Autonomy-of-Experts Models

Ang Lv Ruobing Xie Yining Qian Songhao Wu Xingwu Sun
Zhanhui Kang Di Wang Rui Yan

ICML 2025

Research Question

Overlooked! The Router–Expert Separation in MoE Models

Routers assign tokens to experts without knowing their true capabilities—essentially predicting without labels. Poor routing leads to misaligned tokens, increasing loss. To reduce the loss:

- Experts may overfit to mismatched tokens, drifting from their specialization.
- OR, routers must improve through costly trial-and-error.

Motivation

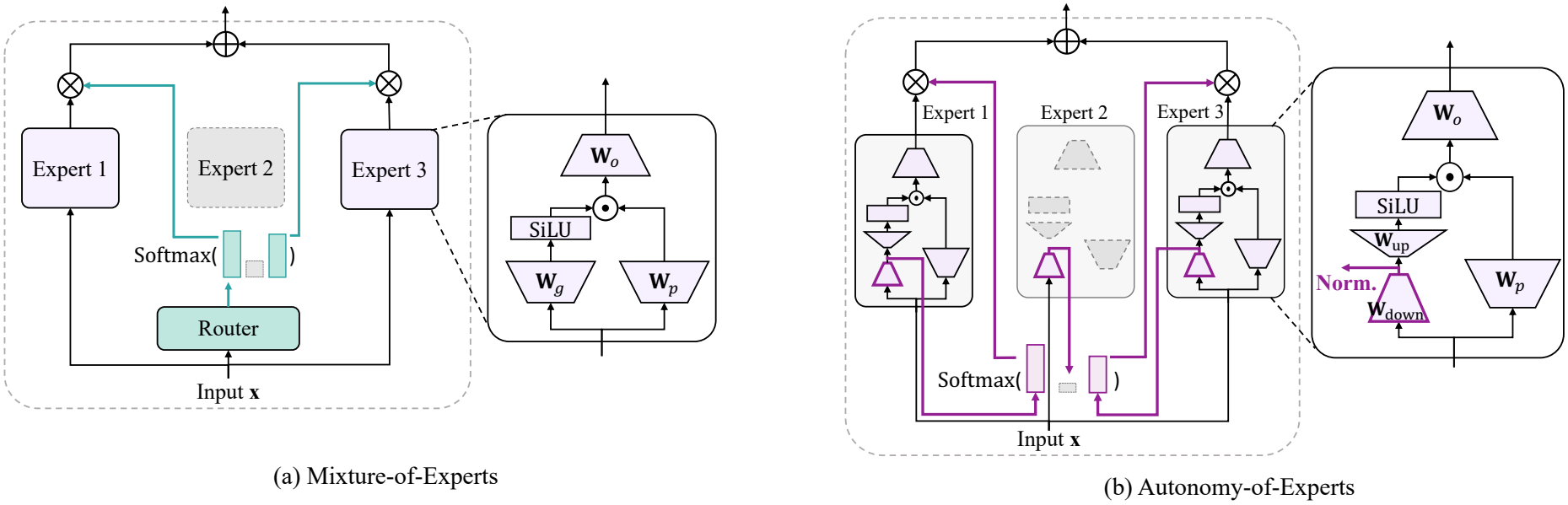
Experts "know" what they're good at—their activation norm reflects this. We remove routers from Mixtral 8×7B and select experts during inference based on the internal activation norms of specific nodes in the computational graph. The MMLU accuracy (5-shot) and time cost in minutes are given. Without any parameter updates, selecting experts by norms can largely preserve accuracy. However, this naive approach results in dense activation, leading to significantly higher computational cost.

Table: Experts "know" what they're good at.

