

SoftMoE: Soft Differentiable Routing for Mixture-of-Experts in LLMs

Mikołaj Zasada, Łukasz Struski, Jacek Tabor, Marcin Kurdziel

ICML 2026 July 6th – 11th



Github

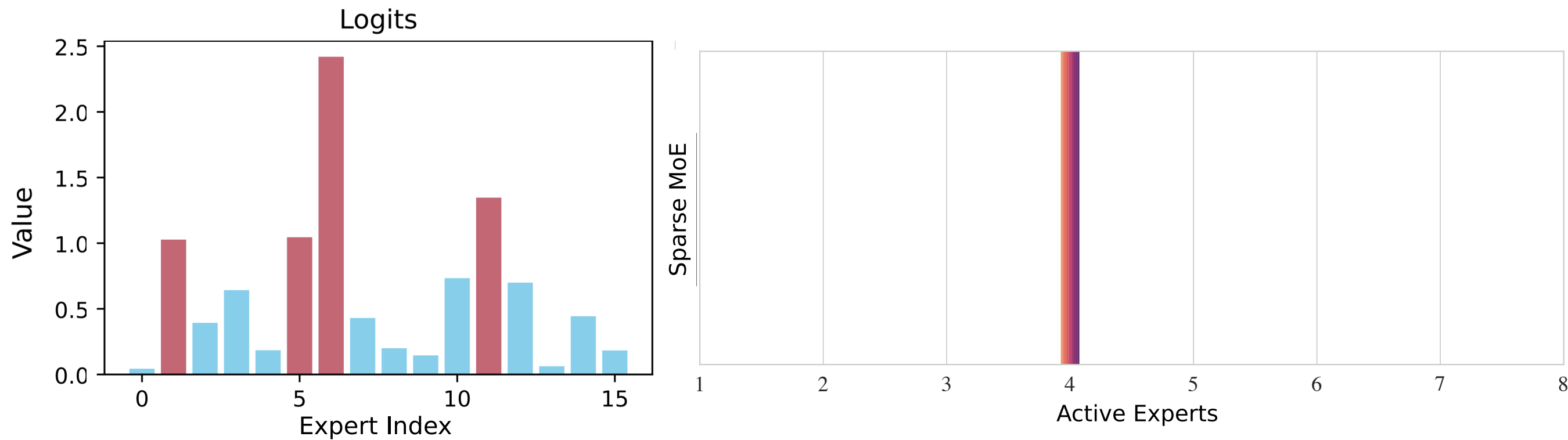


Motivation

Limitations:

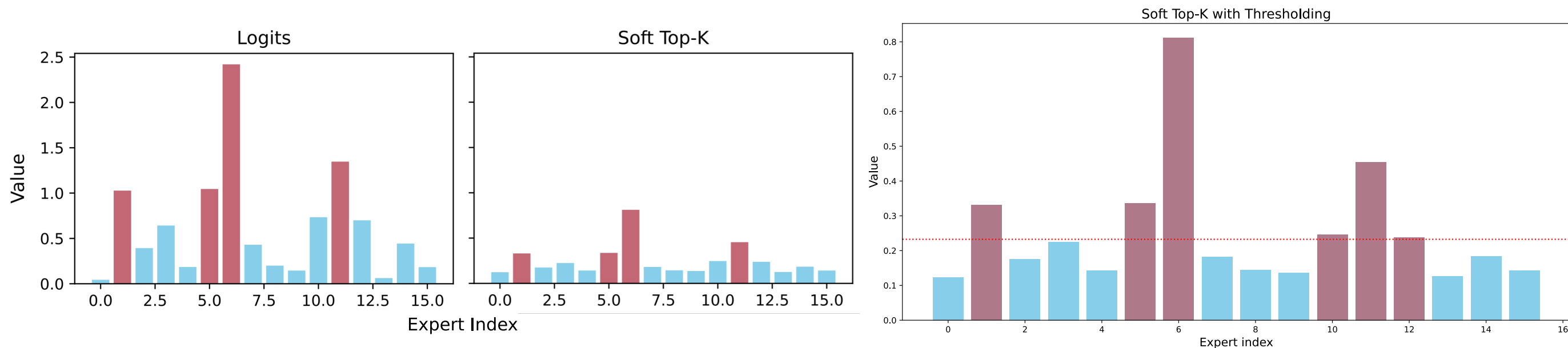
- Activates exactly K experts per token
- Compute cannot adapt to input complexity
- The optimal value of K must be chosen manually
- Expert capacity cannot be allocated across layers

