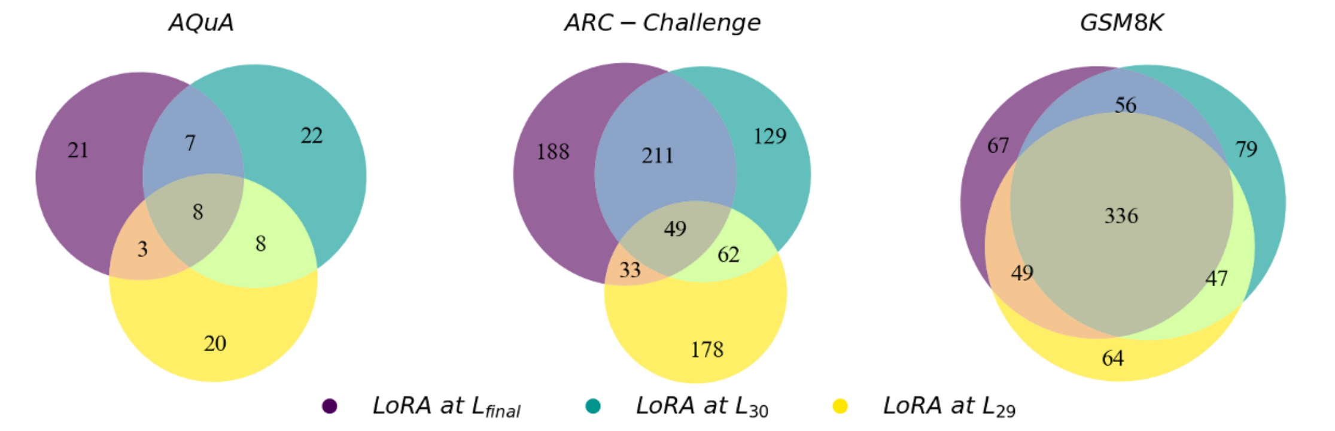
IMPERIAL

Tuning Language Models by Mixture-of-Depths Ensemble

Haoyan Luo Lucia Specia
h.luo23@imperial.ac.uk l.specia@imperial.ac.uk
Department of Computing, Imperial College London

