中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Domain-Specific Pruning of Large Mixture-of-Experts Models with Few-shot Demonstrations

Zican Dong[1,2]*, Han Peng[1,2]*, Peiyu Liu[3]†, Wayne Xin Zhao[1,2]†,
Dong Wu[5], Feng Xiao[4], Zhifeng Wang[4]

[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3] University of International Business and Economics
[4] YanTron Technology Co. Ltd [5] EBTech Co. Ltd

1. Background

- Mixture-of-Experts (MoE) architectures have demostrated efficiency of scaling parameters without proportional computational overhead.
- The deployment of large MoE models imposes substantial memory requirements.
  - DeepSeek-R1 (671B)
  - BF16: 1500GB -> 4 $\times$ 8 A800/H800
  - FP8: 750GB -> 2 $\times$ 8 H800
- Necessary of MoE compression techniques.

1. Background

MoE Architecture

$$\bar{\boldsymbol{h}}_t^l = \sum_{i=1}^{N} g_{i,t}^l \cdot \mathrm{E}_i^l(\boldsymbol{h}_t^l), \quad \tilde{\boldsymbol{h}}_t^l = \boldsymbol{h}_t^l + \bar{\boldsymbol{h}}_t^l$$

Expert Metrics

Frequency

$$f_i^l = \sum_{n=1}^{M} \sum_{t=1}^{T_n} (g_{i,n,t}^l > 0)$$

Gating Score

$$r_i^l = \sum_{n=1}^{M} \sum_{t=1}^{T_n} g_{i,n,t}^l$$

2. Empirical Analysis

高瓴人工智能学院
Gaoling School of Artificial Intelligence

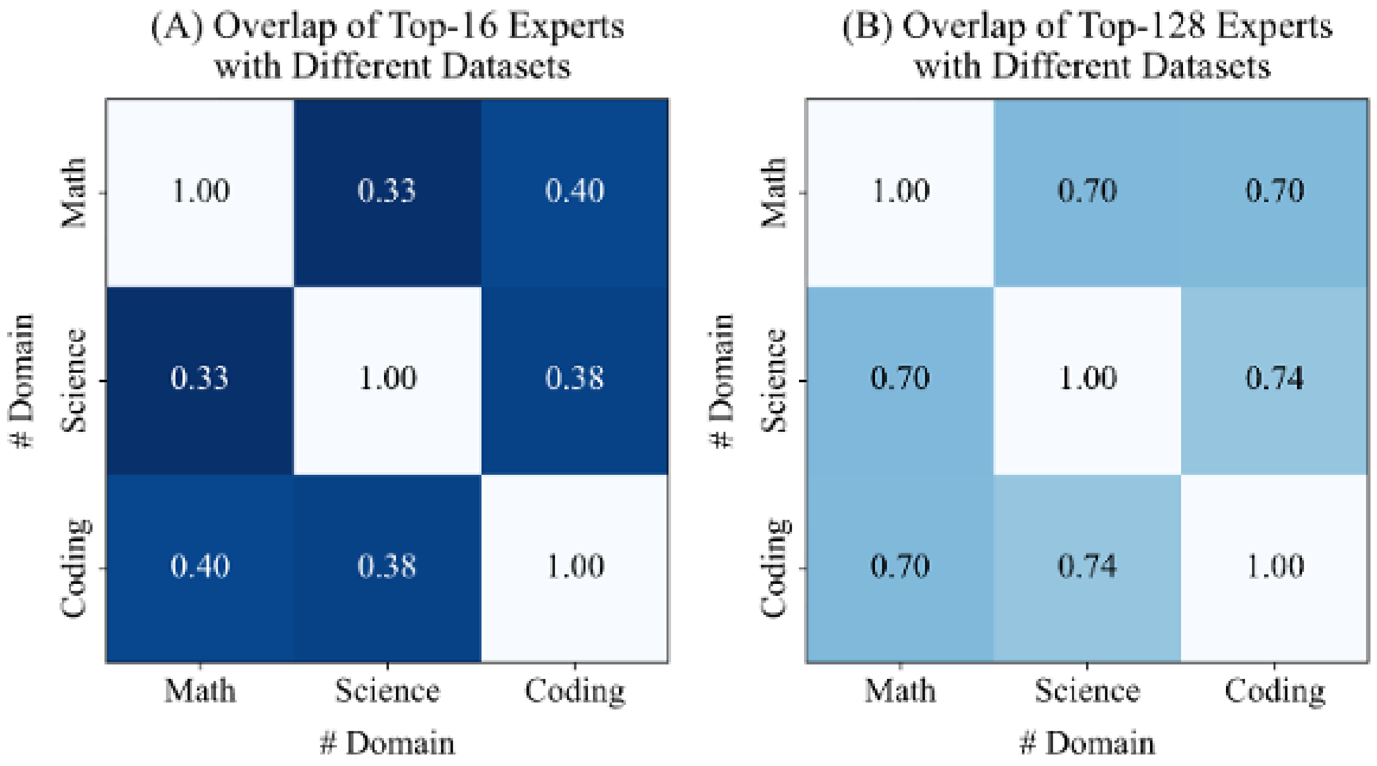
Expert Specialization Across Domains

Domain	AIME24	GPQA	LiveCodeBench
Full	77.08	70.91	63.32
Math	67.33 **(-9.75)**	69.19 (-1.72)	65.27 (+1.95)
Code	78.67 (+1.59)	71.72 (+0.81)	**55.68 (-6.64)**
Science	79.33 (+2.25)	**59.09 (-11.82)**	61.07 (-2.25)

Large MoE models contain domain-specialized experts that are predominantly activated in their respective domains.