Hard Top-K Routing



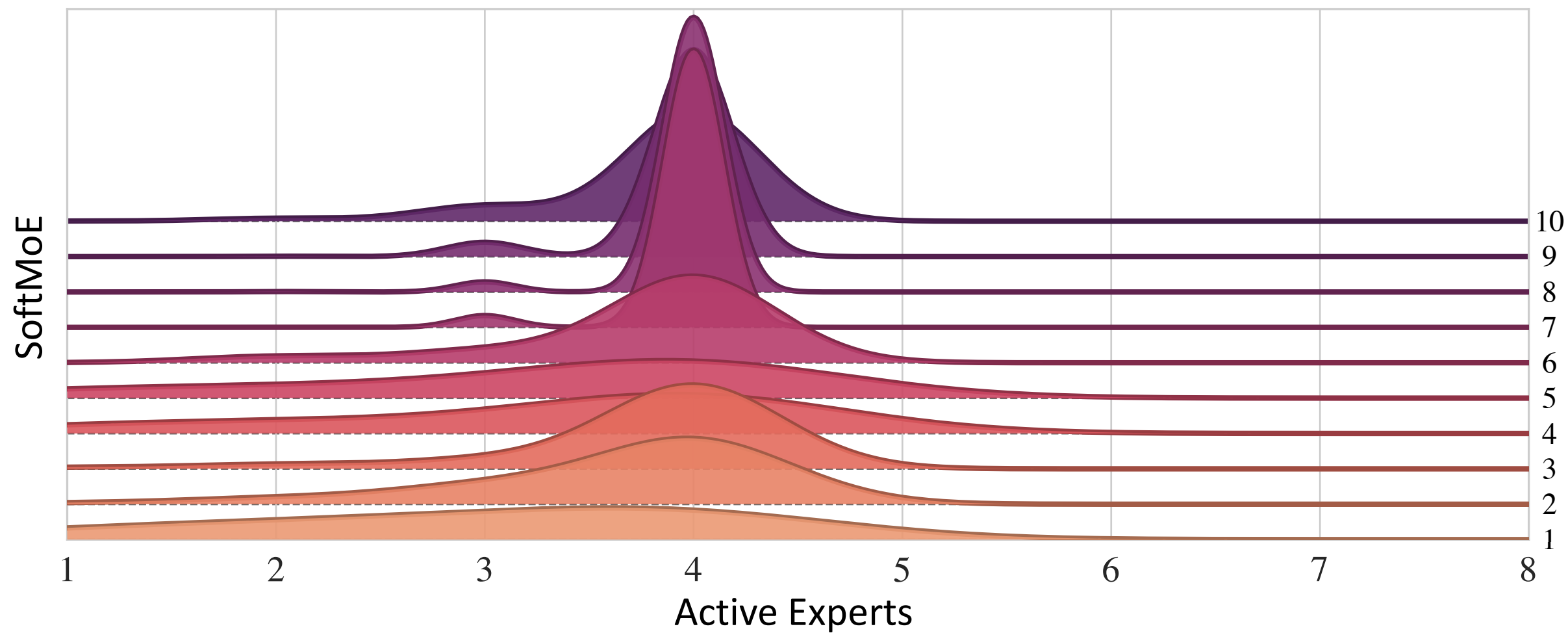
Relaxed Top-K operator

$$G(x) = \text{SoftTopK}(xW_g, \alpha, k)$$



Struski et al. "LapSum--One Method to Differentiate Them All: Ranking, Sorting and Top-k Selection." ICML 2025