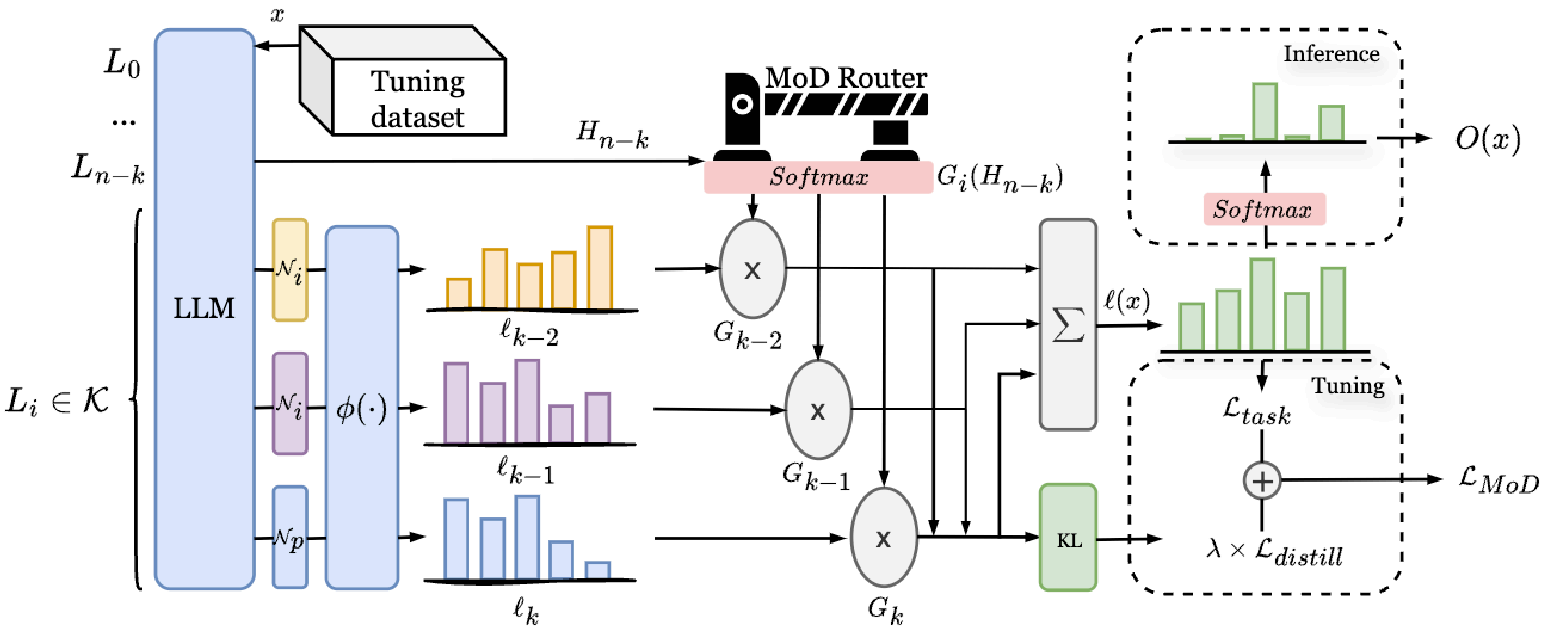
Introduction

Traditional LLMs focus on final-layer loss and representations, overlooking predictive power in intermediate layers. We propose Mixture-of-Depths (MoD) framework, which leverages late layers as ensembles with learned additional routing weights. We adapt late layer by adding distillation loss and normalization modules. MoD integrates with existing tuning methods, improving performance on reasoning tasks with minimal parameter increase. Remarkably, MoD can achieves similar performance with 97% fewer trainable parameters.



Motivation: Intersection of solved problems by tuning loss layers on the AQuA, ARC-Challenge, and GSM8K datasets, showing complementary test-time performance when tuning loss for late layers.

Mixture-of-Depths (MoD)



Applying trained language model head to late layers for (early exit):

$$\ell(x_t \mid x_{<t}) = \phi(\mathcal{N}_p(h_t^{(N)}))_{x_t}, \quad x_t \in \mathcal{V}.$$

Combining the ensemble logits using routing linear network:

$$\ell(x_t \mid x_{<t}) = \sum_{i=0}^{k-1} G(x)_i \cdot \ell_i(x).$$

Adapting late layers with distillation loss:

$$\mathcal{L}_{distill} = \sum_{i=0}^{k-2} \text{KL}(P_i \parallel P_n),$$

and $\mathcal{L}_{tuned} = \mathcal{L}_{task} + \lambda\mathcal{L}_{distill}$. Normalization modules $\mathcal{N}_k$ are trained individually for each ensemble layer.