2. Empirical Analysis

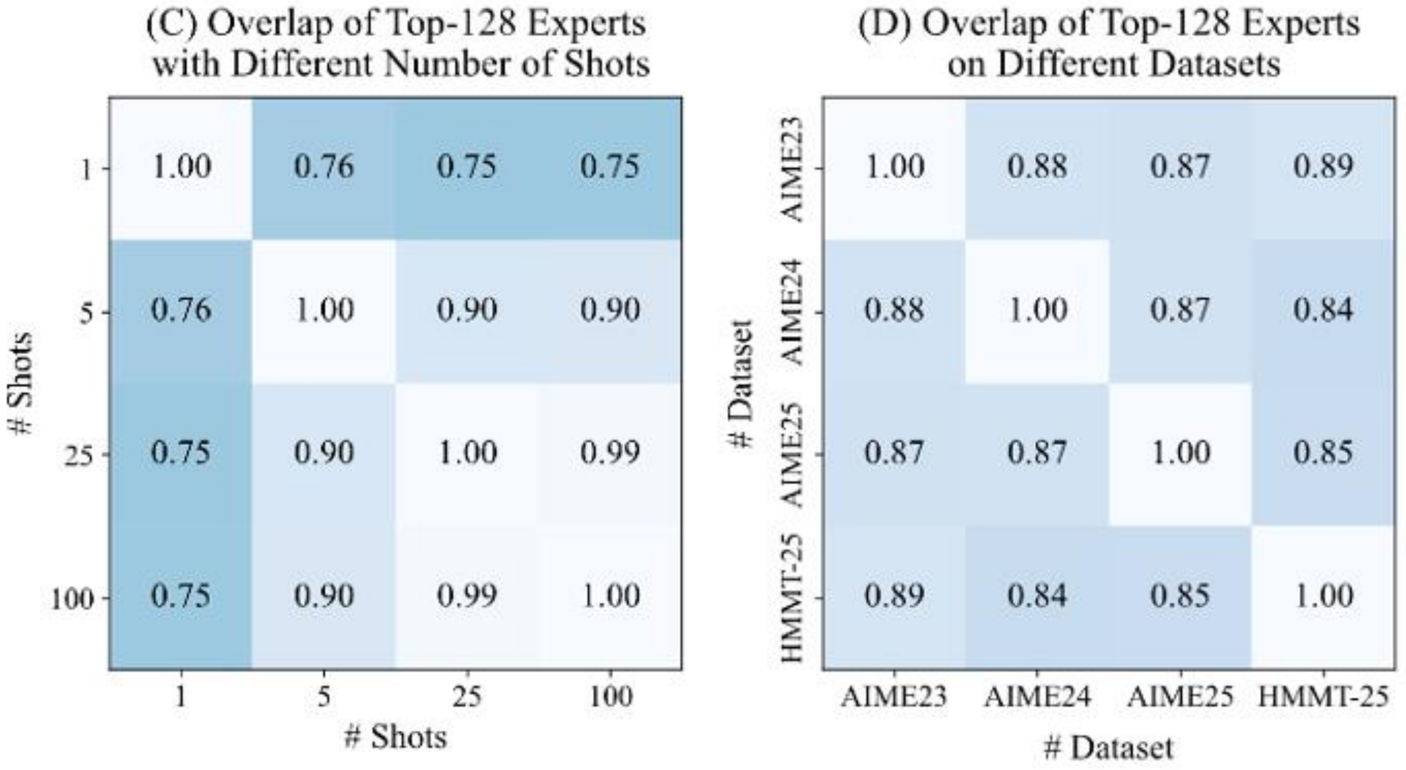
Expert Specialization Across Domains



Domain-specific experts play a critical role in the relevant domain but are redundant for other domains.

