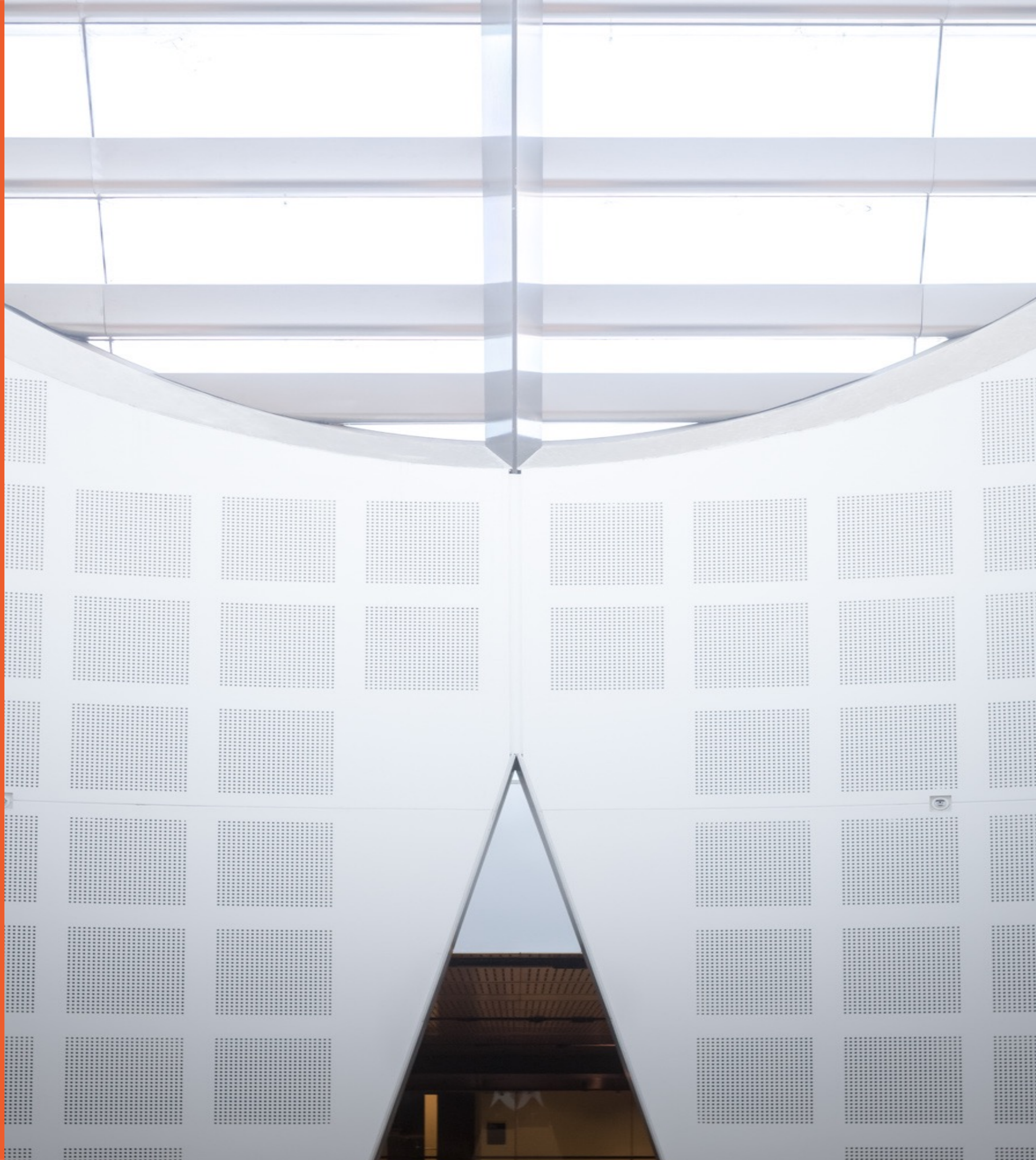
Mixture of Experts for GenAI

Anh-Dung Dinh
anh-dung.dinh@sydney.edu.au
School of Computer Science

THE UNIVERSITY OF SYDNEY



General Assumptions

- More parameters, greater performance

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	**37.5**	**71.9**	**63.6**	**69.4**	**35.2**	**68.9**	**51.2**	**54.2**