Soft routing naturally produces variable expert usage



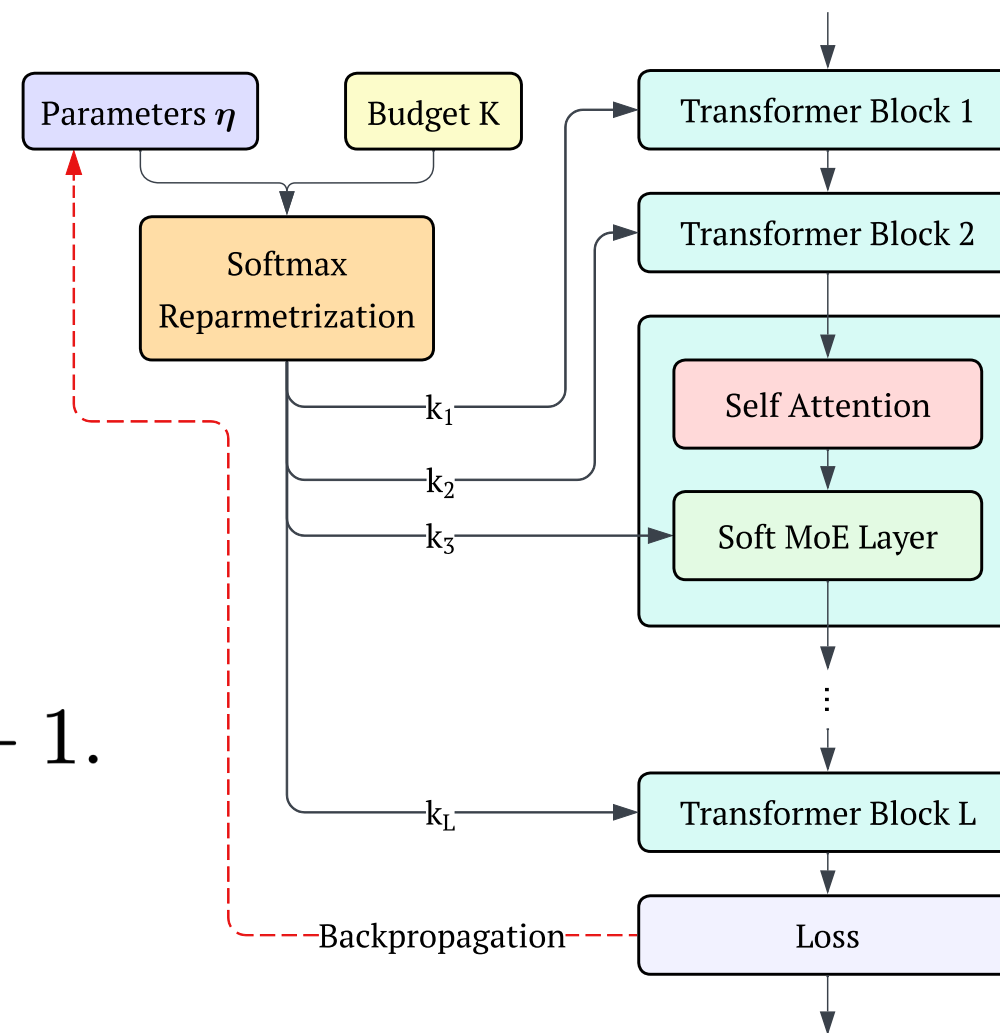
Can the model learn how much compute it needs?