Node for Norm	Acc. (Time)
$\mathbf{x}\mathbf{W}_g$	64.23 (42.70)
$\mathbf{x}\mathbf{W}_p$	62.06 (42.73)
$\texttt{SiLU}(\mathbf{x}\mathbf{W}_g)$	61.71 (43.88)
$\texttt{SiLU}(\mathbf{x}\mathbf{W}_g) \odot \mathbf{x}\mathbf{W}_p$	66.64 (75.53)
Experts' Final Outputs	66.66 (76.15)
Performance w. Router	70.35 (24.30)

A New MoE Paradigm: Autonomy of Experts

Based on this insight, AoE introduces structural changes for both efficiency (maintaining sparsity) and effectiveness.



(a) Mixture-of-Experts

(b) Autonomy-of-Experts

Figure: In an AoE model, experts operate autonomously. They are ranked based on their internal activation norms, and only the top-activated experts continue processing, while the others are terminated.

Improved Performance, Lower Loss, More Balanced, Comparable Efficiency

Table: AoE variants outperform the best traditional MoE. (247M/732M active).

Configuration	ARC-E	PIQA	SIQA	WINO	HELLA	MNLI	QNLI	SST2	AVG.
1 Traditional MoE	39.90	58.43	35.67	52.09	27.98	33.09	49.28	49.66	43.28
2 + $\mathcal{L}_{\text{aux}}$	40.74	58.49	36.13	51.30	28.11	32.67	50.23	51.83	43.68
3 + $\mathcal{L}_{\text{aux}}$ + Factorized $\mathbf{W}_g$	40.45	58.65	36.75	52.09	28.03	32.55	50.08	51.03	43.70
4 + $\mathcal{L}_{\text{aux}}$ + Large Router	41.41	57.62	36.64	52.33	28.34	33.18	49.53	50.69	43.71
5 AoE ($d_{\text{low}} = 64$)	39.77	58.71	35.31	52.33	28.29	32.78	50.27	52.98	**43.81**
6 + $\mathcal{L}_{\text{aux}}$	42.17	57.67	36.75	50.75	28.15	34.06	50.49	53.10	**44.12**
7 AoE ($d_{\text{low}} = 128$)	40.70	59.41	36.64	52.09	28.06	34.38	50.69	53.21	**44.39**
8 + $\mathcal{L}_{\text{aux}}$	41.33	58.65	36.80	50.75	28.40	33.71	49.55	53.10	**44.04**
9 AoE ($d_{\text{low}} = 256$)	41.08	58.81	36.44	51.70	28.23	32.24	50.54	53.90	**44.12**
10 + $\mathcal{L}_{\text{aux}}$	41.16	58.32	36.80	53.04	28.37	32.78	50.61	54.59	**44.46**
11 AoE ($d_{\text{low}} = 512$)	40.57	57.89	36.75	50.59	28.38	32.71	49.72	53.56	**43.77**
12 + $\mathcal{L}_{\text{aux}}$	41.16	57.83	36.75	52.09	28.30	34.92	50.67	50.92	**44.08**

Table: Models trained using alternative expert-selection strategies.

Strategy	Model	ARC-E	PIQA	SIQA	WINO	HELLA	MNLI	QNLI	SST2	AVG.
Top-P	Traditional MoE	41.08	57.96	37.46	50.36	28.25	32.79	50.39	52.64	43.87
	AoE	41.04	58.65	36.39	51.07	28.35	32.96	51.46	54.36	**44.29**
Expert-Choice	Traditional MoE	40.91	59.09	37.26	50.75	28.09	32.11	50.12	52.75	43.89
	AoE	41.58	58.22	37.21	53.04	28.44	33.83	50.54	50.46	**44.17**

Table: For 4B-parameter LLMs (with 1.18B active), AoE exhibits better downstream Acc. than MoE models.

