Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

Chufan Shi[1*], Cheng Yang[1*], Xinyu Zhu[2*], Jiahao Wang[3*],

Taiqiang Wu[3], Siheng Li[1], Deng Cai[4], Yujiu Yang[1†], Yu Meng[2†]

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

Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

- **Background**

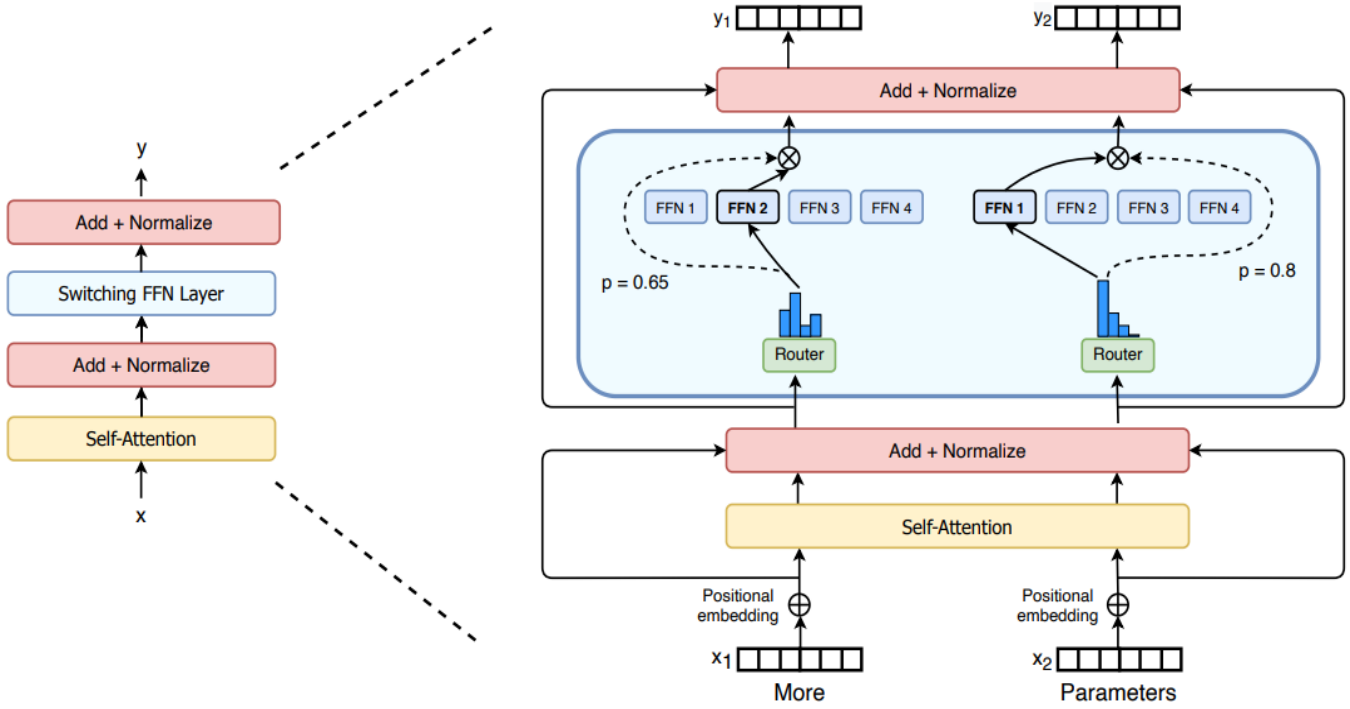


Fig 1. An Illustration of MoE Models

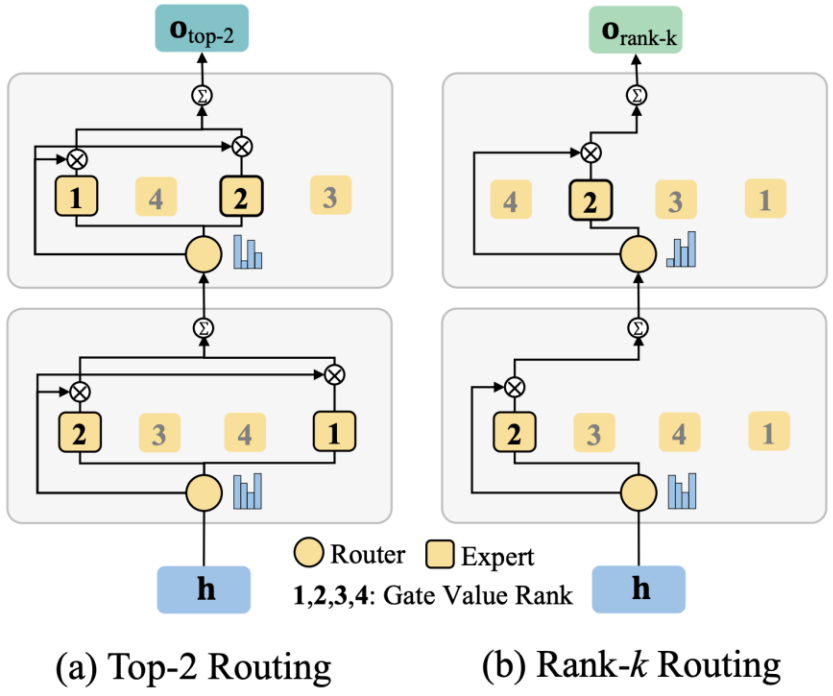


Fig 2. An Illustration of Routing Strategies

Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

- **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.

