



香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

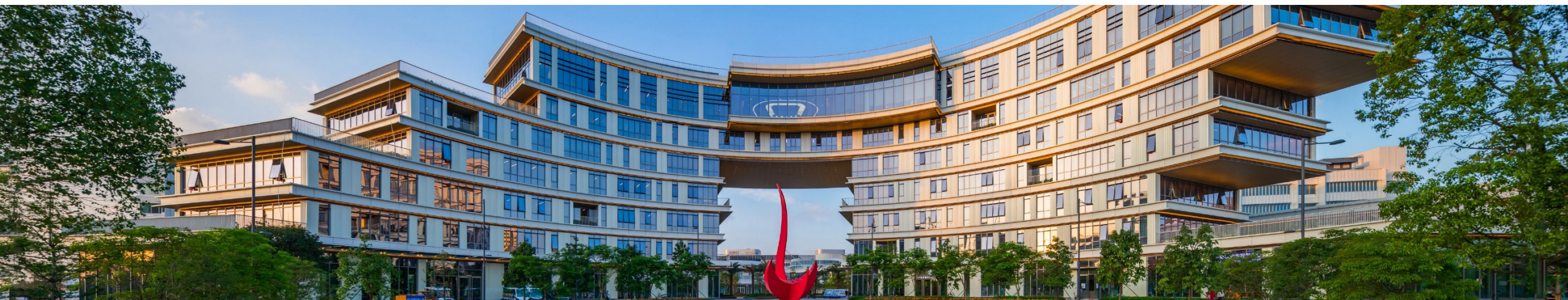


ICML
International Conference
On Machine Learning

Mining Tensor/Neuron-Level Sparsity to Maximize Mixture-of-Experts Potential in Post-Training and Inference

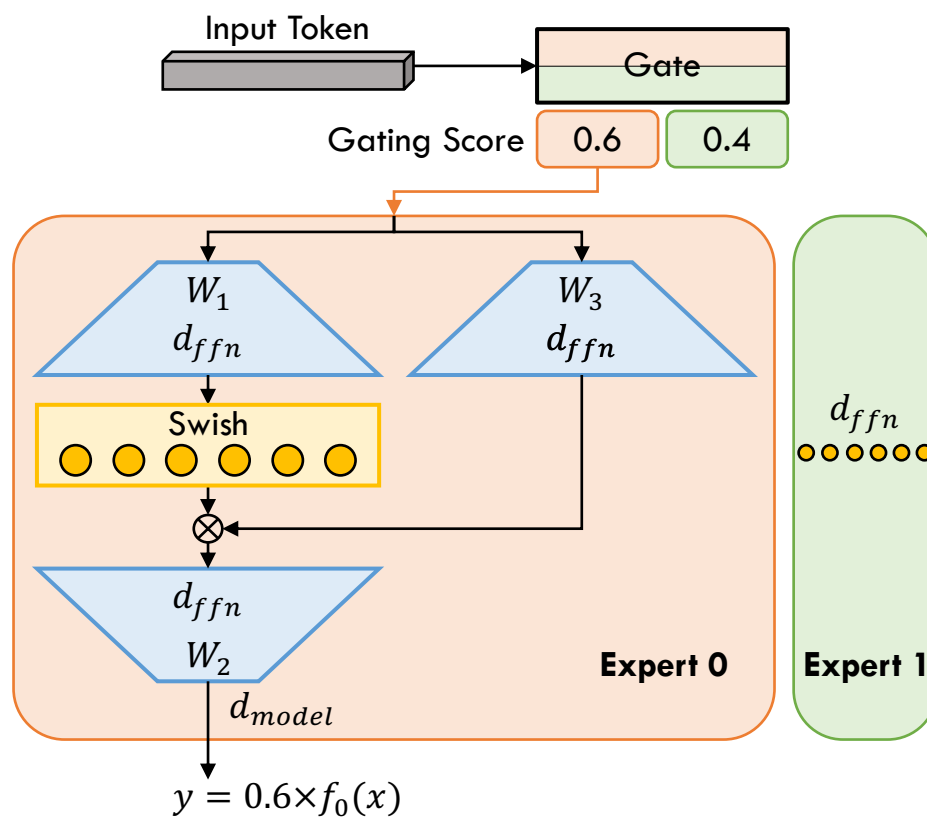
Weilin Cai¹, Le Qin¹, Shwai He², Junwei Cui¹, Ang Li², Jiayi Huang¹⁺

¹The Hong Kong University of Science and Technology (Guangzhou) | ²University of Maryland



Background

Mixture-of-experts (MoE) fixes the **expert granularity** and the **sparsity ratio** at the beginning of pre-training.

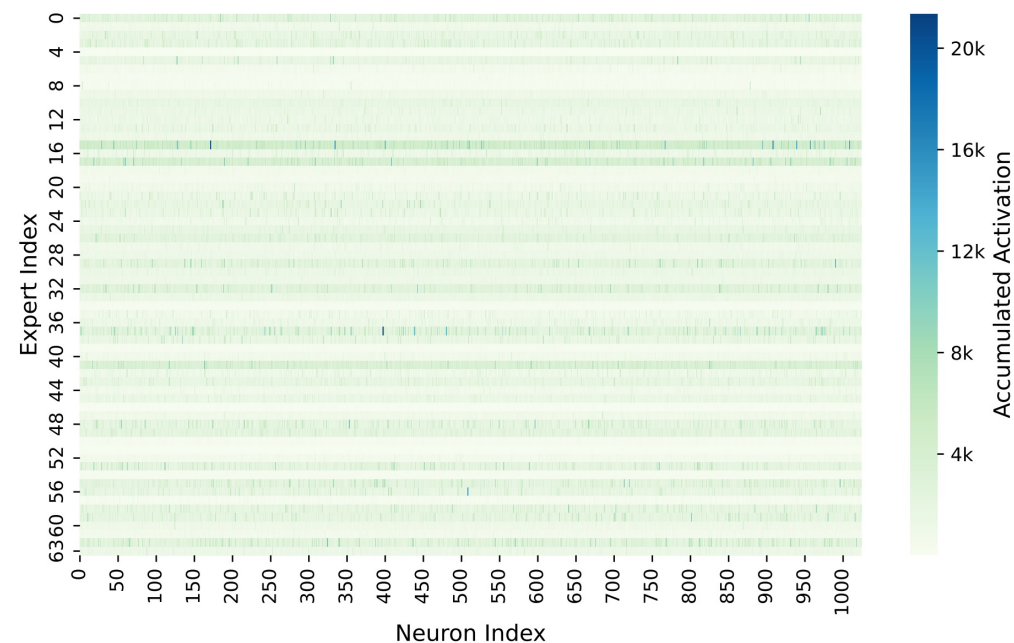


Key Observation — Dual Sparsity

MoE has two types of hidden sparsity:

1. Tensor-level sparsity

- Some activated token-expert pairs have very low gating scores
- These experts may contribute very little



