Chufan Shi^{1*}, Cheng Yang^{1*}, Xinyu Zhu^{2*}, Jiahao Wang^{3*}, Taiqiang Wu³, Siheng Li¹, Deng Cai⁴, Yujiu Yang^{1†}, Yu Meng^{2†}

¹Tsinghua University ²University of Virginia ³The University of HongKong ⁴Tencent AI Lab ^{*}Equal Contribution [†]Corresponding authors









Background

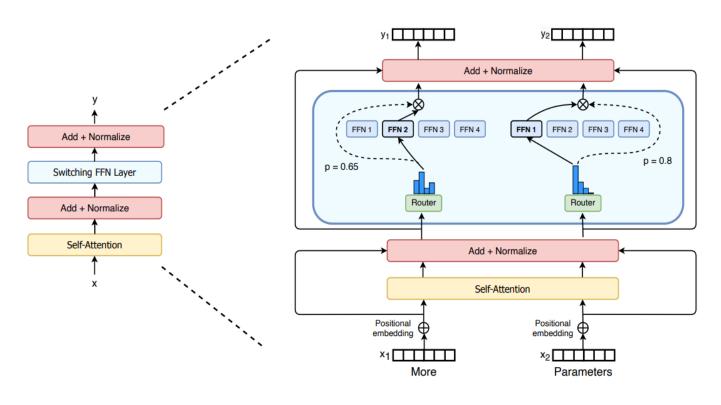


Fig I.An Illustration of MoE Models

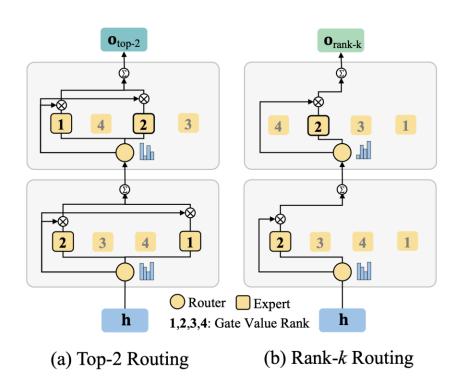


Fig 2.An Illustration of Routing Strategies

Motivation

- At the inference stage, only a small portion of trained experts are used.
- The potential of utilizing more experts during the inference stage to enhance MoE performance remain underexplored.

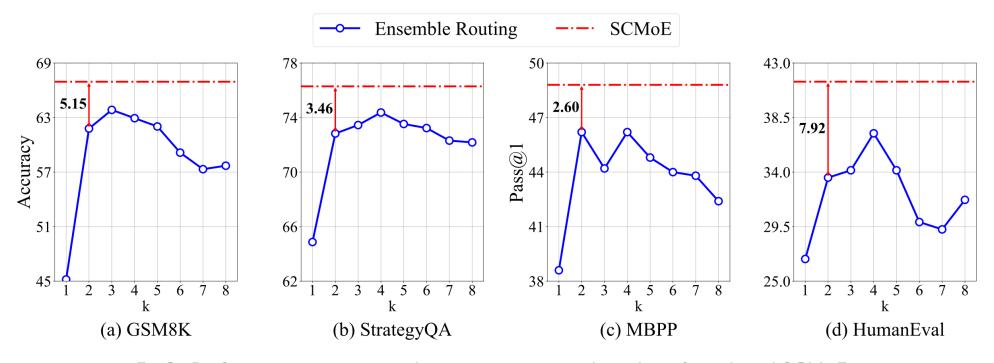


Fig 3. Performance comparison between increasing the value of top-k and SCMoE.

Analysis



Fig 4. Heatmap of KLD between the output distribution of top-2 routing and different rank-k routing strategies

Analysis

Question: Hallie had dance practice for 1 hour on Tuesdays and 2 hours on Thursdays. On Saturdays, she had dance practice that lasted twice as long as Tuesday's night class. How many hours a week did she have dance practice?

Answer: On Tuesdays, she had 1 hour. On Thursdays, she had 2 hours. On Saturdays, she had 2 x 1 = 2 hours. 1 + 2 + 2 = 5. The answer is 5.

Finding I:

The overall next-token prediction capability gap between top-2 and rank-k routing.

Finding 2:

Differences between top-2 and rank-k routing in generating tokens for reasoning.

Finding 3:

Predictions for function or punctuation words pose fewer challenges for rank-k routing.