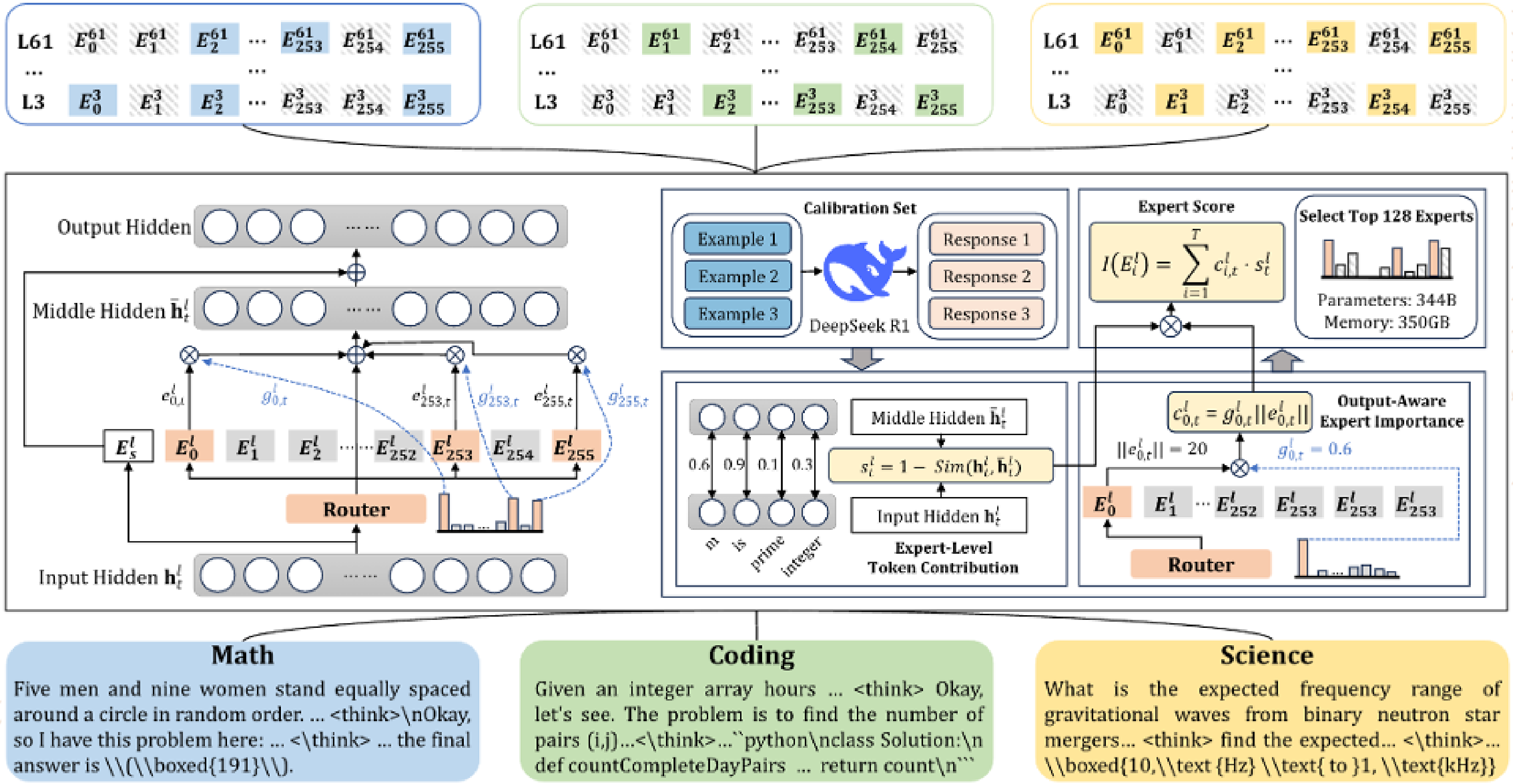
2. Empirical Analysis

Expert Locality Within One Domain



- Domain-specific experts can be identifies with few demonstrations.
- Domain-specific expert activation patterns are largely transferable within the same domain.

3. Method



- Collecting expert activation statistics of the MoE model on target-domain demonstrations and selecting Top-M experts with the largest expert scores.

Expert Score

$$I(\mathrm{E}_i^l) = \sum_{t=1}^{T} c_{i,t}^l \cdot s_t^l.$$

Mixed-Domain Pruning

$$I_{mix}(\mathrm{E}_i^l) = \sum_{\tau \in \mathcal{T}} (I_\tau(\mathrm{E}_i^l) / \sum_{j=1}^{N} I_\tau(\mathrm{E}_j^l)).$$

3. Method

Output-Aware Expert Importance Assessment

Analysis: Each expert's contribution to the final output is bounded by the product of its gating value and the L2 norm of its output.

$$\bar{\boldsymbol{h}}_t^l = \sum_{i=1}^{N} g_{i,t}^l \cdot \boldsymbol{e}_{i,t}^l = \sum_{i=1}^{N} g_{i,t}^l \|\boldsymbol{e}_{i,t}^l\| \cdot \frac{\boldsymbol{e}_{i,t}^l}{\|\boldsymbol{e}_{i,t}^l\|},$$

$$\|\bar{\boldsymbol{h}}_t^l\| \leq \sum_{i=1}^{N} \left\| g_{i,t}^l \|\boldsymbol{e}_{i,t}^l\| \cdot \frac{\boldsymbol{e}_{i,t}^l}{\|\boldsymbol{e}_{i,t}^l\|} \right\| = \sum_{i=1}^{N} g_{i,t}^l \|\boldsymbol{e}_{i,t}^l\|.$$

Output-Aware Expert Importance $\quad c_{i,t}^l = g_{i,t}^l \|\boldsymbol{e}_{i,t}^l\|, \quad \forall g_{i,t}^l > 0.$

3. Method

Expert-Level Token Contribution Estimation

Analysis: When dealing with tokens exhibiting low similarity before and after the MoE module, adjusting their routed experts will induce a substantial distributional shift in their representation