Key Observation — Dual Sparsity

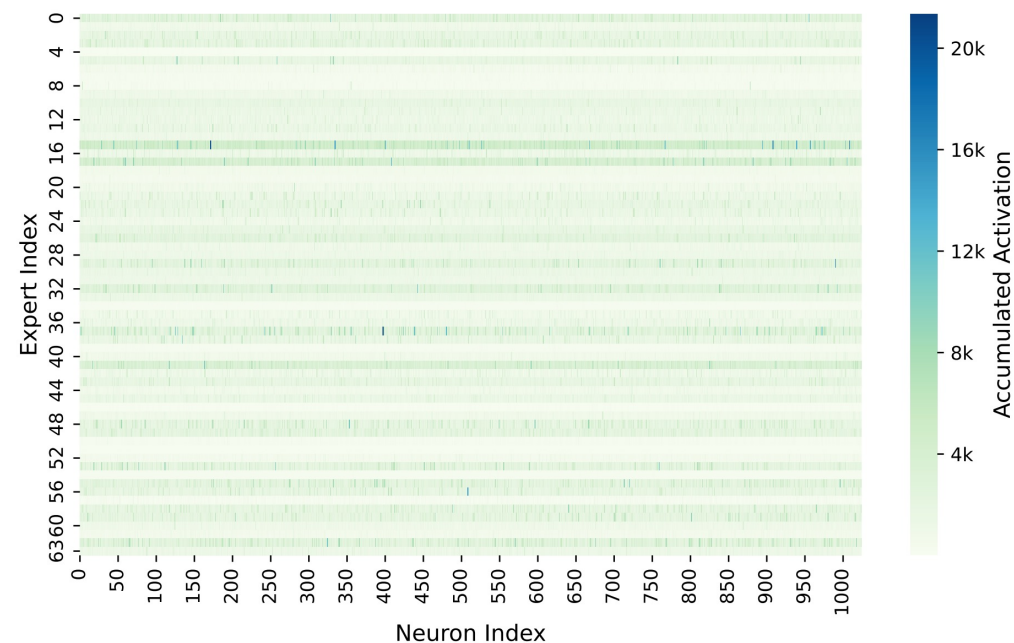
MoE has two types of hidden sparsity:

1. Tensor-level sparsity

- Some activated token-expert pairs have very low gating scores
- These experts may contribute very little

2. Neuron-level sparsity

- Inside each expert, neurons are not equally important
- Some neurons matter much more than others



Key Observation — Dual Sparsity

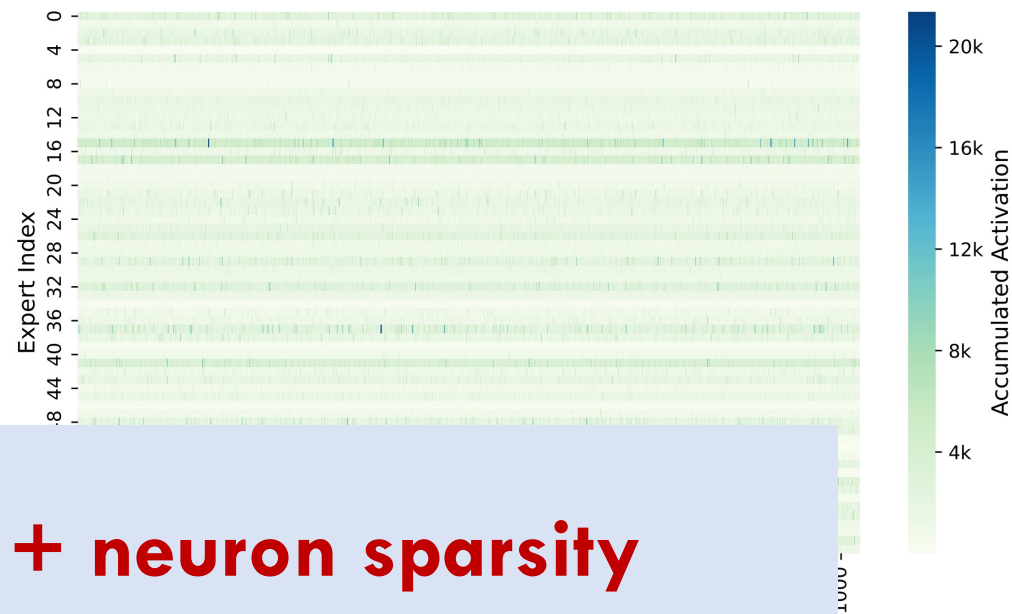
MoE has two types of hidden sparsity:

1. Tensor-level sparsity

- Some activated token-expert pairs have very low gating scores
- These experts may contribute very little

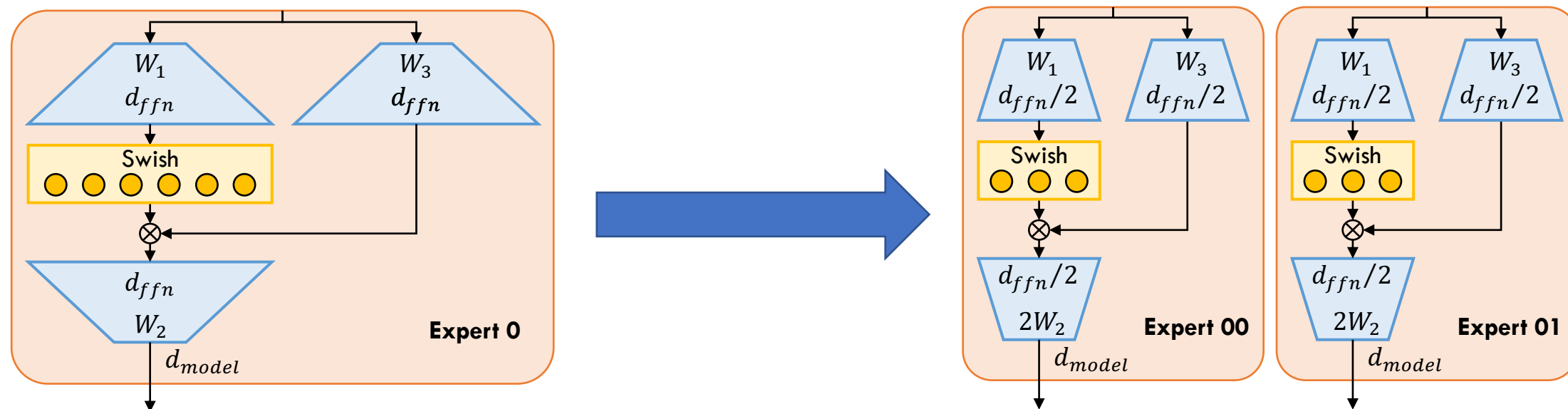
2. Neuron-level sparsity

- Inside each expert, neurons are not equally important
- Some neurons matter much more than others