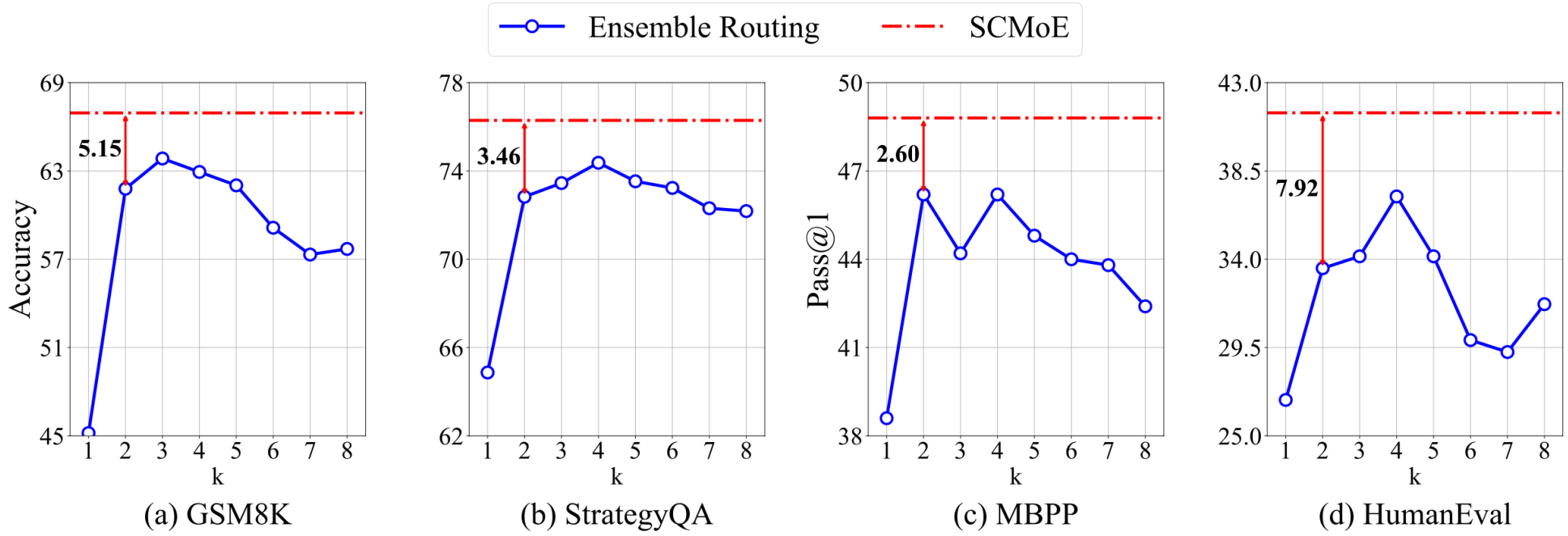


Fig 3. Performance comparison between increasing the value of top-k and SCMoE.

Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