Expert-Level Token Contribution

$$s_t^l = 1 - \mathrm{Sim}(\boldsymbol{h}_t^l, \tilde{\boldsymbol{h}}_t^l).$$

4. Experiment

Model	Method	Mix	#E	AIME-24	AIME-25	FMMT	LiveCode	GPQA	USMLE	FinIQ	A-OS	Avg
DeepSeek-R1	Full	-	256	77.08	66.67	44.38	63.32	70.91	92.66	82.1	40.51	67.20
	Random	×	64	0.00	0.00	0.00	0.00	26.09	0.00	0.00	0.00	3.26
	Frequency	×	64	0.00	0.00	0.00	0.00	17.68	0.00	0.00	2.78	2.58
	Gating Score	×	64	2.67	1.33	2.67	14.97	46.83	0.86	0.00	0.69	8.75
	M-SMoE	×	64	0.00	0.00	0.00	0.00	12.12	0.00	0.00	0.00	1.52
	EASY-EP	×	64	**72.81**	**55.10**	**38.02**	**42.51**	**67.47**	**26.63**	**33.90**	**27.26**	**45.22**
	Random	×	128	8.33	6.67	3.33	20.96	34.95	57.66	0.00	7.64	17.44
	Frequency	×	128	19.33	13.33	7.33	36.08	59.60	61.51	26.40	29.16	31.59
	Gating Score	×	128	70.10	55.52	36.15	47.60	63.78	80.36	66.50	31.94	56.49
	M-SMoE	×	128	5.33	6.00	3.33	25.75	24.75	52.63	39.60	19.44	22.10
	EASY-EP	×	128	**79.17**	**68.33**	**45.31**	**61.11**	**70.12**	**91.67**	**78.80**	**37.92**	**66.55**
	Frequency	✓	128	21.33	10.00	6.00	7.49	41.45	78.55	**62.14**	11.81	29.85
	Gating Score	✓	128	29.33	21.33	18.00	22.75	41.69	62.06	27.29	30.56	31.67
	M-SMoE	✓	128	6.67	2.00	4.67	4.19	32.32	72.00	19.10	6.25	18.40
	EASY-EP	✓	128	**75.94**	**61.98**	**42.50**	**57.63**	**70.36**	**91.20**	57.95	**34.17**	**61.47**
DeepSeek-V3-0324	Full	-	256	55.73	47.71	28.75	48.50	66.87	87.51	64.22	33.33	54.08
	Random	×	64	0.00	0.00	0.00	0.00	26.87	0.39	0.00	0.69	3.49
	Frequency	×	64	31.35	34.06	15.73	1.95	45.25	40.13	61.96	22.74	31.65
	Gating Score	×	64	43.96	25.10	23.12	14.97	51.52	78.68	64.20	0.00	37.69
	M-SMoE	×	64	16.67	13.33	3.33	1.20	22.22	12.18	47.00	21.52	17.18
	EASY-EP	×	64	**53.12**	**41.56**	**28.85**	**27.99**	**57.35**	**84.57**	**72.50**	**27.55**	**49.19**
	Random	×	128	1.33	0.67	0.00	11.38	34.95	53.5	53.66	18.75	21.78
	Frequency	×	128	**55.73**	42.60	30.10	36.08	63.54	84.29	66.84	31.71	51.36
	Gating Score	×	128	55.42	45.10	30.94	**47.60**	63.78	84.62	**67.76**	35.42	53.83
	M-SMoE	×	128	48.00	38.67	28.67	30.53	55.82	86.72	66.60	33.33	48.54
	EASY-EP	×	128	55.21	**46.88**	**31.56**	46.71	**65.25**	**86.72**	63.58	**37.08**	**54.12**
	Frequency	✓	128	51.35	37.60	24.27	17.07	55.90	83.47	66.80	36.25	46.59
	Gating Score	✓	128	53.75	40.10	27.19	28.74	58.88	83.86	67.74	34.58	49.36
	M-SMoE	✓	128	43.33	30.00	20.00	7.19	52.53	82.33	62.20	29.17	40.84
	EASY-EP	✓	128	**57.81**	**46.56**	**33.33**	**40.72**	**64.95**	**85.00**	**72.26**	**38.74**	**54.92**

- EASY-EP can achieve better performances than other method and achieves comparable performances to the full model with half experts.
- Non-reasoning models exhibit greater robustness after pruning.
- EASY-EP preserve performance well under mixed-domain pruning settings.