Dual sparsity = tensor sparsity + neuron sparsity

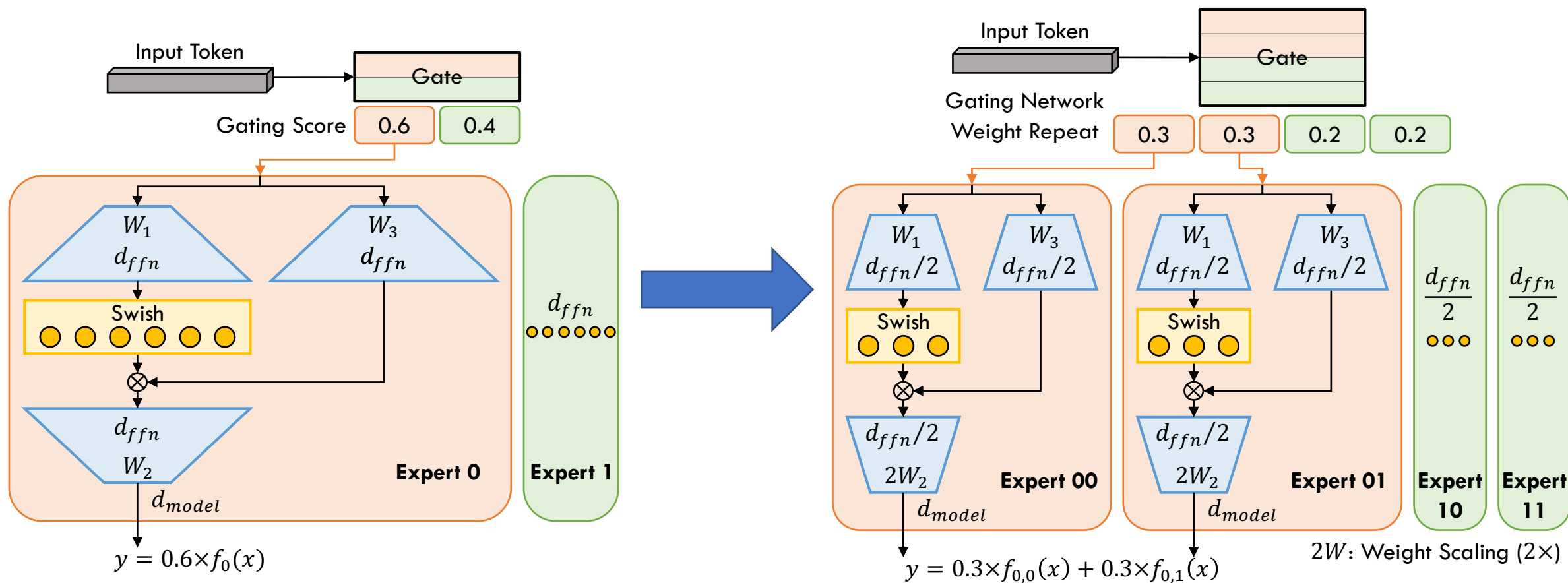
Complete Expert Partition for Post-training



Complete Expert Partition for Post-training



Complete Expert Partition for Post-training



Post-training Accuracy

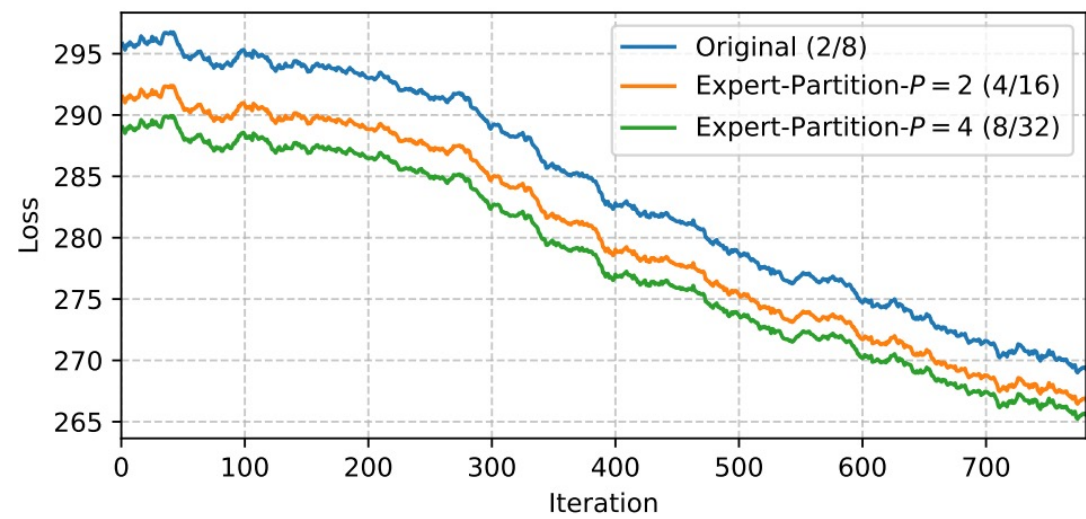
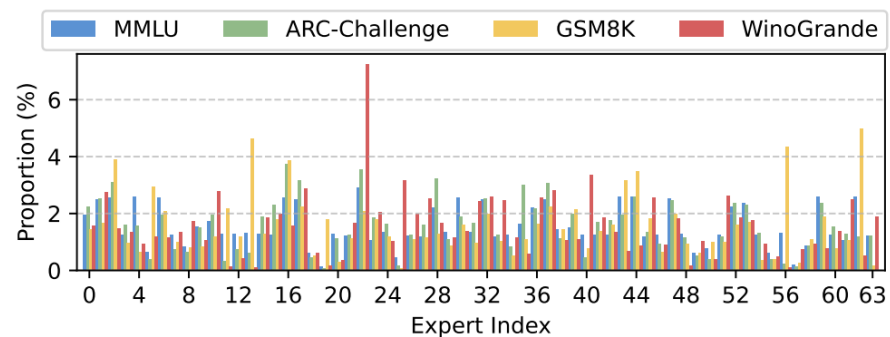


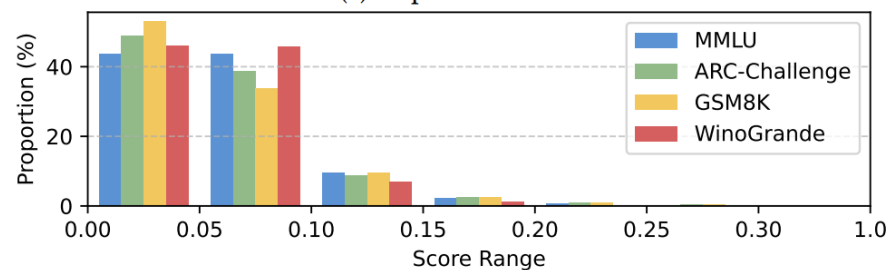
Table 1. Comparison of downstream accuracy between the original Mixtral-8×7B model and its expert-partitioned variant.