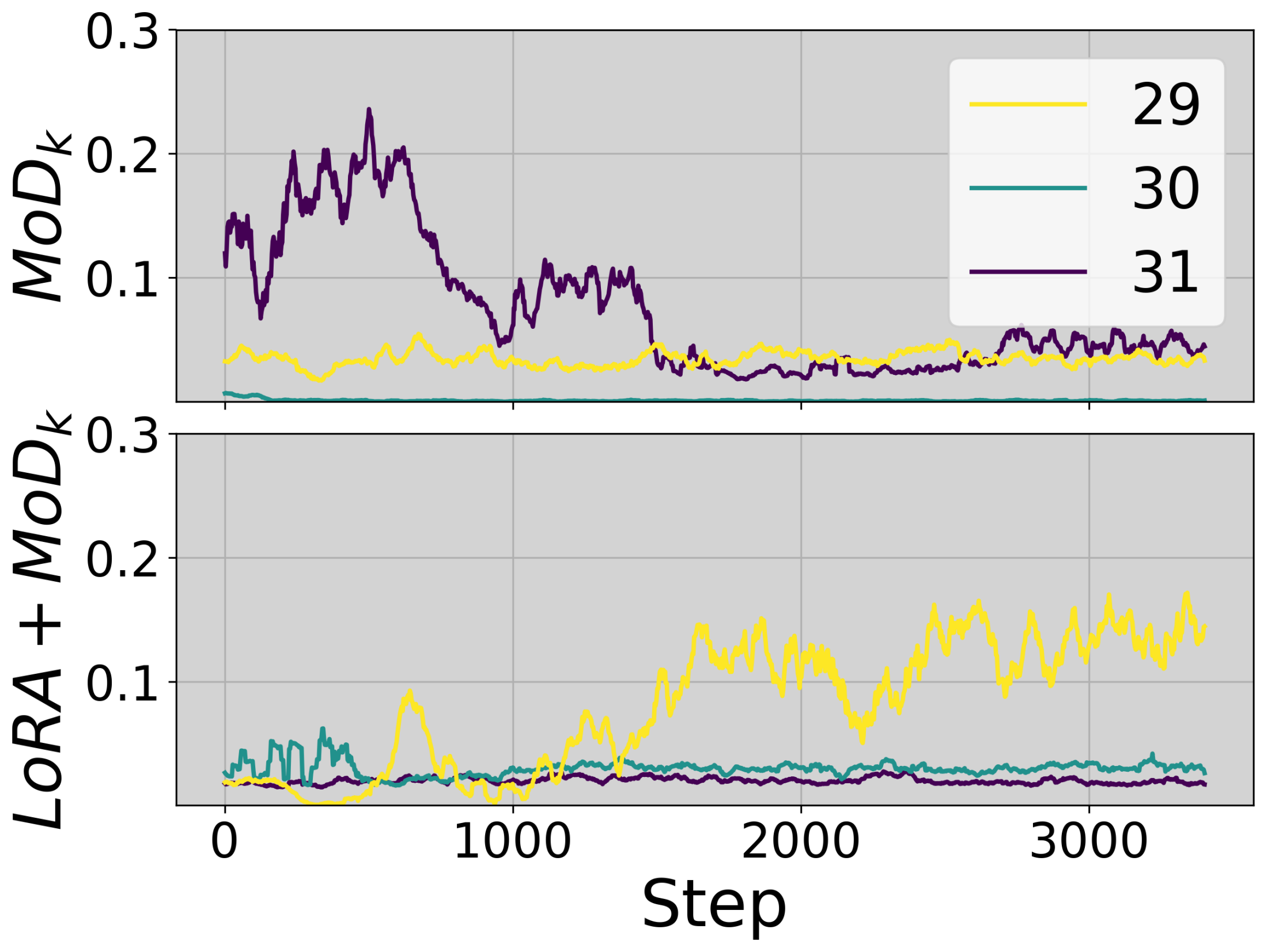
Experiments

We use LoRA for base tuning with LLaMA-1 and LLaMA-2 models (7B parameters), setting k to 3. Baselines are defined as LoRA$_{\neg\mathcal{K}}$ (excluding last k layers) and LoRA$_{\text{all}}$ (all layers). MoD achieves the best performance on eleven reasoning datasets with minimal increase in trainable parameters.

When combined with k LoRA ensemble layers, MoD consistently improves performance with minimal additional parameters (0.04%). Furthermore, MoD alone, with 97% less parameters, provides competitive performance with LoRA$_{\text{all}}$, demonstrating the potential of leveraging late layer predictive powers.