$$G(x) = \text{SoftTopK}(xW_g, \alpha, \boxed{k})$$

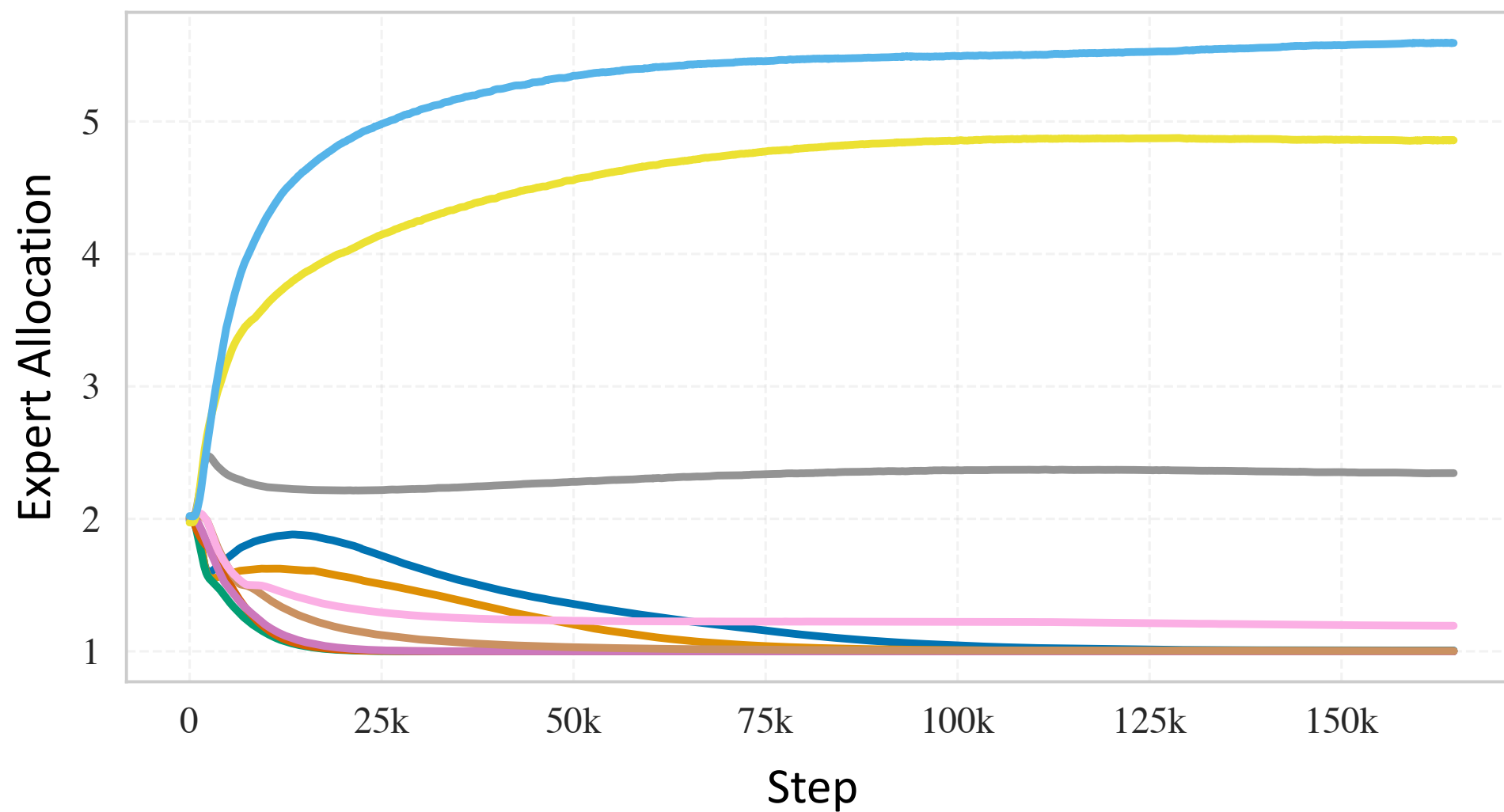
$$k_l \geq 1, \quad l \in \{1, \dots, L\}$$