Model	E -Activ./Total	T^1_{drop}	Drop Rate	ARC-C	BoolQ	GSM8K	HellaSwag	MMLU	OBQA	PIQA	RTE	WinoGrande	AVG.(↑)
Mixtral-8×7B	2/8	-	0	59.47	85.14	58.07	84.05	67.13	47.00	83.79	70.40	76.56	<u>70.18</u>
	4/16 ($P = 2$)	-	0	59.56	85.32	58.30	84.02	67.05	47.20	83.41	70.76	76.01	<u>70.18</u>
	8/32 ($P = 4$)	-	0	59.47	85.26	58.07	83.99	67.22	46.80	83.46	70.76	76.72	<u>70.19</u>
Fine-Tuned Mixtral-8×7B	2/8	-	0	60.58	87.06	60.73	82.99	64.92	46.20	83.62	71.84	76.87	<u>70.53</u>
	4/16 ($P = 2$)	-	0	59.56	87.06	62.85	82.96	65.65	47.00	83.3	72.92	76.48	<u>70.86</u>
	8/32 ($P = 4$)	-	0	60.67	87.55	62.85	83.06	65.10	47.60	83.46	72.92	76.87	<u>71.12</u>
Fine-Tuned Mixtral-8×7B Threshold Drop	2/8	0.30	20.3%	59.39	87.06	61.84	82.72	64.26	46.40	82.86	71.48	76.64	<u>70.29</u>
	4/16 ($P = 2$)	0.15	21.0%	59.64	87.00	63.46	82.58	64.75	46.40	83.13	73.65	76.24	<u>70.76</u>
	8/32 ($P = 4$)	0.08	23.9%	59.73	87.31	62.85	82.75	64.76	47.00	83.03	74.01	76.48	<u>70.88</u>

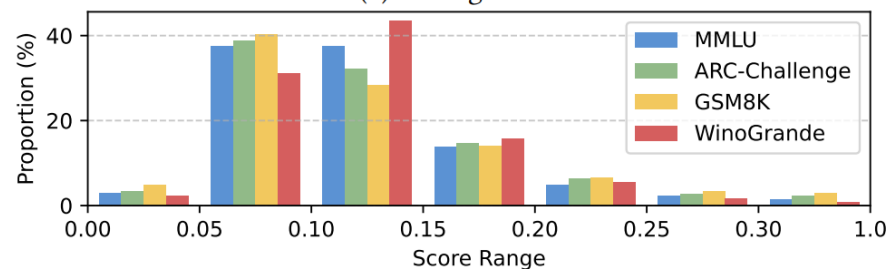
Threshold-based Token-Expert Dropping (1 T-Drop)