- More parameters, larger inference time

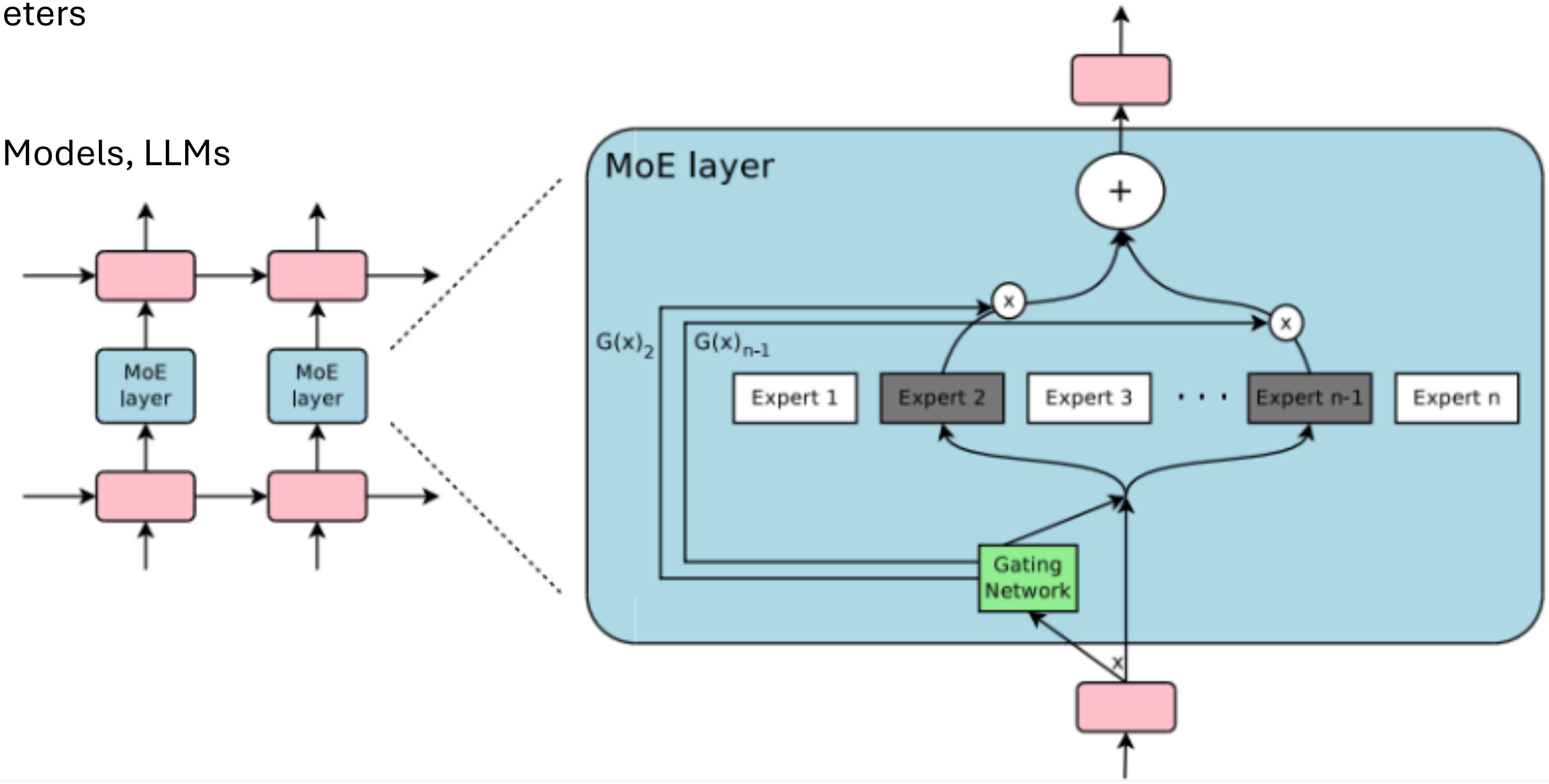
How about NAS, pruning?