-	0.58	0.13	0.00	0.00	0.07	0.08	0.30	0.22	0.01	0.03	0.26	0.15	0.41	0.01	0.00	0.03	0.03	0.03	0.02
1 -	4.47	9.08	0.80	7.49	0.83	2.37	1.82	3.68	10.14	6.22	4.67	5.74	0.98	0.60	1.61	0.32	0.57	0.56	4.24
m -	13.28	21.01	32.38	18.95	2.28	4.01	10.01	3.77	14.40	24.97	7.13	11.05	20.52	14.14	7.24	0.72	1.14	0.72	9.29
[⊼] 4 -	10.25	15.63	42.83	13.24	2.13	8.69	2.13	6.83	19.38	12.78	6.99	10.80	12.69	42.44	33.53	1.25	2.22	1.66	4.71
5	19.47	20.13	41.64	25.90	4.44	11.03	4.57	5.81	8.91	17.86	12.81	14.58	22.04	41.81	25.01	1.59	7.06	3.93	10.35
9 -	13.13	23.31	37.83	29.38	10.13	10.82	13.83	13.06	14.18	28.37	10.11	15.50	18.38	43.11	43.28	2.78	21.61	8.86	10.05
۲ -	14.01	25.12	44.22	38.95	11.00	12.95	25.68	13.43	9.77	24.24	22.05	16.19	30.56	40.17	23.84	7.92	13.07	4.34	12.55
∞ -	17.44	28.31	44.39	39.08	22.93	25.15	22.92	12.74	27.01	27.95	25.31	25.77	28.56	47.76	19.16	19.52	24.71	9.18	18.52
	_On	_T	ues	days	,	_she	_had	_	1	_hour		_On	_Th	urs	days	,	_she	_had	_
										Token									
-	0.00	0.05	0.05	0.24	0.07	0.00	0.00	0.02	0.08	0.11	0.08	0.13	0.36	0.03	0.00	0.09	0.00	0.00	0.03
6 -	0.14	2.21	6.59	6.67	1.65	0.14	3.03	0.70	0.92	0.51	6.65	3.77	3.99	1.12	3.42	5.25	1.25	17.88	4.21
m -	5.26	17.42	8.47	5.95	17.19	27.46	29.35	0.33	0.66	0.79	4.28	8.15	9.72	2.69	14.61	11.23	1.94	22.71	11.89
첫 4 -	2.56	3.25	15.87	7.81	19.88	57.47	20.38	0.77	1.29	3.45	3.46	3.68	8.30	7.03	15.62	18.09	4.16	4.55	25.02
_ ~ -	10.34	5.22	13.25	8.61	17.92	35.02	9.39	1.61	1.54	14.45	7.57	3.53	12.02	14.45	14.18	25.35	7.75	18.49	19.00
9 -	12.56	9.27	11.91	26.61	17.27	44.09	35.89	2.06	26.89	12.76	11.40	8.59	16.22	26.15	25.50	27.20	35.83	20.21	24.88
۲ -	12.64	25.72	16.99	25.64	23.05	51.35	14.18	10.16	12.97	15.87	7.70	6.52	16.72	34.41	29.18	19.97	9.31	8.85	18.95
∞ -	8.33	28.44	22.56	13.88	26.40	49.99	39.33	10.76	18.28	13.59	8.75	2.11	20.45	41.83	23.79	26.50	10.79	18.05	26.50
	2	_hours		_On	_S	atur	days	,	_she	_had	_	2	_x	_	1	_=	_	2	_hours
										Token									
	0.11	0.23	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.11	0.06	0.01	0.00	0.00	0.00	0.09
6 -	7.59	3.59	10.36	8.58	0.25	0.11	18.14	1.15	4.38	8.48	0.43	15.52	1.13	2.80	0.20	0.66	0.95	1.02	2.44
m -	3.43	6.78	12.83	13.76	1.15	5.11	19.74	2.72	13.61	6.19	2.03	14.49	5.94	5.31	0.61	0.88	3.76	8.08	2.67
[⊼] 4 -	16.16	7.98	15.31	18.47	1.30	1.58	19.35	3.72	8.46	8.07	4.83	17.82	14.10	5.57	4.53	2.11	10.26	21.14	3.86
_ v -	5.80	7.85	9.87	23.86	4.19	9.81	23.53	15.80	13.13	12.40	3.85	15.49	18.82	11.91	8.14	0.65	7.20	23.01	10.11
9 -	8.10	13.57	15.68	24.17	6.80	32.52	32.04	13.25	39.56	16.31	4.77	21.47	20.94	13.39	10.49	9.27	9.84	22.76	7.47
7	14.89	22.39	13.63	27.97	25.11	2.68	28.18	14.38	11.06	17.52	10.43	17.54	17.56	16.11	14.00	9.06	18.09	21.76	19.00
∞ -	12.70	19.22	20.47	30.22	11.78	33.42	28.72	9.64	20.80	27.48	11.88	13.19	21.87	24.01	35.70	2.29	13.41	38.88	22.66
			1	+		2	+		2	=		5		The	answer	ic		5	
	•		1	_+			'					3	•	_1116	_answer	_18	_	3	•