(a) Expert Selection

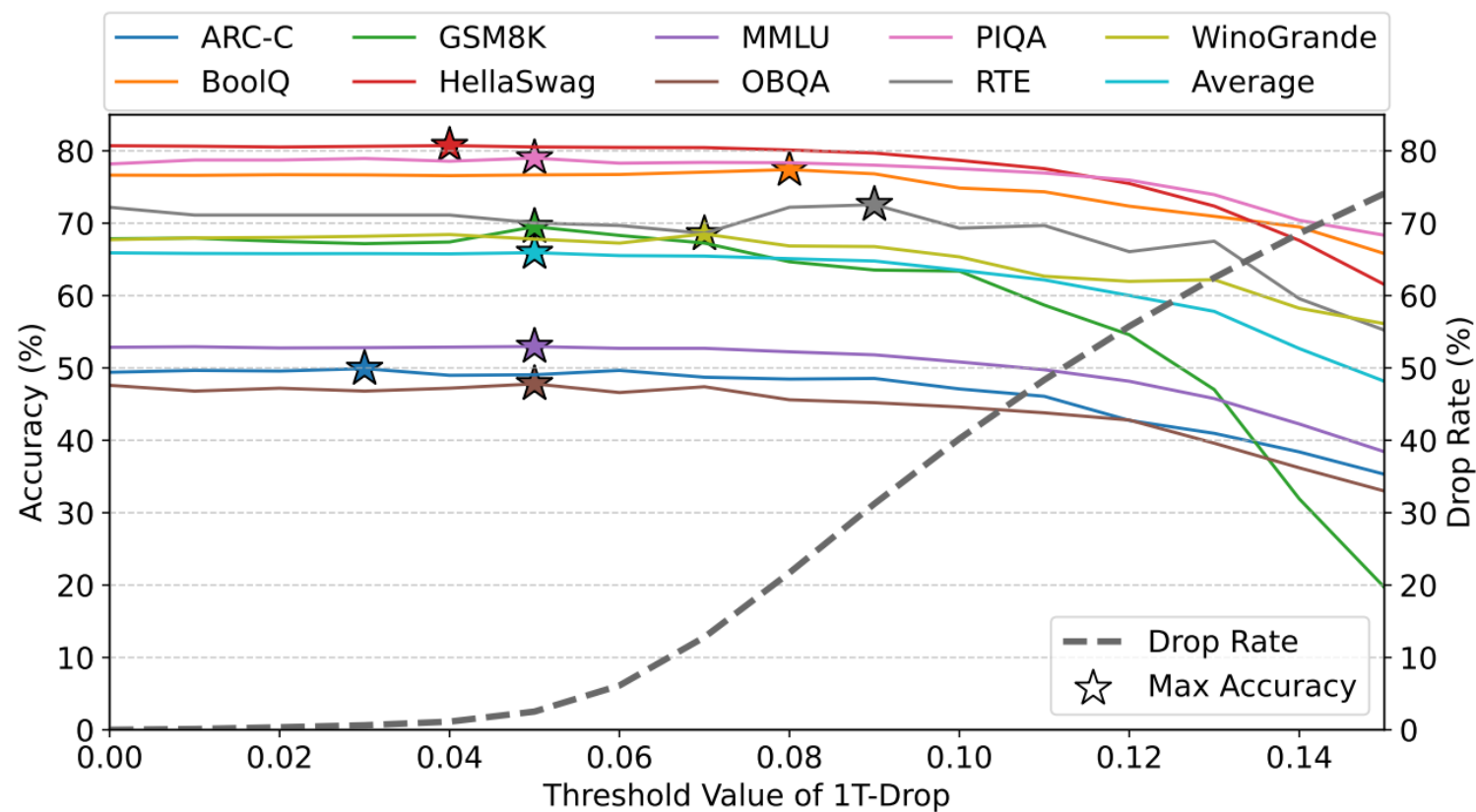


(b) Gating Score

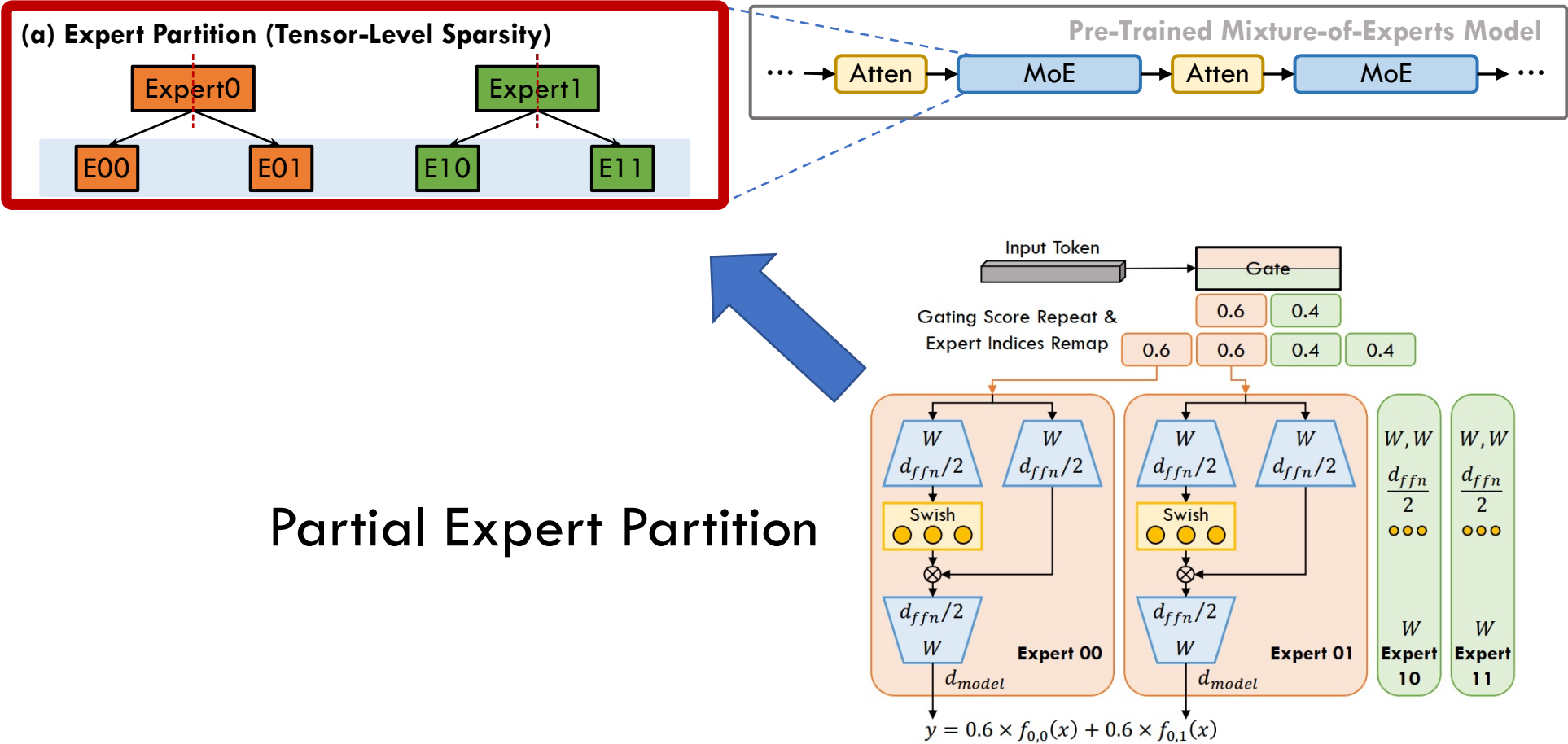


(c) Normalized Gating Score

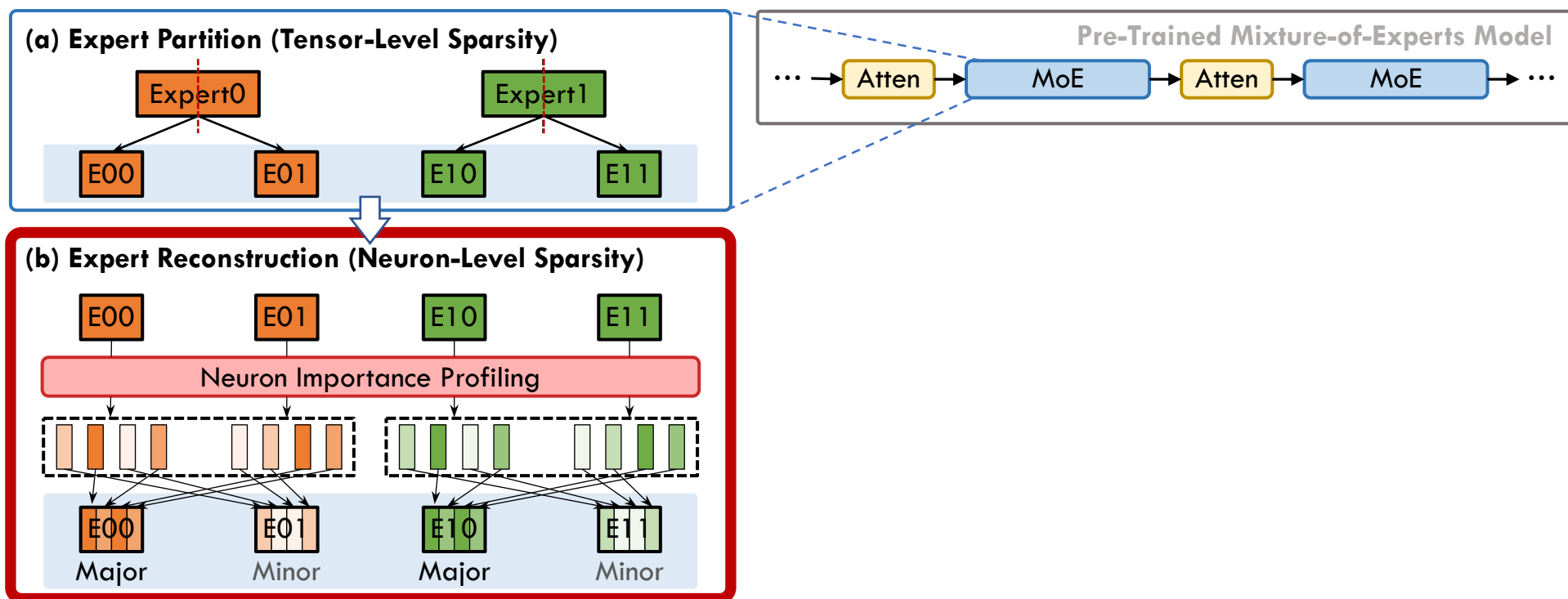
Threshold-based Token-Expert Dropping (1T-Drop)



