Math-specialized Llama 3.1 8B; 200 prompts per dataset. Distance from the final token to preceding tokens.

3 Covariance: beyond scale

In regression, shift x from $\mathcal{U}(0, 1)$ to $\mathcal{U}(-1, 0)$: **same magnitude, different distribution.** Trace-normalized covariances isolate shape and orientation. A score of zero means matching covariance.

$$M = \widetilde{\Sigma}_{\text{ID}}^{-1/2} \widetilde{\Sigma}_{\text{OOD}} \widetilde{\Sigma}_{\text{ID}}^{-1/2}, \qquad S = \frac{1}{d} \sum_j |\log \lambda_j(M)|.$$

Final-layer mismatch	Full	Pruned
ID vs. ID	0.132 ± 0.027	0.139 ± 0.018
ID vs. OOD	$\mathbf{0.642} \pm 0.029$	$\mathbf{0.520} \pm 0.024$

Removing layer 1 brings OOD covariance closer to ID. The gap decreases, but remains above the ID–ID reference. This supports alignment of variance structure, beyond a change in activation scale.

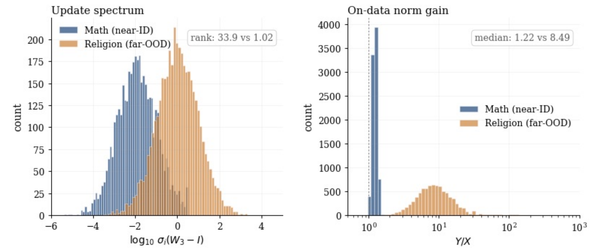
4 The pattern holds at scale

Math specialists, fine-tuned on NuminaMath-CoT. Accuracy (%, mean ± SD); gains in percentage points.

Model	Evaluation	Full	TALE	Gain
Llama 3.1 8B	MATH500 (ID)	87.5 ± 0.3	87.5 ± 0.2	0.0
Llama 3.1 8B	MMLU Math	36.7 ± 0.5	44.1 ± 0.4	**+7.4**
GPT-OSS 120B	MMLU Math	47.0 ± 0.4	54.0 ± 0.3	**+7.0**
Llama 3.1 8B	BoolQ	83.7 ± 0.4	86.7 ± 0.3	+3.0
GPT-OSS 120B	BoolQ	93.0 ± 0.3	94.0 ± 0.2	+1.0

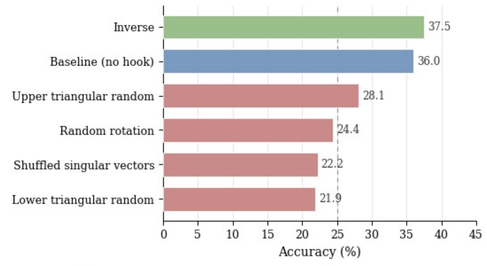
5 Same layer, different action

Fit a linear surrogate to layer 3 on two MMLU slices. With the same weights, far-OOD inputs produce a near rank-one update and much larger norm gain.



6 Test the geometry directly

An inverse surrogate after layer 24 reaches **37.5%**, close to deletion (37.8%; full model: 36.0%). Random maps with similar norm ratios destroy accuracy.



2-shot MMLU high-school mathematics. Norm ratios: 1.04–1.14. **Task-relevant structure matters, not norm alone.**

Layer usefulness depends on both the task and the input distribution.

[1] Naim et al. *TELL-TALE: Task Efficient LLMs with Task Aware Layer Elimination.* Findings of ACL 2026, pp. 22616–22638.

Supported by ANITI, SARER, Summ-RE, COCOBOTS and CALMIP.

Scan for the paper

Code github.com/omyokun/TAPIOCA