- Losing performance is not cool
- Too complicated when training with LLMs, Diffusion
- Some important parameters will be totally removed after NAS/pruning which is hindering the generalization

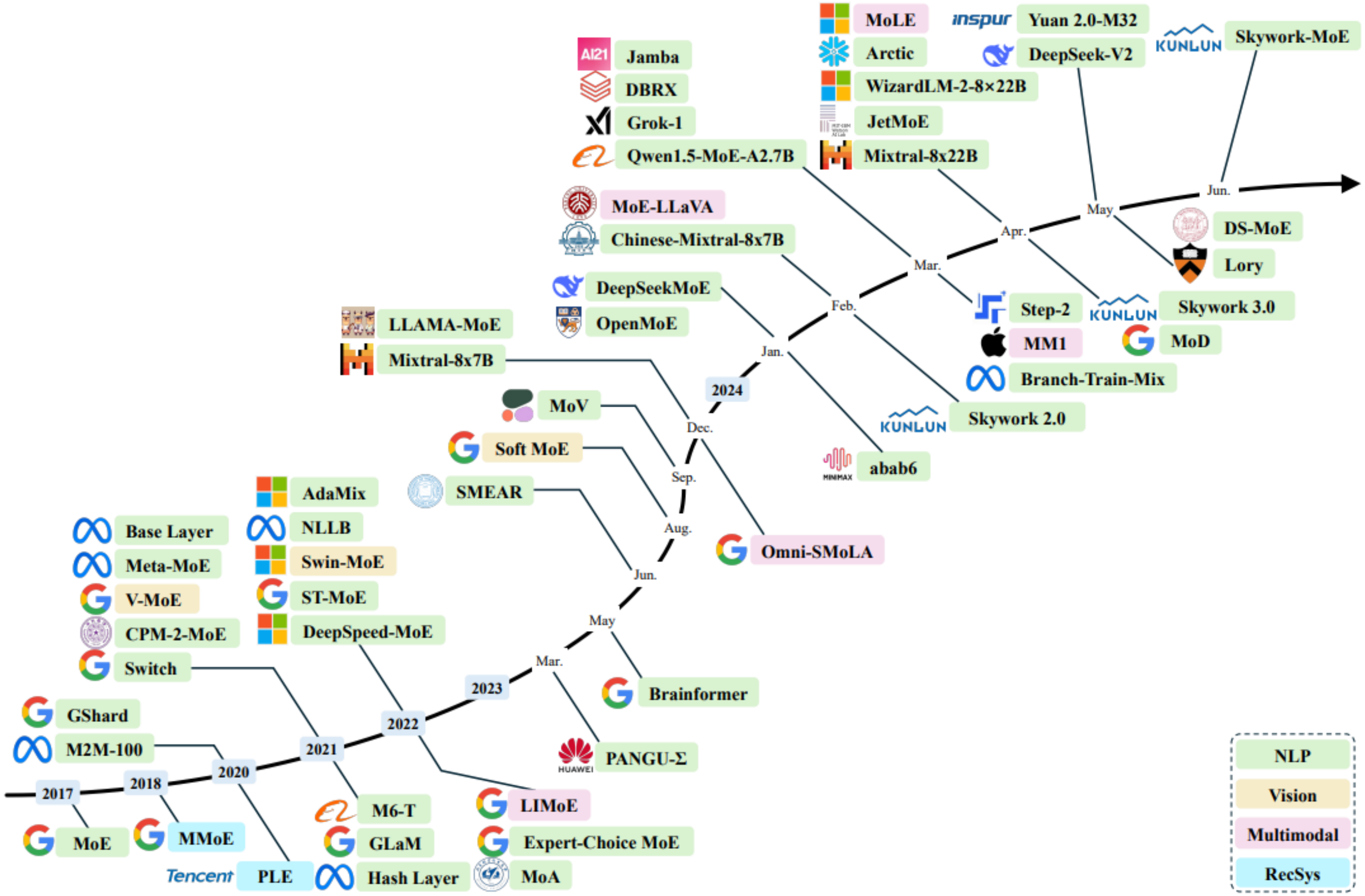
Mixture of Experts

What we have:

1. Large number of parameters
2. Low inference time
3. Good performance
4. Applicable to Diffusion Models, LLMs



A timeline of MoE



MoE details - Sparsity

1. We add some noise

$$H(x)_i = (x \cdot W_{\mathrm{g}})_i + \mathrm{StandardNormal}() \cdot \mathrm{Softplus}((x \cdot W_{\mathrm{noise}})_i)$$

2. We only pick the top k

$$\mathrm{KeepTopK}(v, k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v, \\ -\infty & \text{otherwise.} \end{cases}$$

