Mixture of Experts (MoEs)

Exploring trade-offs in MoE routing designs

Florian Kowarsch

nuseum

Florian Kowarsch

2021-2023

TU WIEN

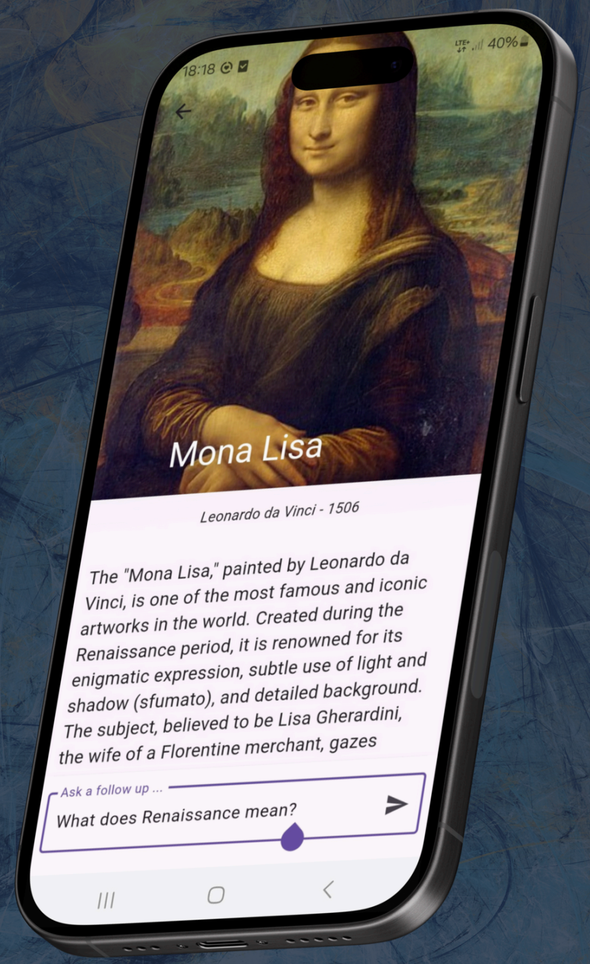
CVL Computer Vision Lab

2023-2025

SEVENTH SENSE

nuseum

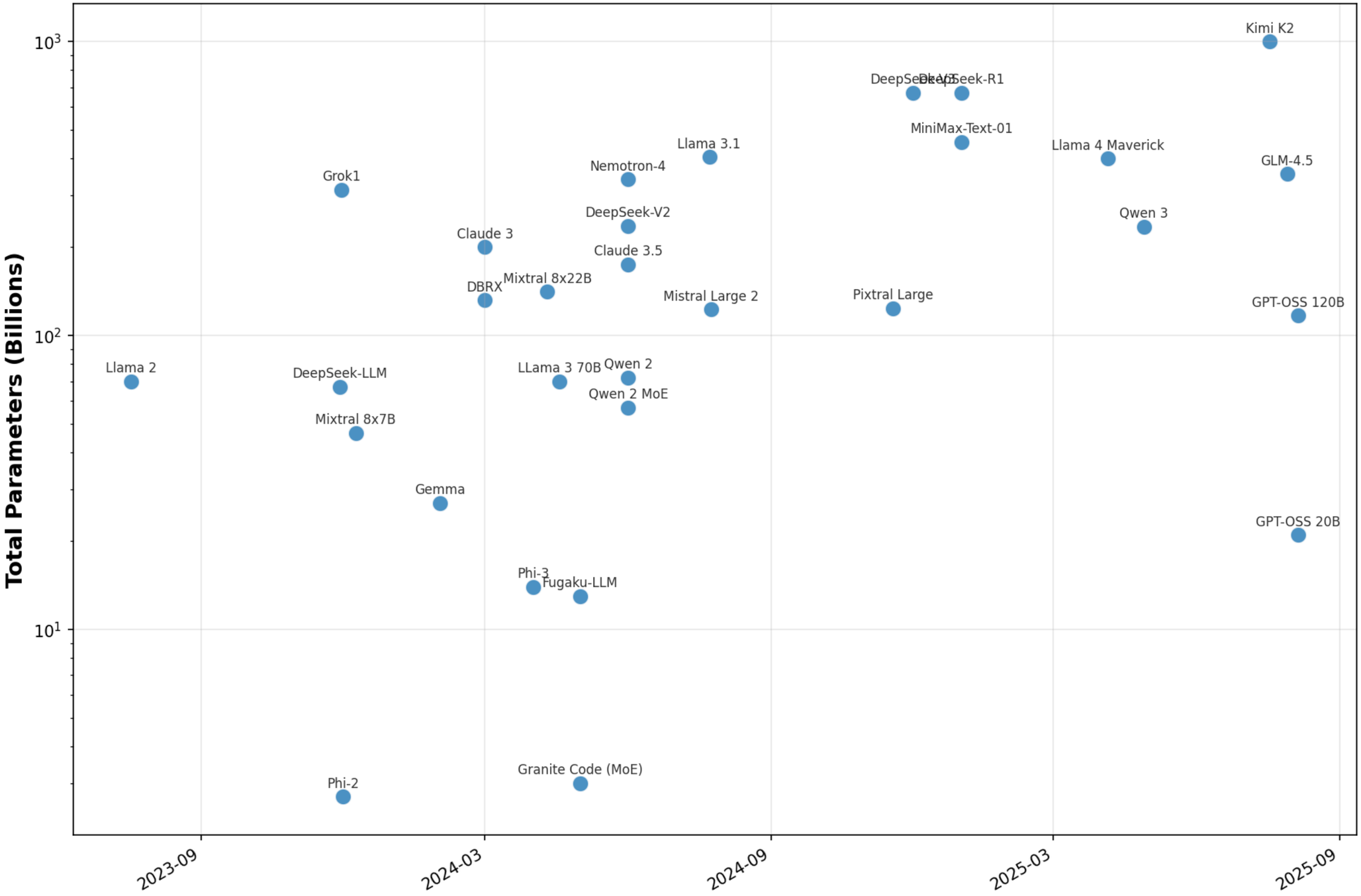
Reducing Barriers,
unleashing curiosity.