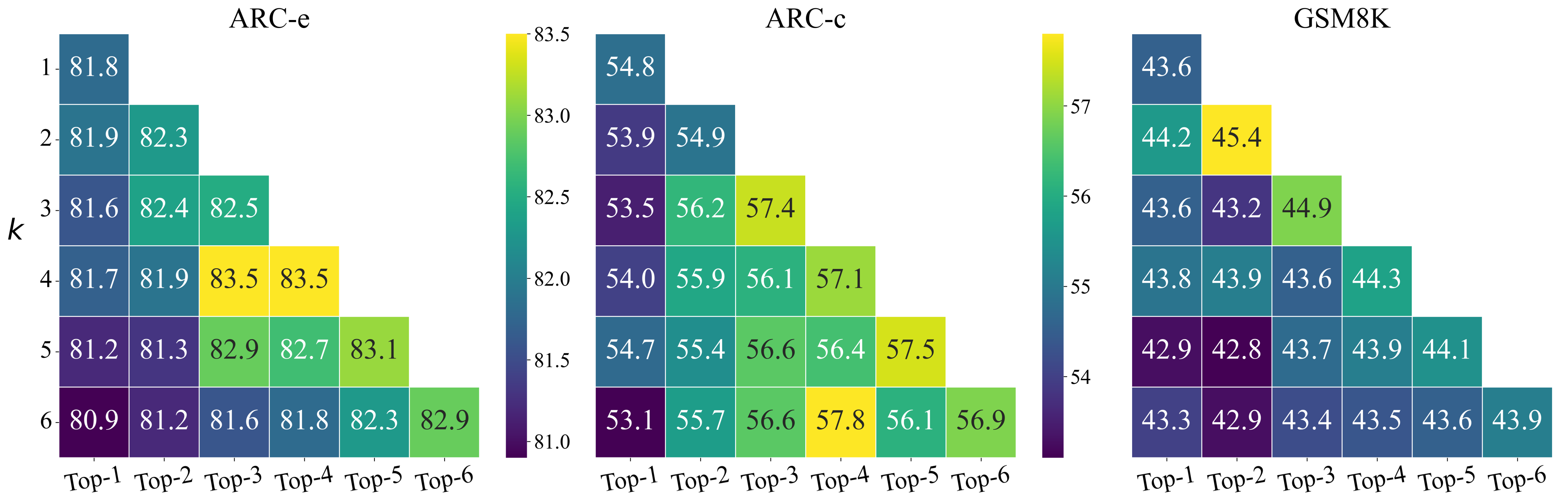
Method	AddSub	AQuA	GSM8K	MAWPS	MultiArith	SingleEq	SWAMP	Avg.
LLaMA-7B								
LoRA$_{\neg\mathcal{K}}$	38.7	13.4	37.3	56.3	78.2	59.8	42.3	46.6
+ $L_{\text{LoRA}} \times \lvert\mathcal{K}\rvert$ (+10.3%)	41.3	15.4	38.5	58.0	81.0	**62.9**	44.2	48.8
+ $L_{\text{LoRA}} \times \lvert\mathcal{K}\rvert$ + **MoD$_{\mathcal{K}}$** (+10.4%)	**42.0**	15.8	**39.1**	**58.5**	**81.3**	**62.9**	**44.9**	**49.2**
+ **MoD$_{\mathcal{K}}$** (+0.04%)	41.5	**16.1**	38.2	58.4	80.7	62.3	43.8	48.7
LLaMA2-7B								
LoRA$_{\neg\mathcal{K}}$	46.3	20.5	39.7	60.6	81.4	62.0	43.2	50.5
+ $L_{\text{LoRA}} \times \lvert\mathcal{K}\rvert$ (+10.3%)	51.1	24.4	43.6	62.6	84.2	66.9	47.7	54.5
+ $L_{\text{LoRA}} \times \lvert\mathcal{K}\rvert$ + **MoD$_{\mathcal{K}}$** (+10.4%)	**51.2**	**25.5**	**43.9**	63.1	**84.3**	**67.3**	**48.0**	**54.8**
+ **MoD$_{\mathcal{K}}$** (+0.04%)	50.1	24.3	43.4	**63.7**	82.2	66.8	47.5	54.0

Sparsity of Routing Network



We measured sparsity of routing weights across training tokens , setting sparsity threshold $\epsilon = 1 \times 10^{-5}$. MoD with and without k LoRA layers showed distinct sparsity patterns. The model can learn pattern of favoring which ensemble route during tuning.