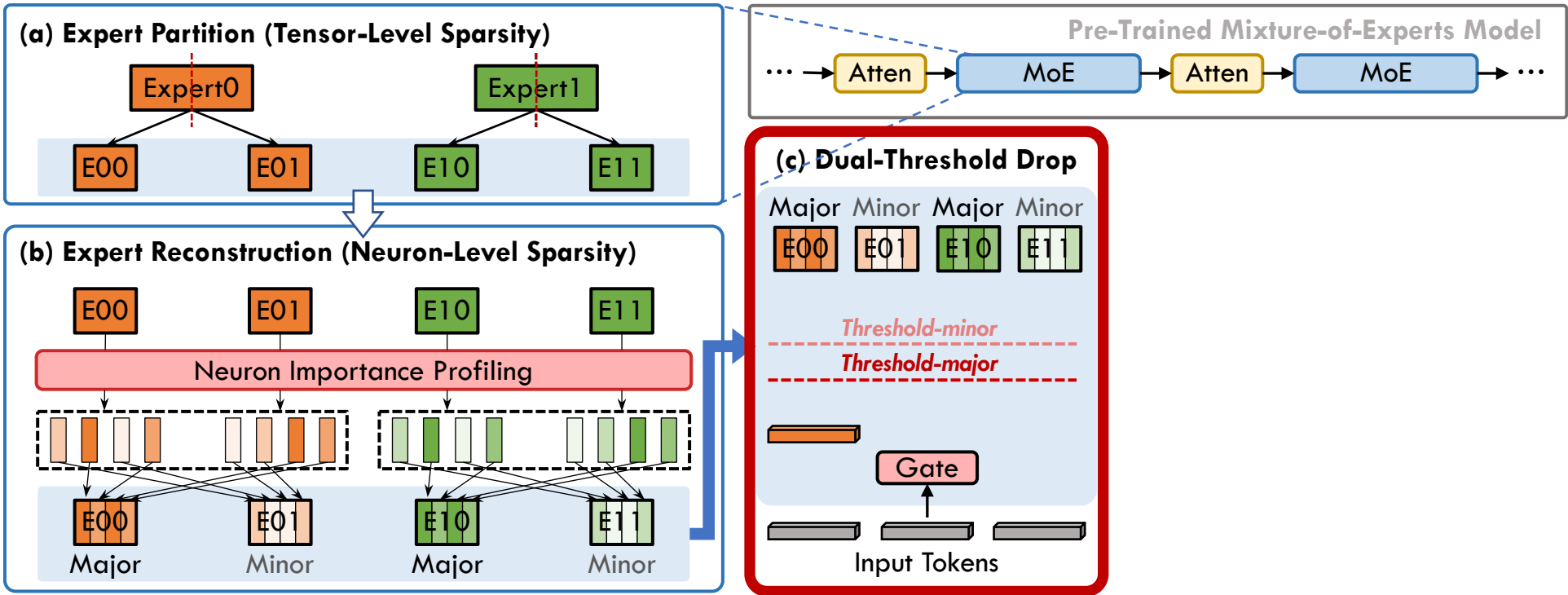
Dual-Threshold Token-Expert Dropping (2T-Drop)



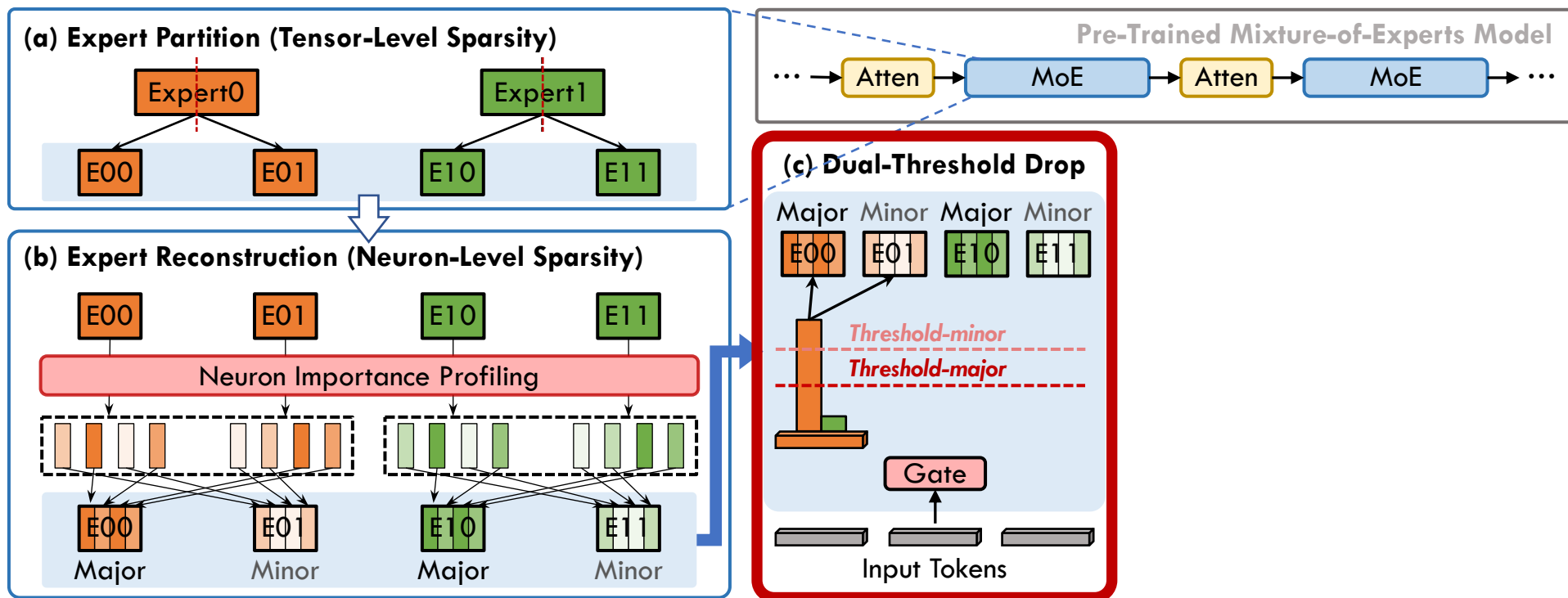
Dual-Threshold Token-Expert Dropping (2T-Drop)