3. We apply the softmax.

$$G(x) = \mathrm{Softmax}(\mathrm{KeepTopK}(H(x), k))$$

MoE details – Load Balancing

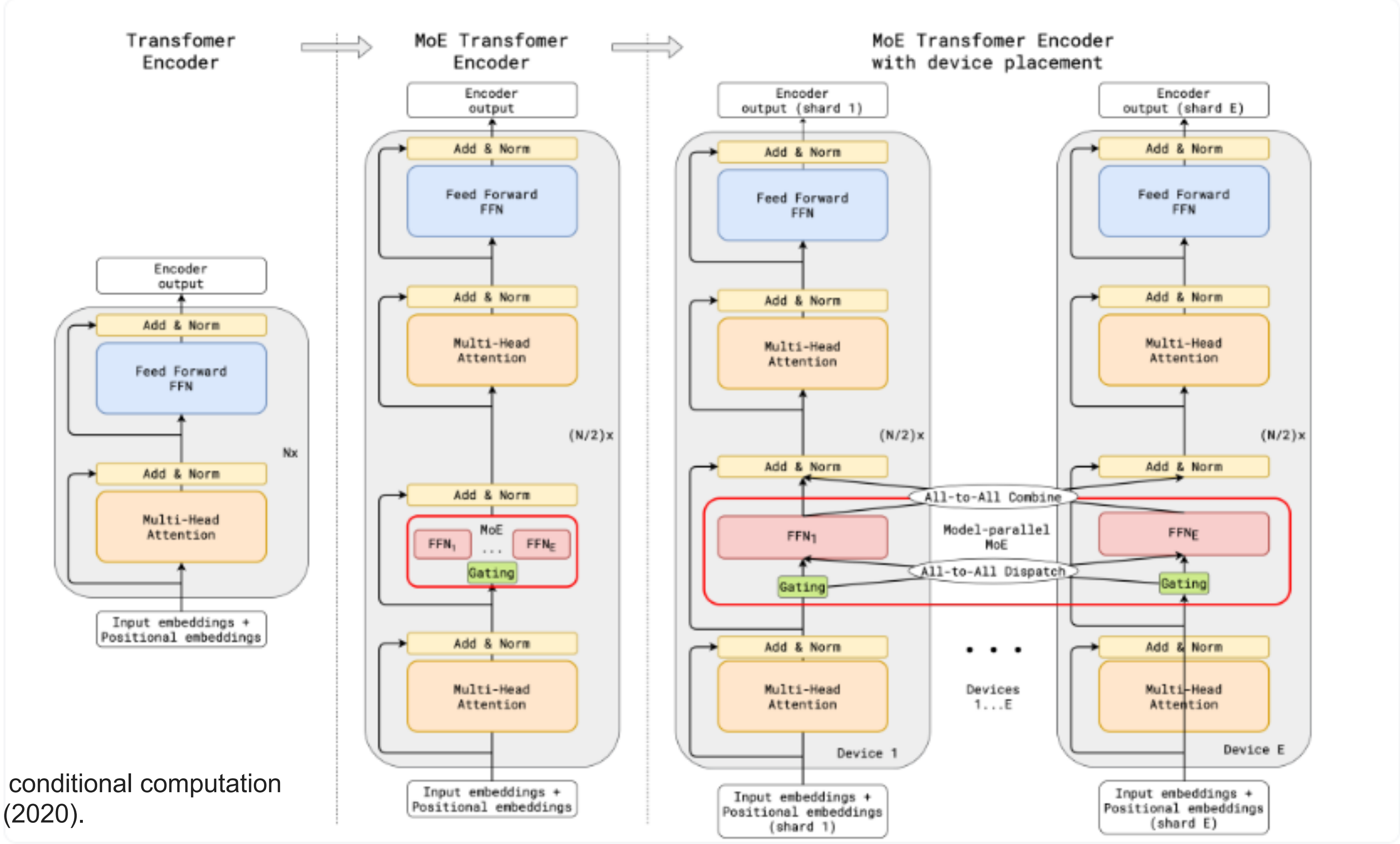
Gshard [1] replaced the FFN layer with MoE layer :

- **Random routing**: Top-2 setup, pick the top one and the second top is picked by proportional probability