Model	ARC-E	PIQA	SIQA	WINO	HELLA	MNLI	QNLI	SST2	AVG.
Traditional MoE	53.70	65.40	39.10	51.54	35.80	32.19	49.77	57.00	48.06
AoE	55.98	65.61	39.87	52.57	36.77	35.39	50.05	61.93	**49.80**

Lower Loss

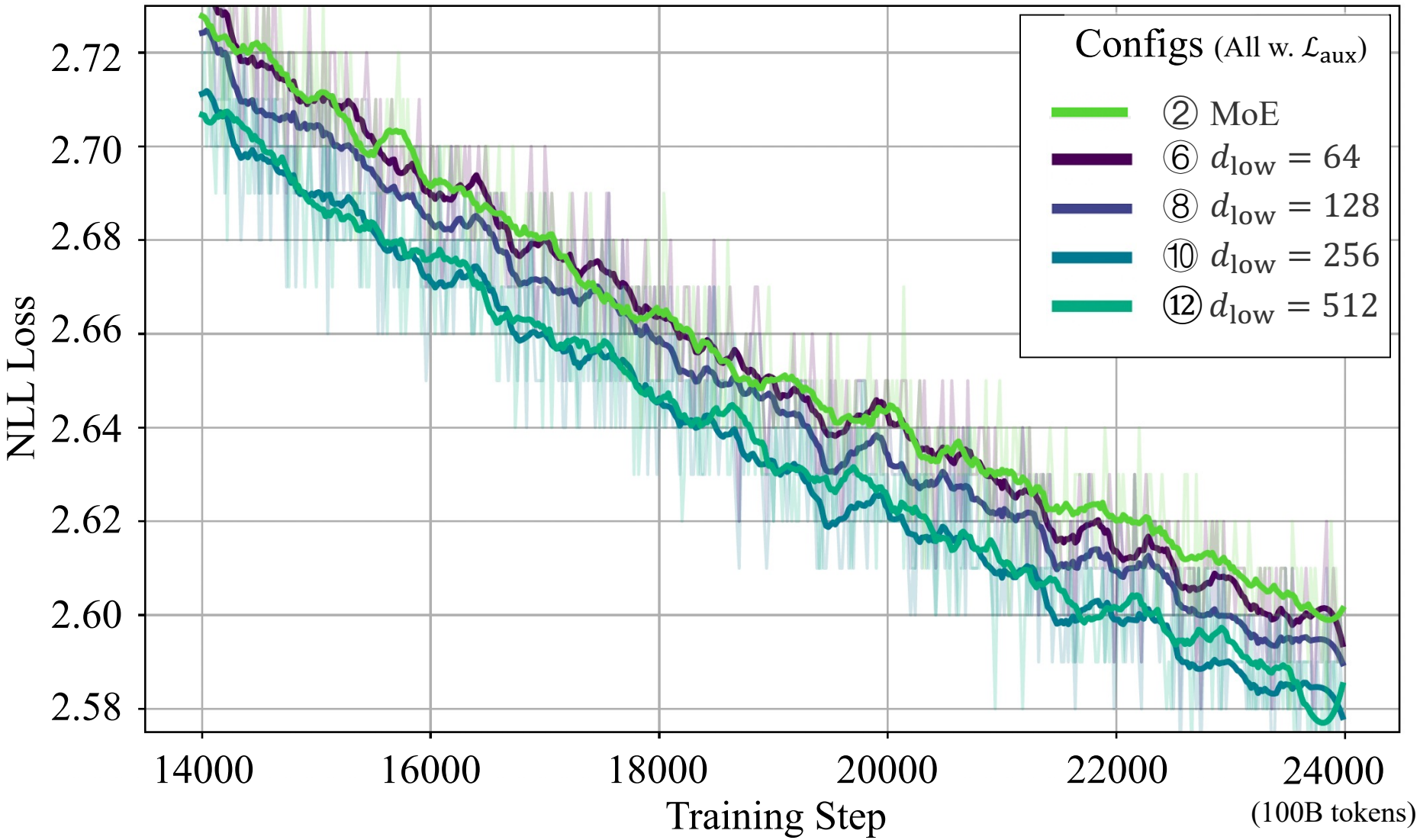


Figure: Pre-training LM loss.