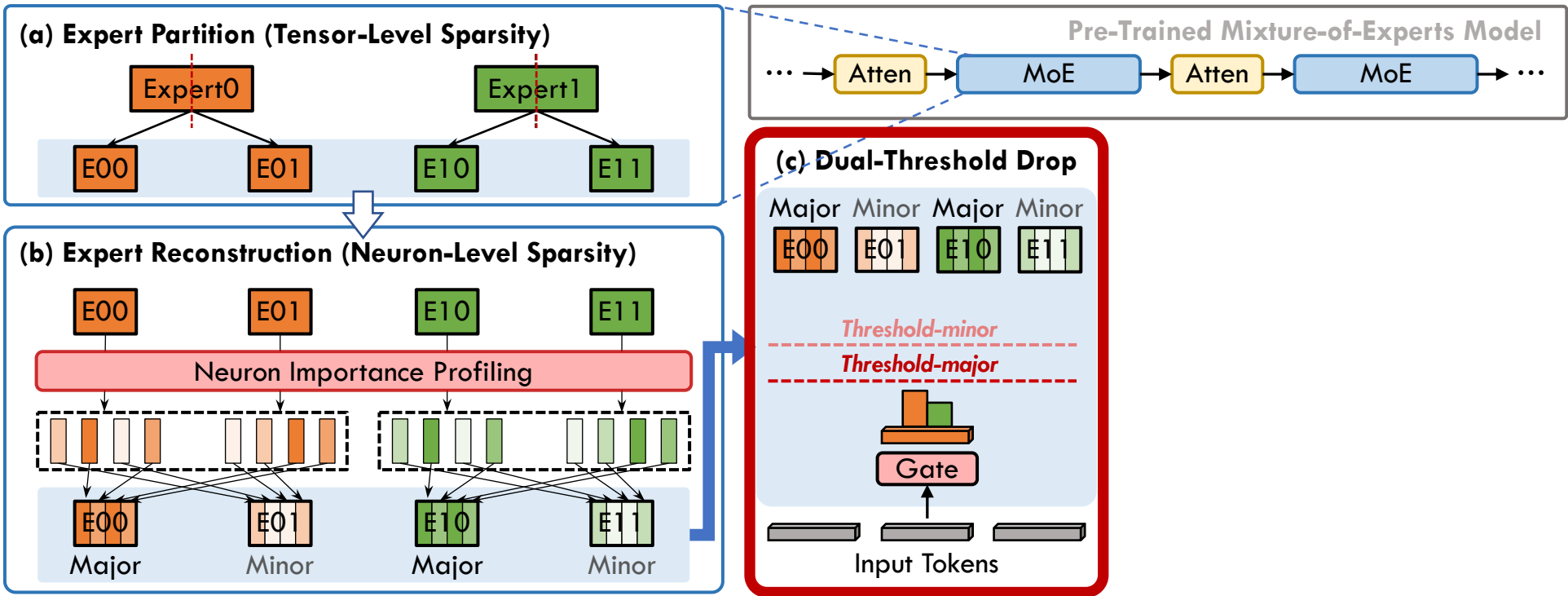
Dual-Threshold Token-Expert Dropping (2T-Drop)



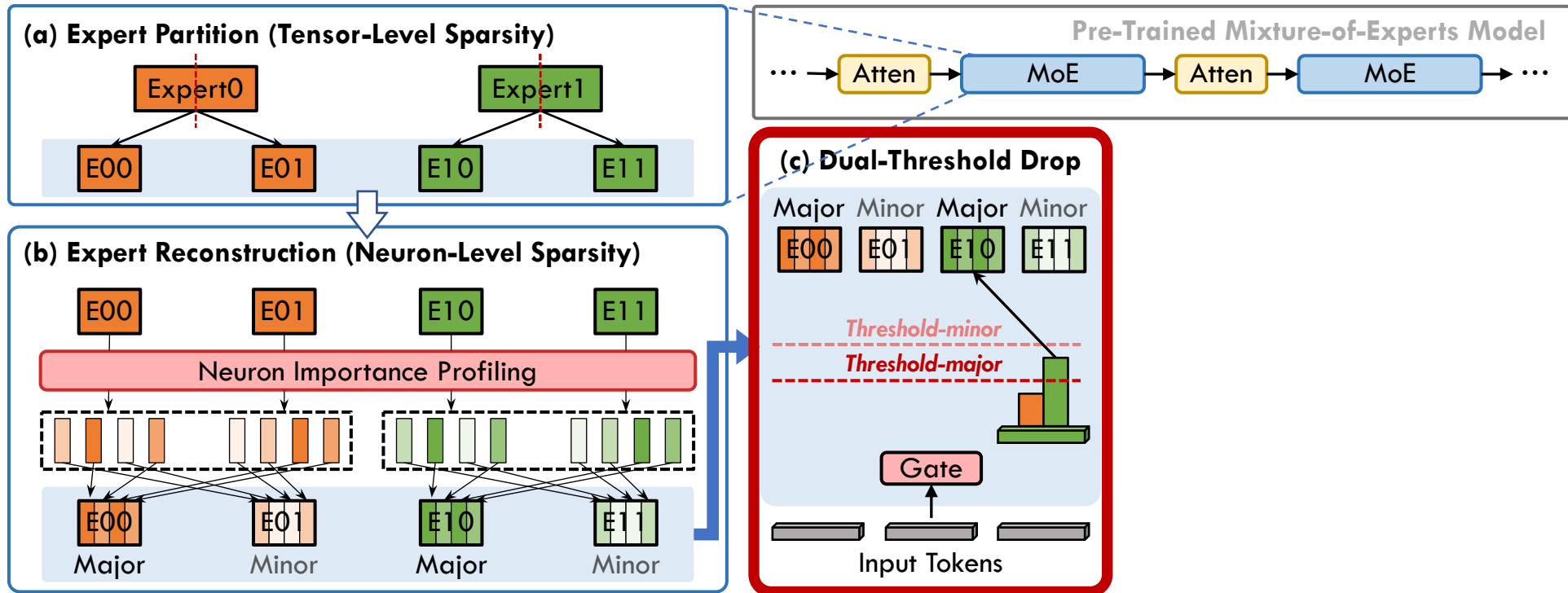
Dual-Threshold Token-Expert Dropping (2T-Drop)



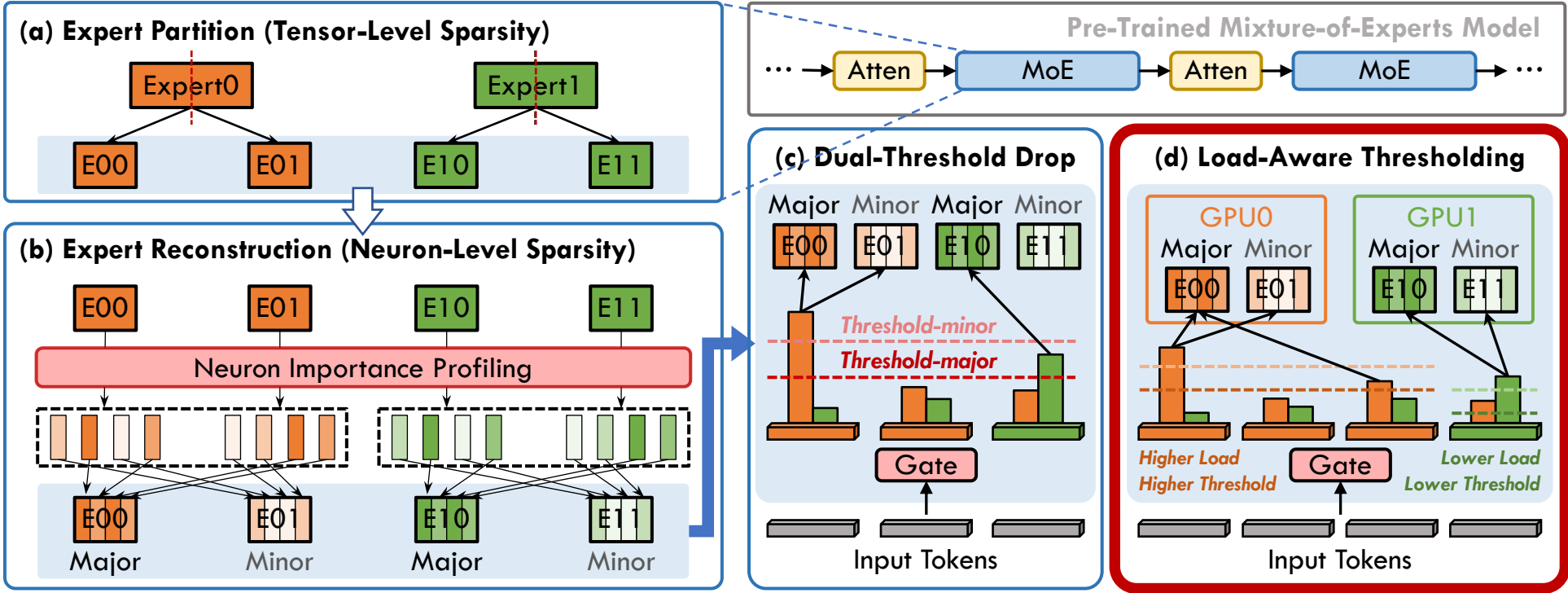
Dual-Threshold Token-Expert Dropping (2T-Drop)