$$\sum_{l=1}^L k_l = K$$

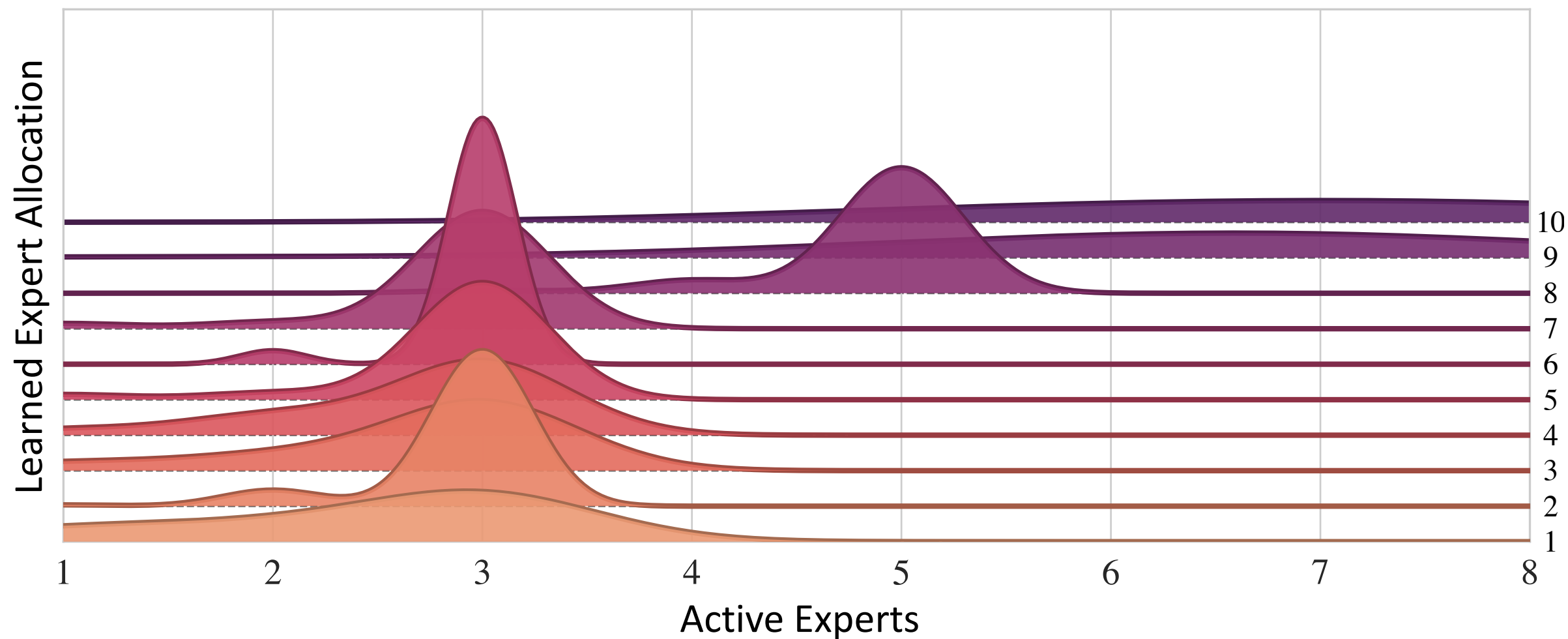
$$\pi = \text{softmax}(\eta), \quad k_l = \pi_l \cdot (K - L) + 1.$$