4. Experiment

Ablation study

Both components in EASY-EP are important.

Method	Metric	Experts	AIME-24	AIME-25	HMMT	LiveCode	GPQA	A-OS
Ours	$g^l_{i,t}\|e^l_{i,t}\| \cdot s^l_t$	64	72.81	55.33	36.00	42.51	67.47	27.26
w/o Token	$g^l_{i,t}\|e^l_{i,t}\|$	64	65.33	49.33	31.33	27.54	56.57	21.53
w/o norm	$g^l_{i,t} \cdot s^l_t$	64	70.00	40.00	23.33	19.76	61.11	18.75
w/o both	$g^l_{i,t}$	64	2.67	1.33	2.67	0.00	20.20	0.69

Generalization Capacities

A certain generalization capy, especially in similar domains.

Domain	AIME24	LiveCodeBench	GPQA	Agent-OS	USMLE	FinIQ
Math	**79.17**	46.11	46.91	3.47	46.43	58.20
Coding	38.00	**61.11**	39.90	15.97	41.79	53.00
Science	64.64	53.59	**70.12**	4.17	75.88	57.50

4. Experiment

Ablation study

Method	Metric	Experts	AIME-24	AIME-25	HMMT	LiveCode	GPQA	A-OS
Ours	$g^l_{i,t}\|e^l_{i,t}\| \cdot s^l_t$	64	72.81	55.33	36.00	42.51	67.47	27.26
w/o Token	$g^l_{i,t}\|e^l_{i,t}\|$	64	65.33	49.33	31.33	27.54	56.57	21.53
w/o norm	$g^l_{i,t} \cdot s^l_t$	64	70.00	40.00	23.33	19.76	61.11	18.75
w/o both	$g^l_{i,t}$	64	2.67	1.33	2.67	0.00	20.20	0.69

Both components in EASY-EP are important.

Generalization Capacities

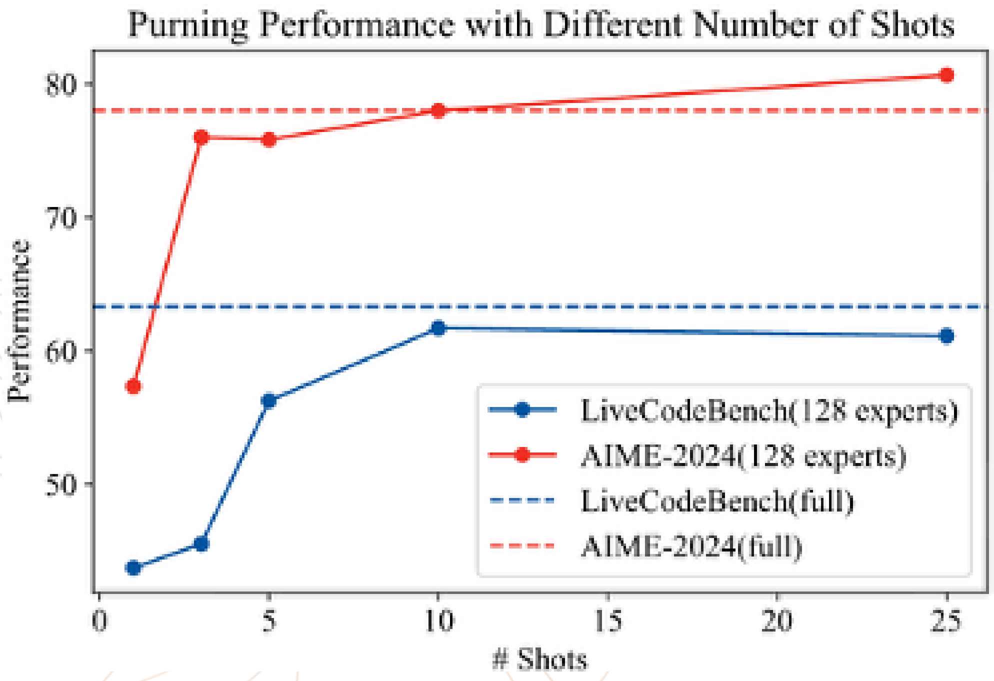
Domain	AIME24	LiveCodeBench	GPQA	Agent-OS	USMLE	FinIQ
Math	**79.17**	46.11	46.91	3.47	46.43	58.20
Coding	38.00	**61.11**	39.90	15.97	41.79	53.00
Science	64.64	53.59	**70.12**	4.17	75.88	57.50