- **Analysis**

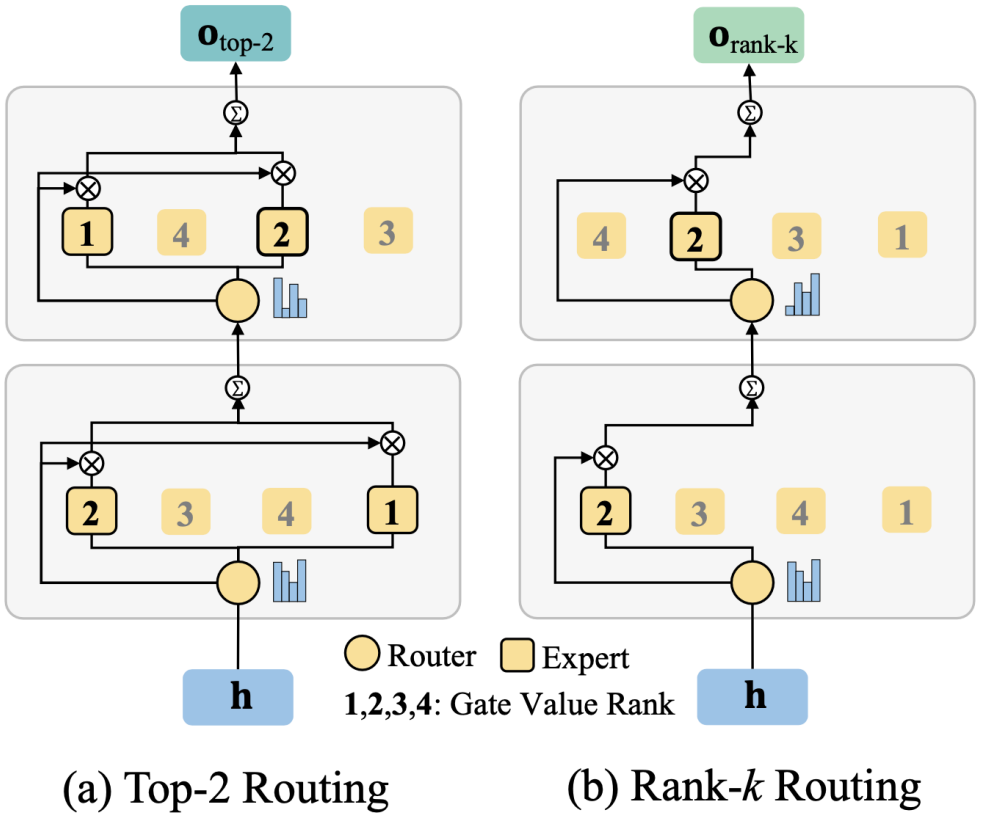


Fig 4. Heatmap of KLD between the output distribution of top-2 routing and different rank-k routing strategies

Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

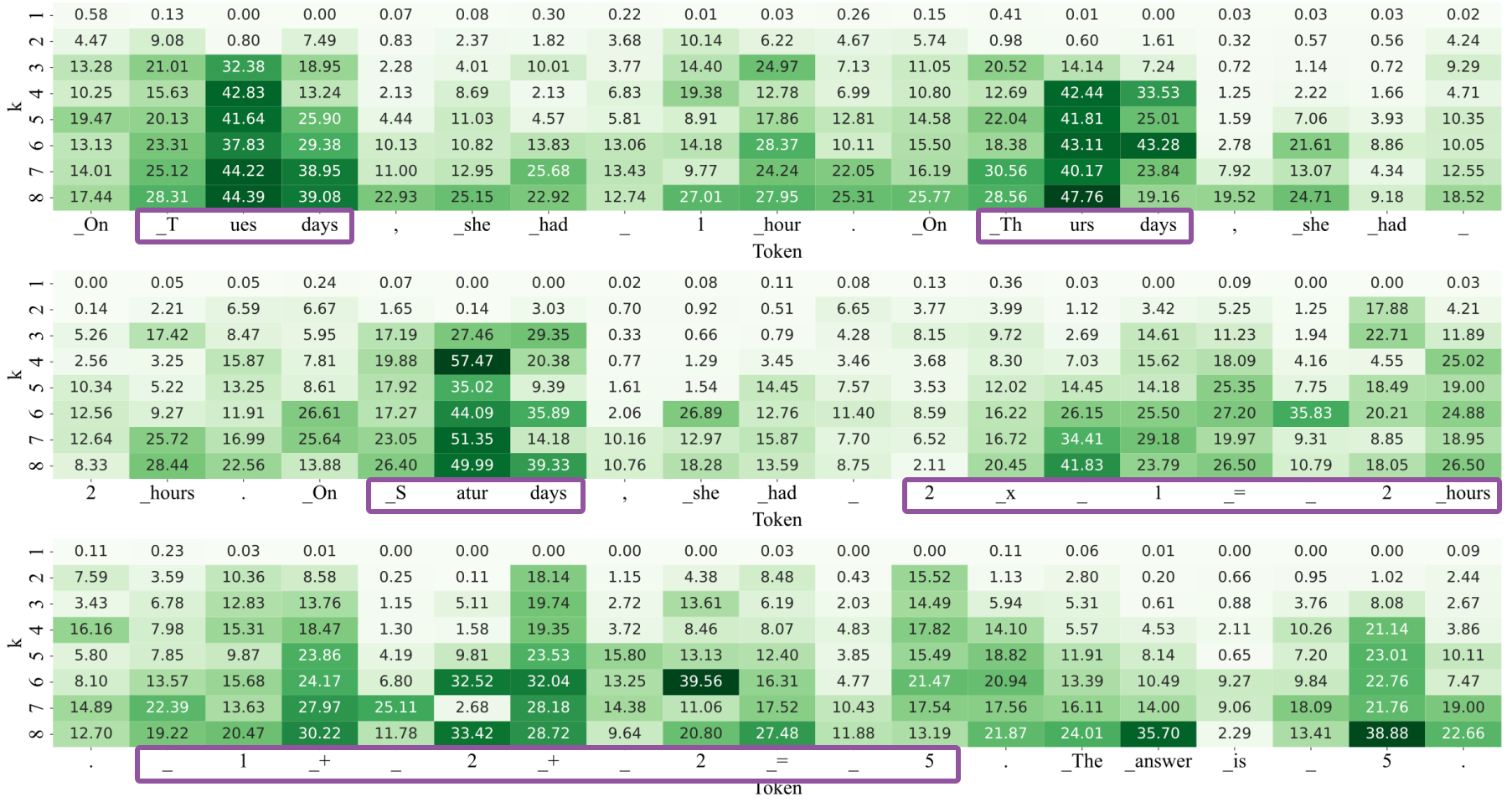
- **Analysis**

Finding 1:
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.

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.



Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

- **Method**

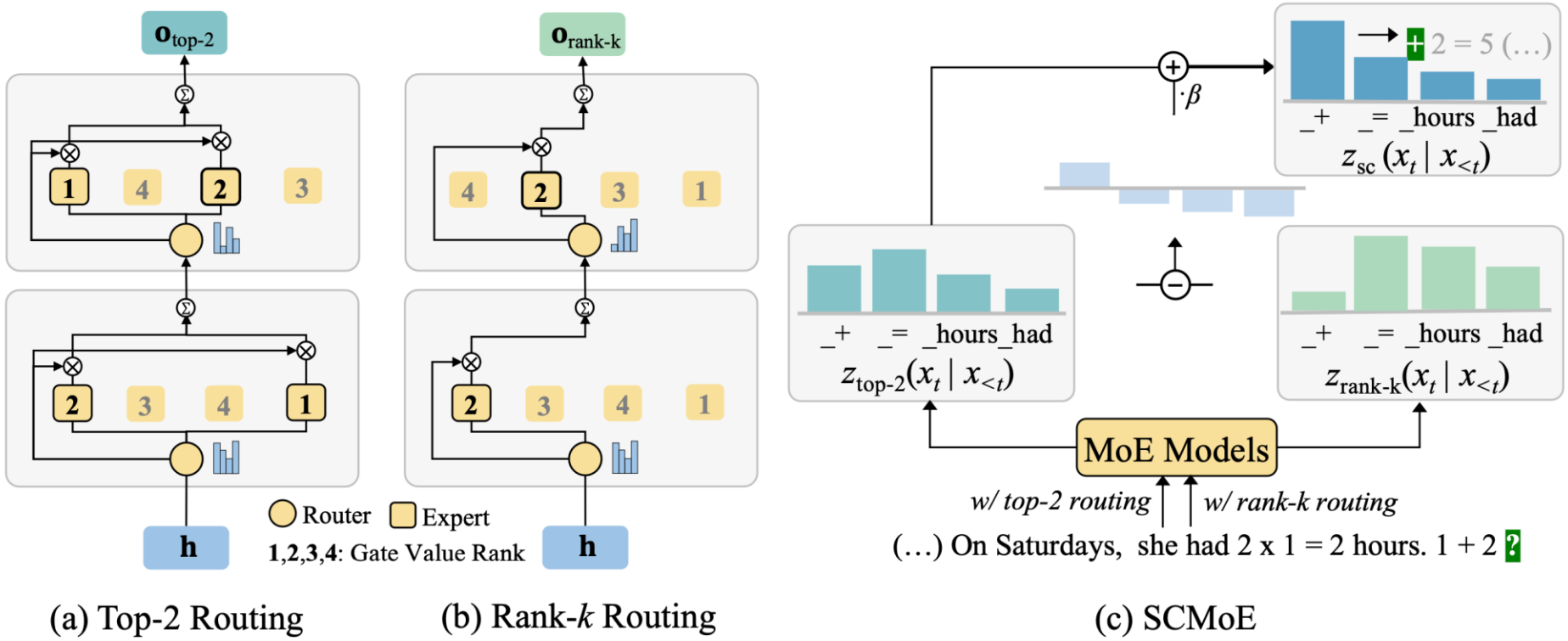


Fig 5. Illustration of our method——SCMoE

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

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

Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

- **Method**

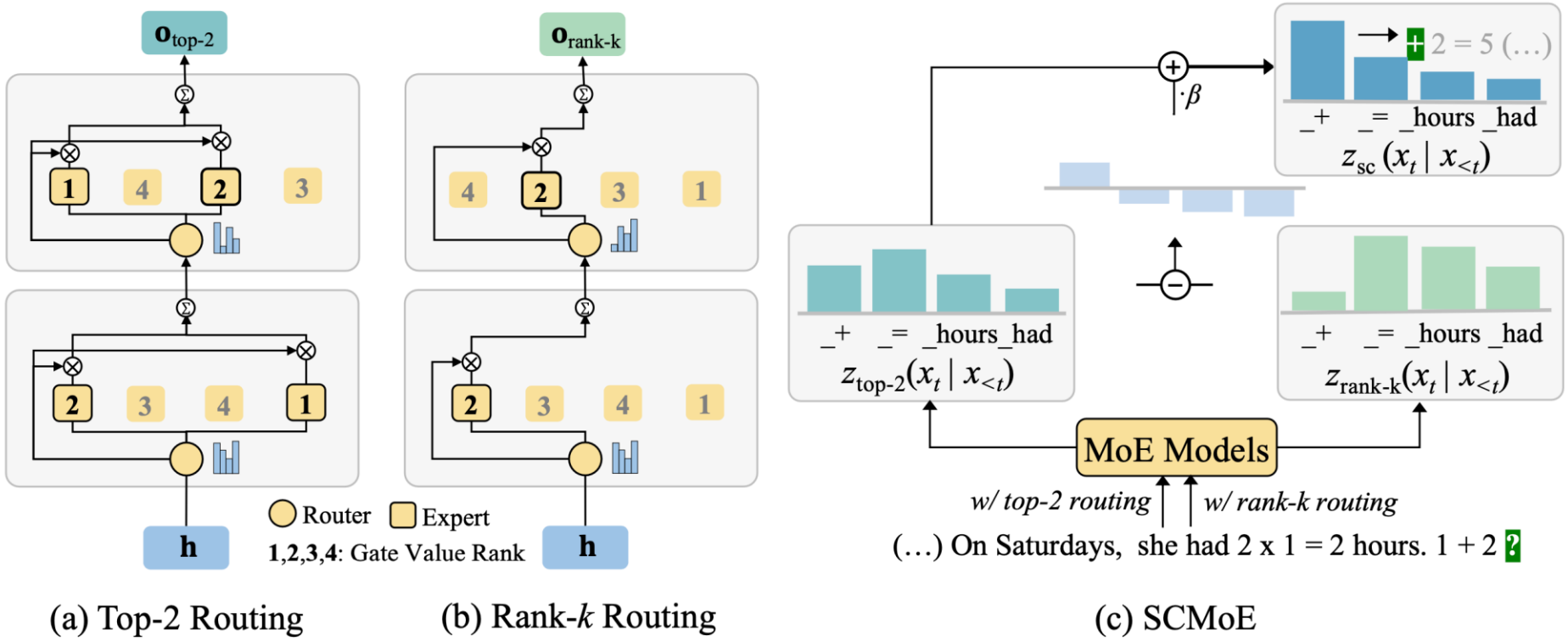


Fig 5. Illustration of our method——SCMoE

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