Method

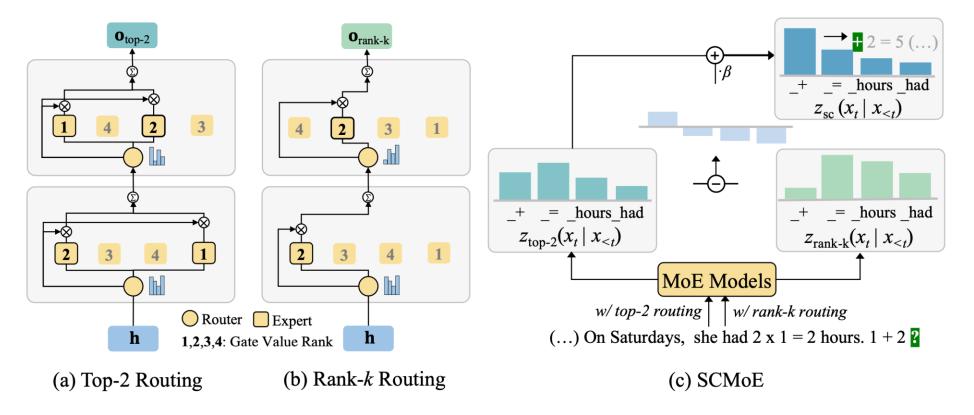


Fig 5. Illustration of our method——SCMoE

$$z_{sc}(x_t = i|x_{< t}) = \begin{cases} (1 + \beta) \cdot z_{top-2}(x_t = i|x_{< t}) & -\beta \cdot z_{rank-k}(x_t = i|x_{< t}) & i \in \mathcal{V}_{valid} \\ -\infty & i \notin \mathcal{V}_{valid} \end{cases}$$

$$\mathcal{V}_{valid} = \{ i \mid z_{top-2}(x_t = i|x_{< t}) \geq \log \alpha + \max_{j \in \mathcal{V}} z_{top-2}(x_t = j|x_{< t}) \}$$

Method

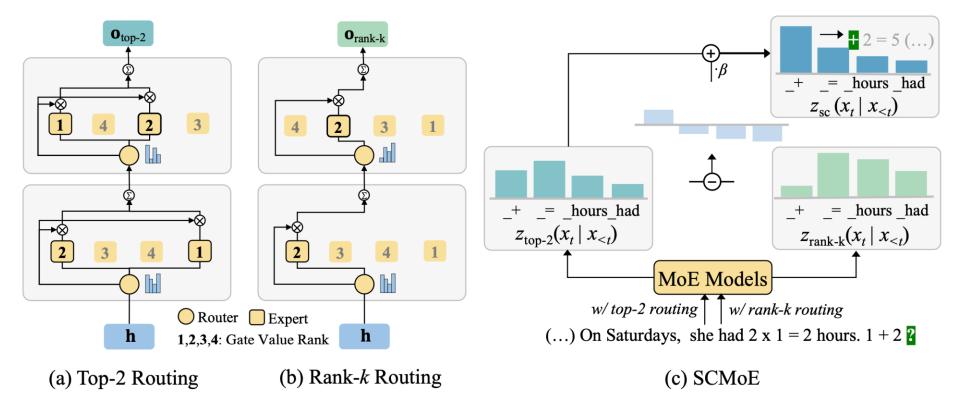


Fig 5. Illustration of our method——SCMoE

Experimental Results

Method	GSM8K	StrategyQA	MBPP	HumanEval							
Greedy	61.79	72.83	46.20	33.54							
Routing-based											
Dynamic Routing	61.11	74.41	47.80	38.41							
Ensemble Routing	63.84	74.37	46.20	37.20							
	Search-based										
Contrastive Search	60.96	74.85	46.20	36.59							
DoLa	49.96	71.04	33.00	12.80							
Contrastive Decoding	62.24	74.45	45.20	35.98							
SCMoE	66.94	76.29	48.80	41.46							

Our method (SCMoE) is effective across four reasoning benchmarks and computationally lightweight.

Method	Greedy	Ensemble	Dynamic	CS	DoLa	CD	SCMoE
Latency (s / 512 tokens) Latency Ratio	50.32	59.82	54.85	81.73	53.30	72.04	65.47
	x1.00	x1.19	x1.09	x1.62	x1.06	x1.43	x1.30

Experimental Analysis

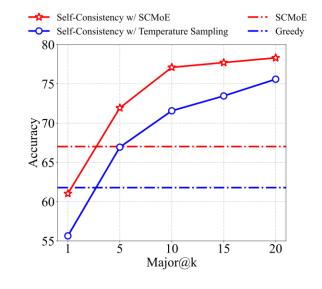
(I) Impact of Weak & Strong Activation

By carefully selecting activation combinations, we can achieve even better performance.

Method	GSM8K	StrategyQA	MBPP	HumanEval
SCMoE	66.94	76.29	48.80	41.46
SCMoE w/ ideal strong activations	68.92	76.42	50.60	41.46

(2) Combination with Self-Consistency

SCMoE works as a decoding strategy and works well with self-consistency.



(3) Proportion of Unchosen Experts

SCMoE can effectively utilize nearly 50% or more of unchosen experts.

rank-k	1	2	3	4	5	6	7	8
unchosen expert ratio (%)	2.81	46.21	72.62	80.54	84.61	87.79	90.44	90.96

Thank you:)