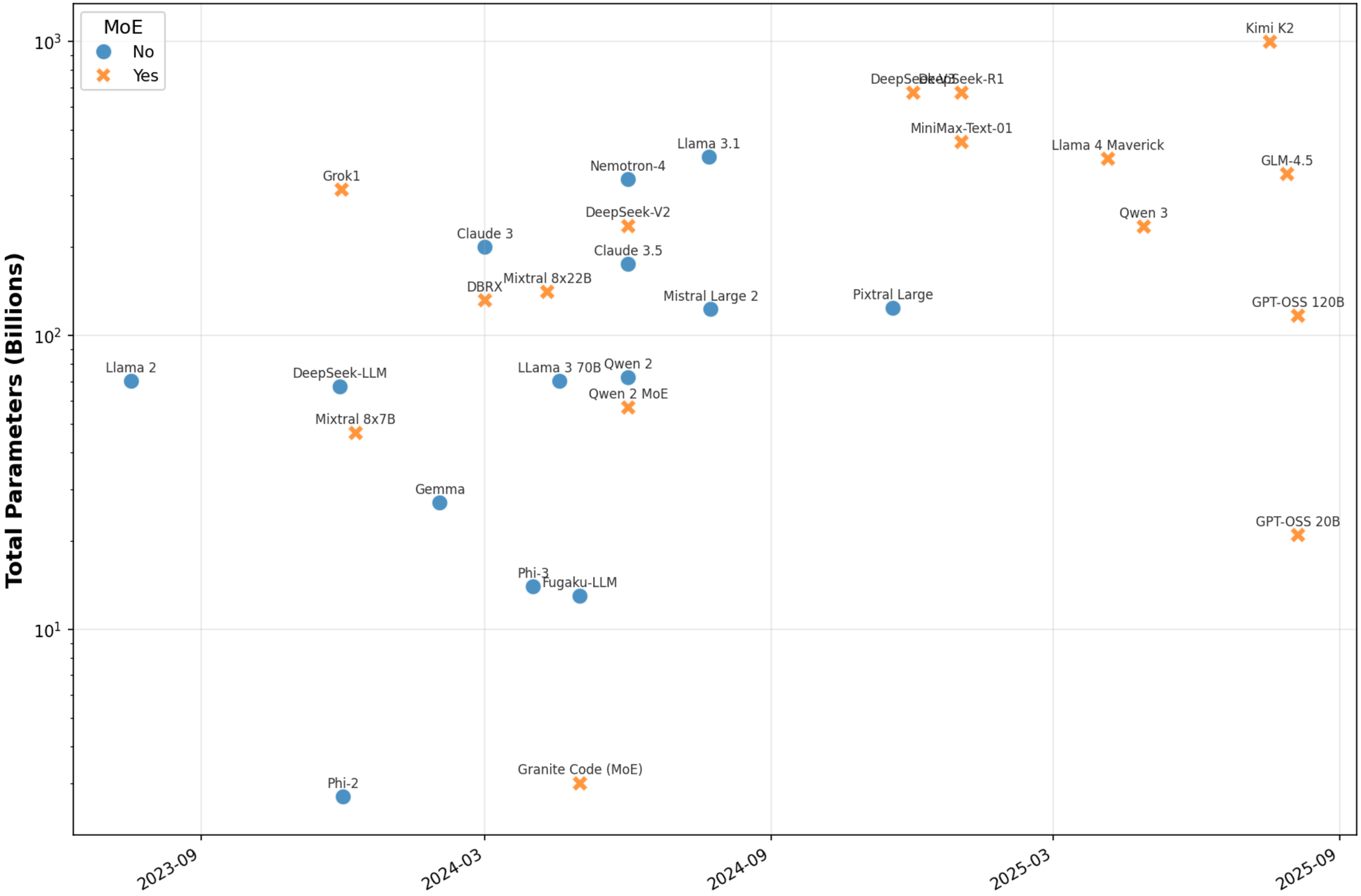


The guide