Better Load Balance

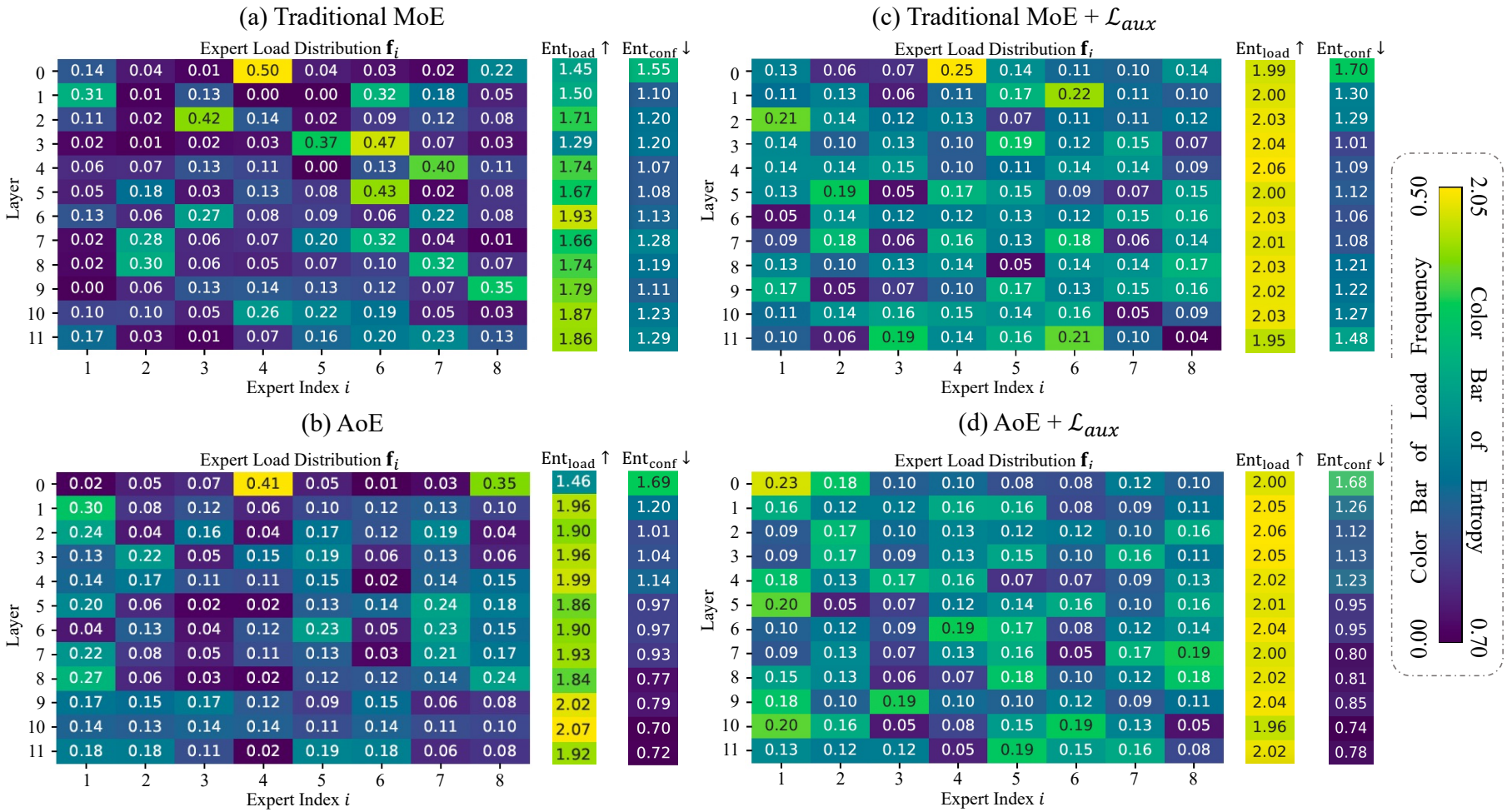


Figure: AoE achieves a more balanced expert load, evidenced by: (1) $\mathbf{f}_i$ values closer to 0.125, (2) higher load entropy (Entload), and (3) lower confidence entropy (Ent_{conf}).

Thank you!