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

Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

- **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)	50.32	59.82	54.85	81.73	53.30	72.04	65.47
Latency Ratio	x1.00	x1.19	x1.09	x1.62	x1.06	x1.43	x1.30

Unchosen Experts Can Contribute Too: Unleashing MoE Models' Power by Self-Contrast

- **Experimental Analysis**

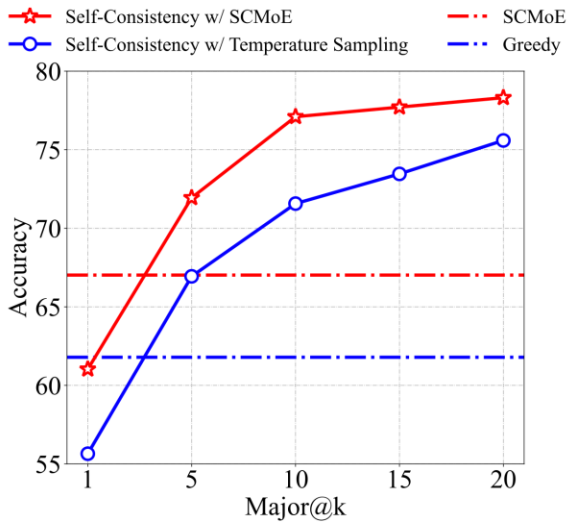
(1) Impact of Weak & Strong Activation

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 :)