Dual-Threshold Token-Expert Dropping (2T-Drop)

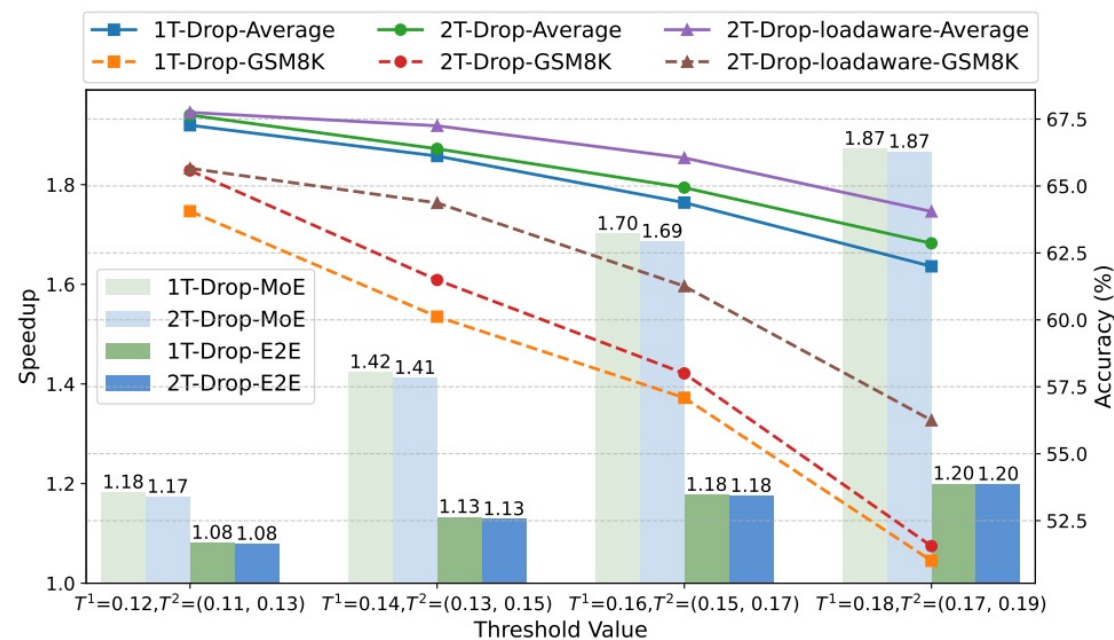
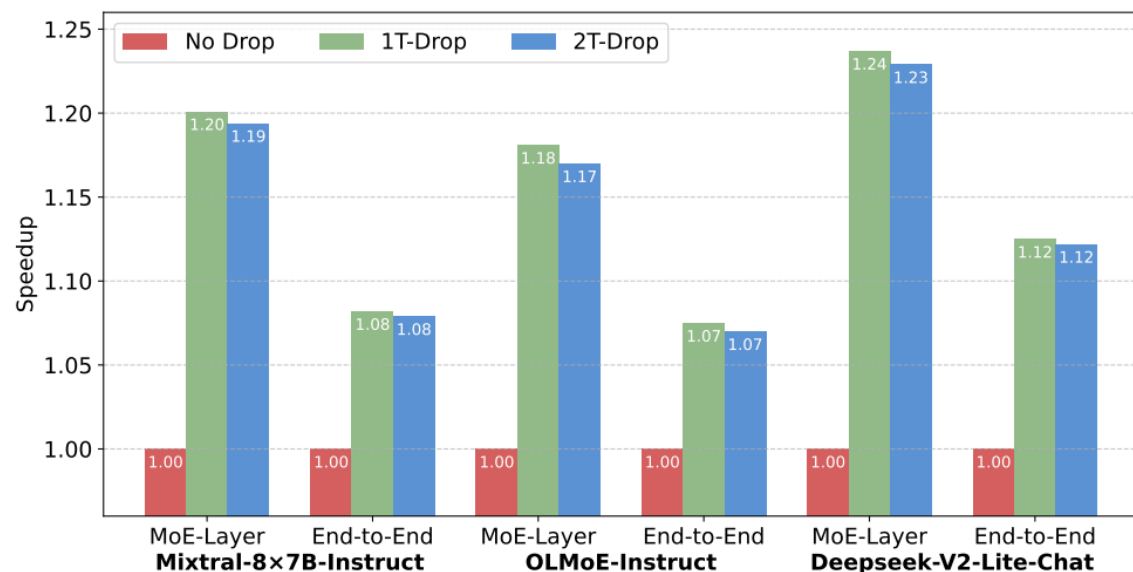


Dual-Threshold Token-Expert Dropping (2T-Drop)



Inference Evaluation

Model	Drop Method	T_{major}^2	T_{minor}^2	Drop Rate	ARC-C	BoolQ	GSM8K	HellaSwag	MMLU	OBQA	PIQA	RTE	WinoGrande	AVG.(↑)
OLMoE-Instruct	No Drop	-	-	0	49.40	76.64	67.85	80.70	52.86	47.60	78.18	72.20	67.72	<u>65.91</u>
	1T-Drop	0.08	0.08	21.7%	48.46	77.40	64.67	80.11	52.23	45.60	78.35	72.20	66.85	<u>65.10</u>
	2T (Reconstruct)	0.07	0.09	22.0%	50.00	77.00	67.22	80.13	52.45	47.80	79.38	71.12	65.59	<u>65.63</u>
DeepSeek-V2-Lite-Chat	No Drop	-	-	0	53.92	82.91	65.05	80.81	56.91	45.20	81.12	72.56	71.98	<u>67.83</u>
	1T-Drop	0.12	0.12	27.0%	51.79	82.91	63.61	80.30	55.18	44.40	81.61	74.37	71.27	<u>67.27</u>
	2T (Reconstruct)	0.11	0.13	26.9%	52.47	82.94	64.37	80.37	55.58	44.80	81.39	74.73	72.22	<u>67.65</u>



Thank You

Email: wcai738@connect.hkust-gz.edu.cn



香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)