A certain generalization capacity, especially in similar domains.

4. Experiment

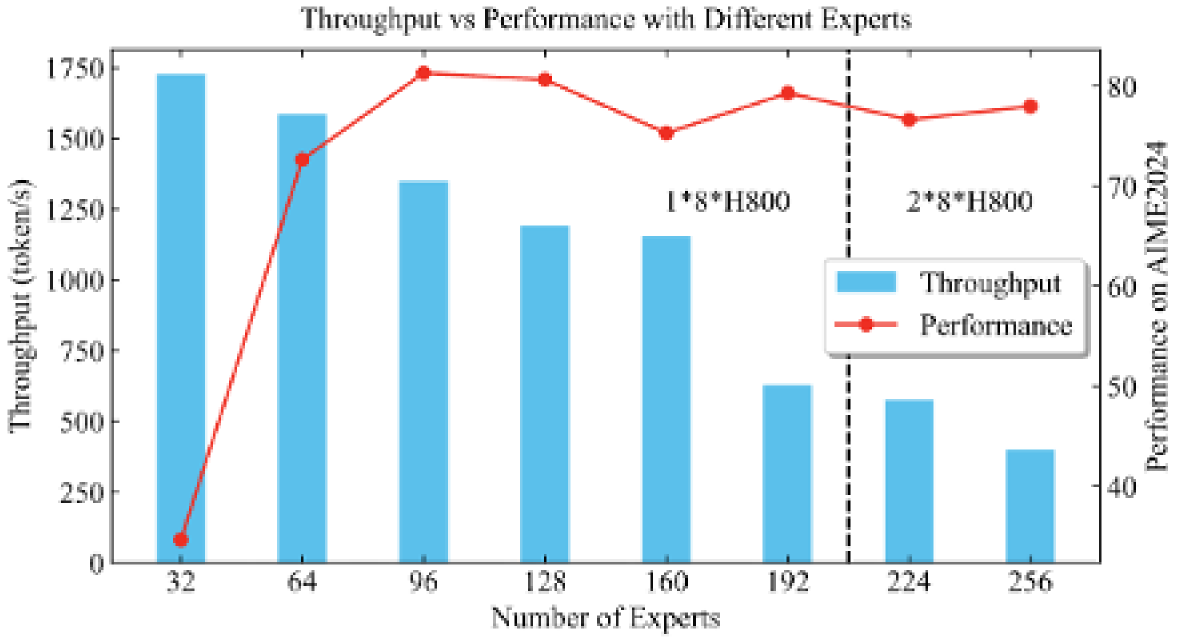
Number of Demonstrations

Only few demonstrations can achieve comparable performance with the baselines.



Throughput

Compared to the full model, the pruned models presents improved throughputs.



THANKS!