- **Expert capacity**: Each expert can only process a number of tokens. If the maximum is reached, the token will be forward toward the next expert.
- **Auxiliary loss**

$$Importance(X) = \sum_{x \in X} G(x)$$

$$L_{importance}(X) = w_{importance} \cdot CV(Importance(X))^2$$

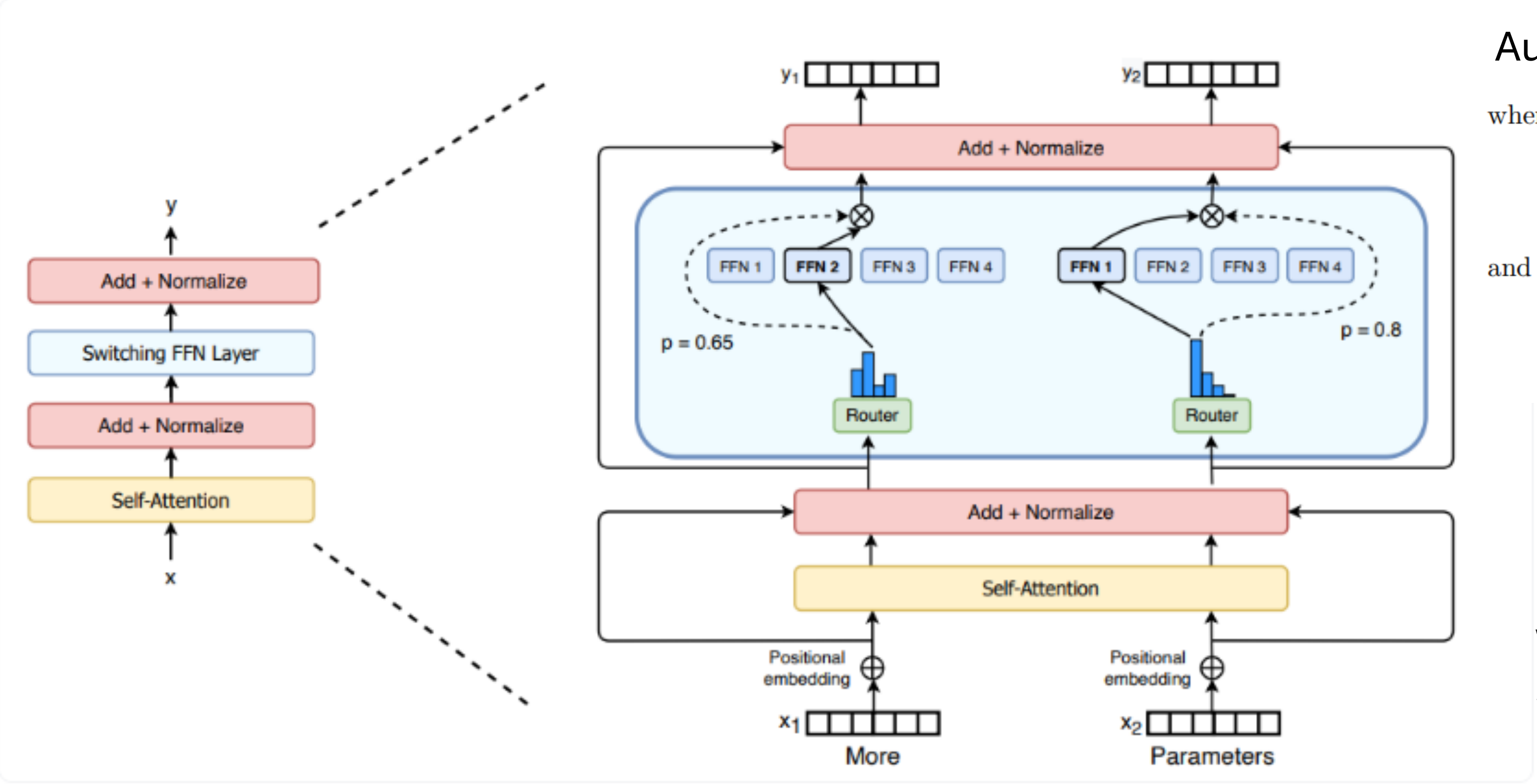


Lepikhin, Dmitry, et al. "Gshard: Scaling giant models with conditional computation and automatic sharding." *arXiv preprint arXiv:2006.16668* (2020).