MoD Sparse Routing



For efficiency, we investigate sparse routing with Top-K activation:

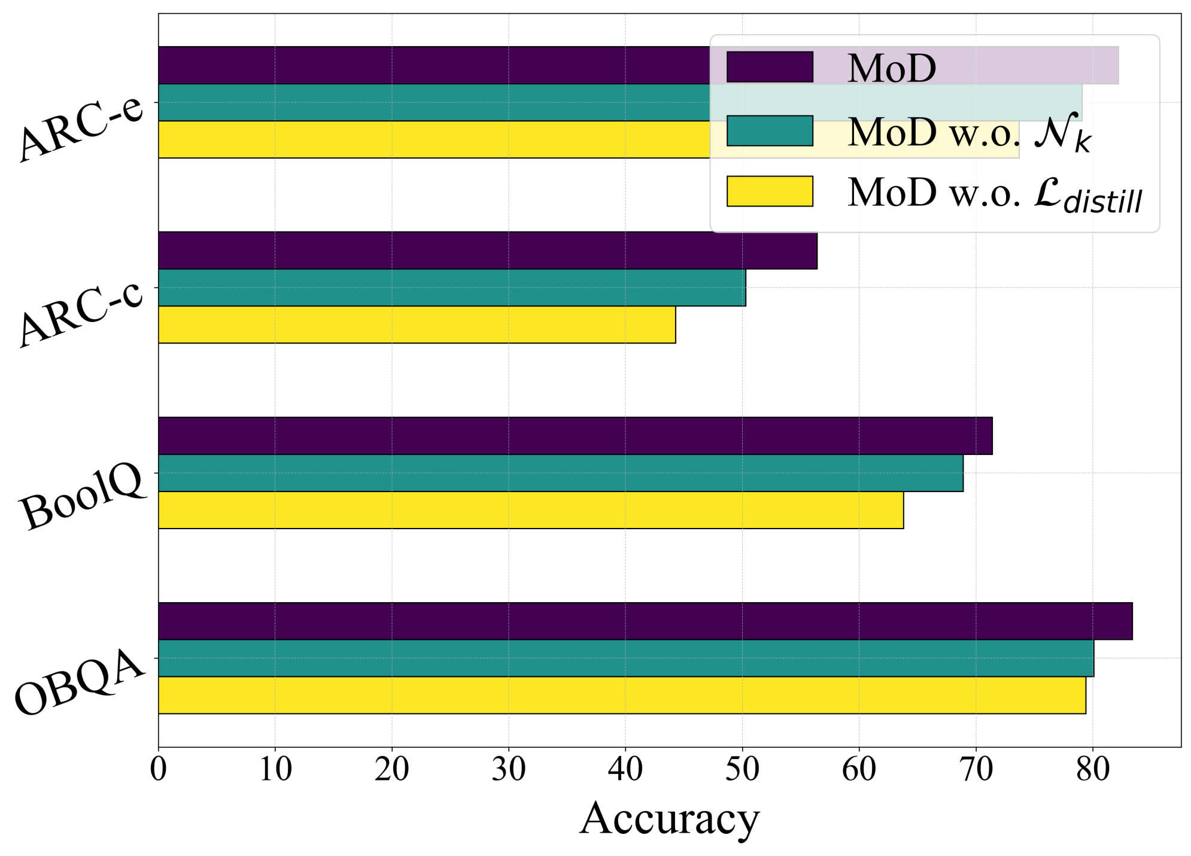
$$G_{\text{TopK}}(x) := \text{Softmax}(\text{TopK}(x \cdot W_g)).$$

Optimal k value ranges from 3 to 4. Top-K activation slightly reduces performance but still outperforms the baseline when $k = 1$.

Dataset	Top-2	Top-3	Top-4	Top-5
ARC-e	1.4×	1.5×	1.4×	1.2×
ARC-c	1.6×	1.4×	1.3×	1.1×

Larger Top-K increase acceleration ratios, suggesting a trade-off between utilizing the predictive power and efficiency.

Ablation Study



Both $\mathcal{L}_{distill}$ and $\mathcal{N}_k$ enhance tuning performance. Strong task signals from the last layer are important for effective adaptation, while routing network also demonstrates strong performance independently.