The model prefers later layers



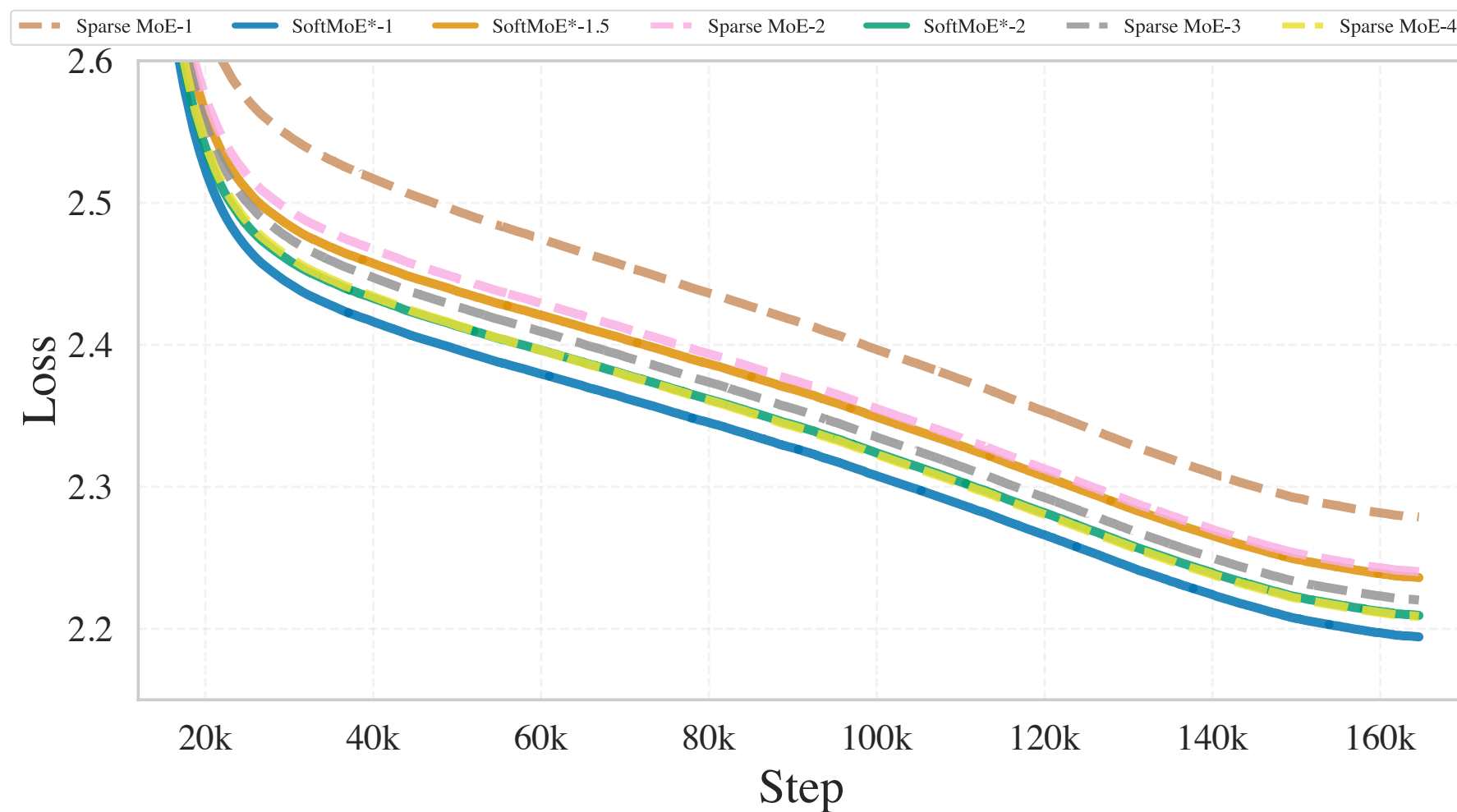
The model prefers later layers



Better quality with fewer active experts

Data	Model	k	α	Train-AE	Infer-AE	Active-Param	PIQA	HELLA	ARC-E	Loss	#Expert
C4	Sparse MoE	1	-	1	1	107.74	69.75	43.14	42.5	2.27	= 1
	Sparse MoE	2	-	2	2	156.94	71.98	45.41	40.74	2.24	≤ 2
	SoftMoE*	1.5	2	1.65	1.81	147.60	71.06	45.49	39.15	2.23	≤ 2
	Sparse MoE	3	-	3	3	206.14	71.60	46.36	42.50	2.22	≤ 3
	SoftMoE	1.5	2	1.96	2.19	166.29	71.38	46.79	43.39	2.22	≤ 3
	Sparse MoE	4	-	4	4	255.34	71.60	47.23	43.56	2.20	≤ 4
	SoftMoE*	1	4	3.60	3.82	246.49	72.91	48.48	43.21	2.19	≤ 4
	SoftMoE*	2	2	2.91	3.62	236.65	71.60	48.05	42.86	2.21	≤ 4
	SoftMoE	2	2	3.35	3.76	243.54	72.03	47.48	42.15	2.20	≤ 4

Better quality with fewer active experts



Key takeaways

- We propose SoftMoE, a routing layer with relaxed Top-K selection based on LapSum operator.
- SoftMoE adapts compute to input tokens and redistribute active expert budgets between layer.
- SoftMoE matches hard top-k in language modelling and downstream tasks while activating significantly fewer experts.



Github