MoE details – Load Balancing

Switch Transformer [2] (2048 experts):

$$\text{Expert Capacity} = \left(\frac{\text{tokens per batch}}{\text{number of experts}}\right) \times \text{capacity factor}$$



Auxiliary loss: $$\text{loss} = \alpha \cdot N \cdot \sum_{i=1}^{N} f_i \cdot P_i$$

where f_i is the fraction of tokens dispatched to expert i,

$$f_i = \frac{1}{T} \sum_{x \in \mathcal{B}} \mathbb{1}\{\text{argmax}\, p(x) = i\}$$

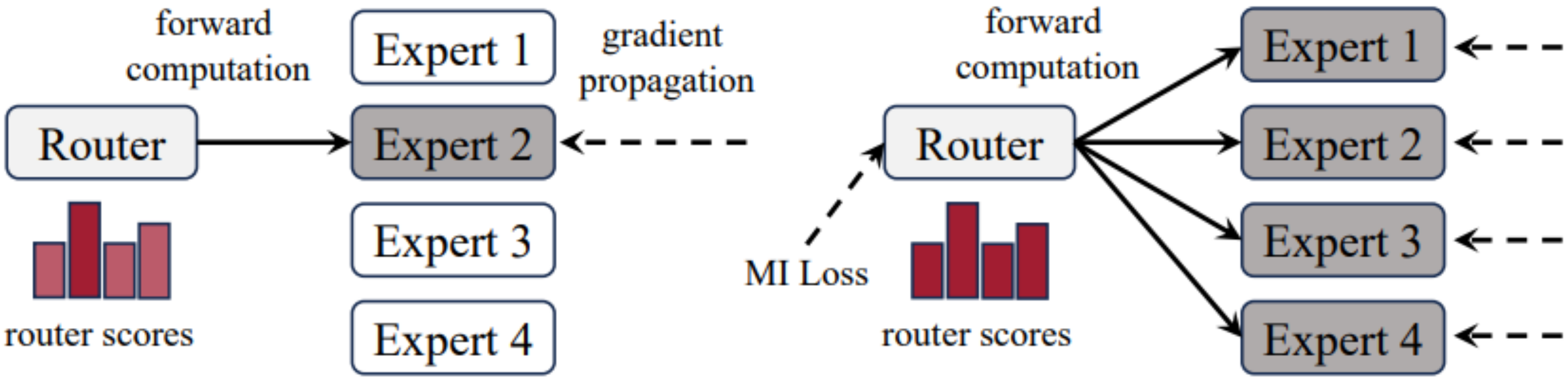
and P_i is the fraction of the router probability allocated for expert i,

$$P_i = \frac{1}{T} \sum_{x \in \mathcal{B}} p_i(x).$$

Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." *Journal of Machine Learning Research* 23.120 (2022): 1-39.

MoE details – Recent aproaches

DS-MoE: Dense training, but sparse inference

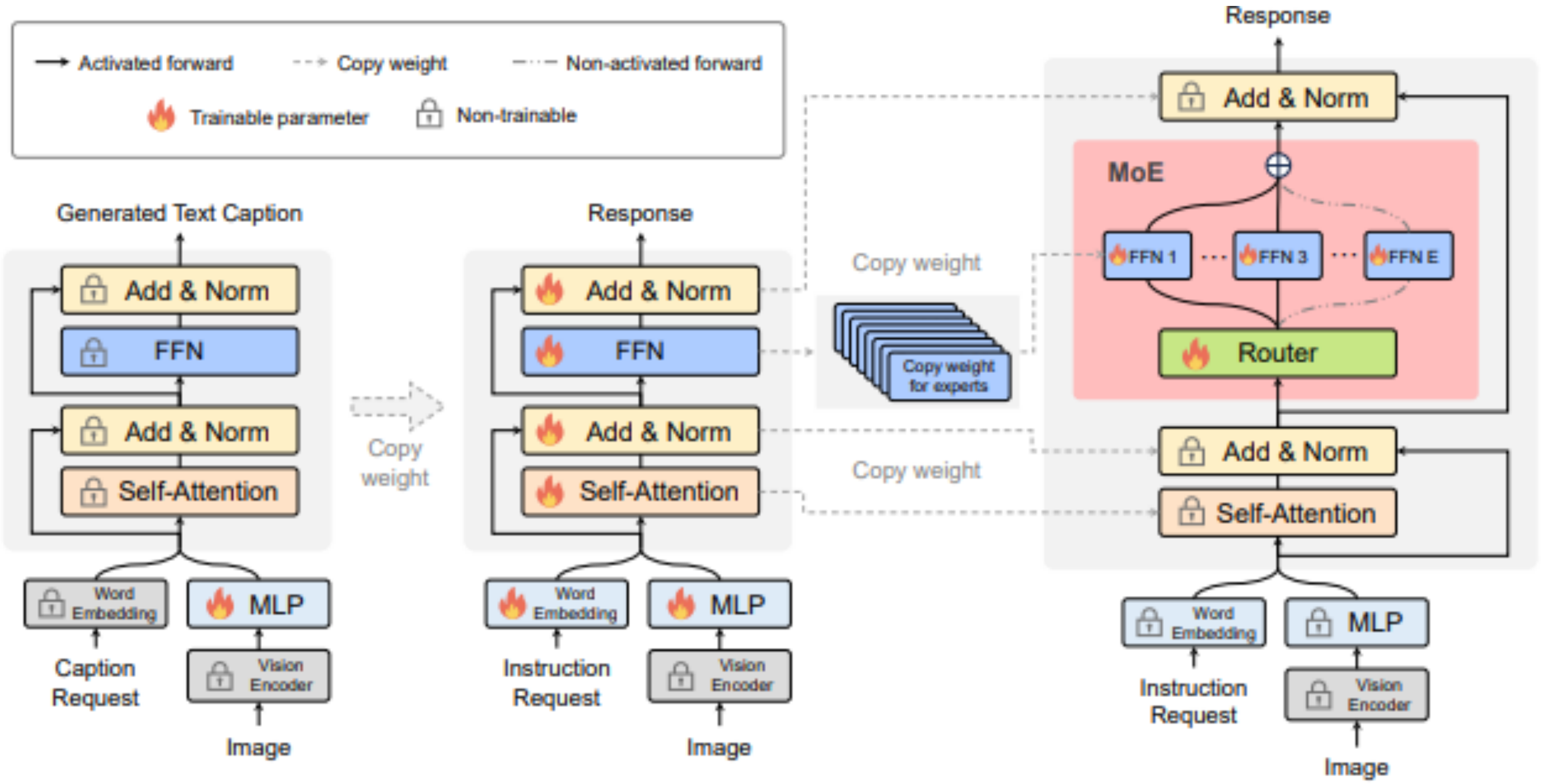


(a) Sparse Training (b) Dense Training

Pan, Bowen, et al. "Dense Training, Sparse Inference: Rethinking Training of Mixture-of-Experts Language Models." *arXiv preprint arXiv:2404.05567* (2024).

MoE details – Recent aproaches

MoE-Llava: Mixture of Experts for Large Vision-Language Models



Lin, Bin, et al. "Moe-llava: Mixture of experts for large vision-language models." *arXiv preprint arXiv:2401.15947* (2024).

MoE applications – Mistral of Experts

Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
context_len	32768
vocab_size	32000
num_experts	8
top_k_experts	2

	LLaMA 2 70B	GPT-3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	**70.6%**
HellaSwag (10-shot)	**87.1%**	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	**85.8%**
WinoGrande (5-shot)	**83.2%**	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	**60.7%**
GSM-8K (5-shot)	53.6%	57.1%	**58.4%**
MT Bench (for Instruct Models)	6.86	**8.32**	8.30

[3] Jiang, Albert Q., et al. "Mixtral of experts." *arXiv preprint arXiv:2401.04088* (2024).

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c	NQ	TriQA	HumanE	MBPP	Math	GSM8K
LLaMA 2 7B	7B	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	17.5%	56.6%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	13B	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	16.7%	64.0%	18.9%	35.4%	6.0%	34.3%
LLaMA 1 33B	33B	56.8%	83.7%	76.2%	82.2%	79.6%	54.4%	24.1%	68.5%	25.0%	40.9%	8.4%	44.1%
LLaMA 2 70B	70B	69.9%	**85.4%**	**80.4%**	82.6%	79.9%	56.5%	25.4%	**73.0%**	29.3%	49.8%	13.8%	69.6%
Mistral 7B	7B	62.5%	81.0%	74.2%	82.2%	80.5%	54.9%	23.2%	62.5%	26.2%	50.2%	12.7%	50.0%
Mixtral 8x7B	13B	**70.6%**	84.4%	77.2%	**83.6%**	**83.1%**	**59.7%**	**30.6%**	71.5%	**40.2%**	**60.7%**	**28.4%**	**74.4%**

MoE applications – SegMoE

SegMoE4x2: 4 experts, 2 selected

Combine between Stable Diffusion and Stable Diffusion XL

https://blog.segmind.com/introducing-segmoe-segmind-mixture-of-diffusion-experts/

three green glass bottles

SegMoE2x1 SegMoE4x2 Baseline

panda bear with aviator glasses on its head