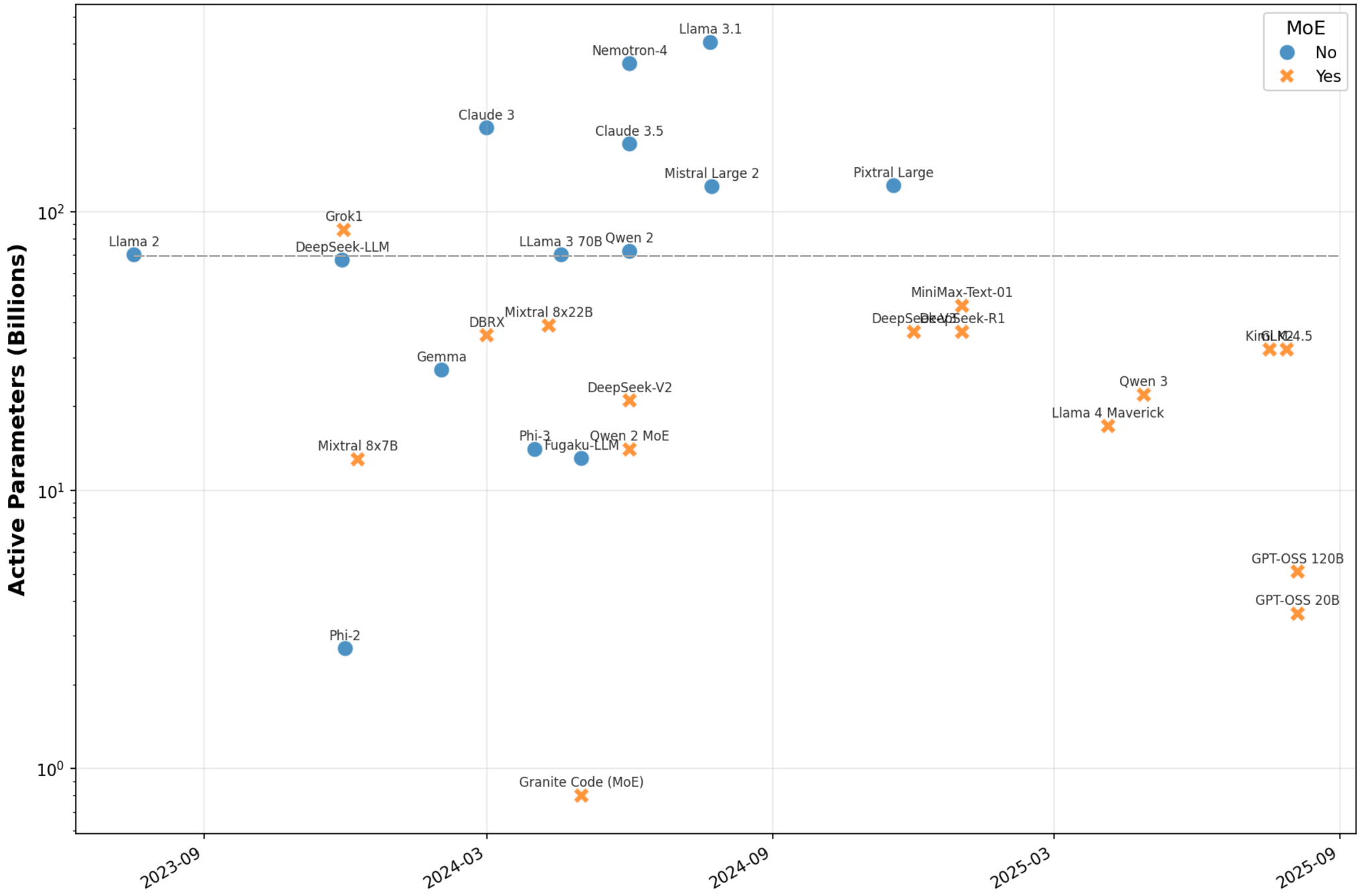
your visitors can have a conversation with.

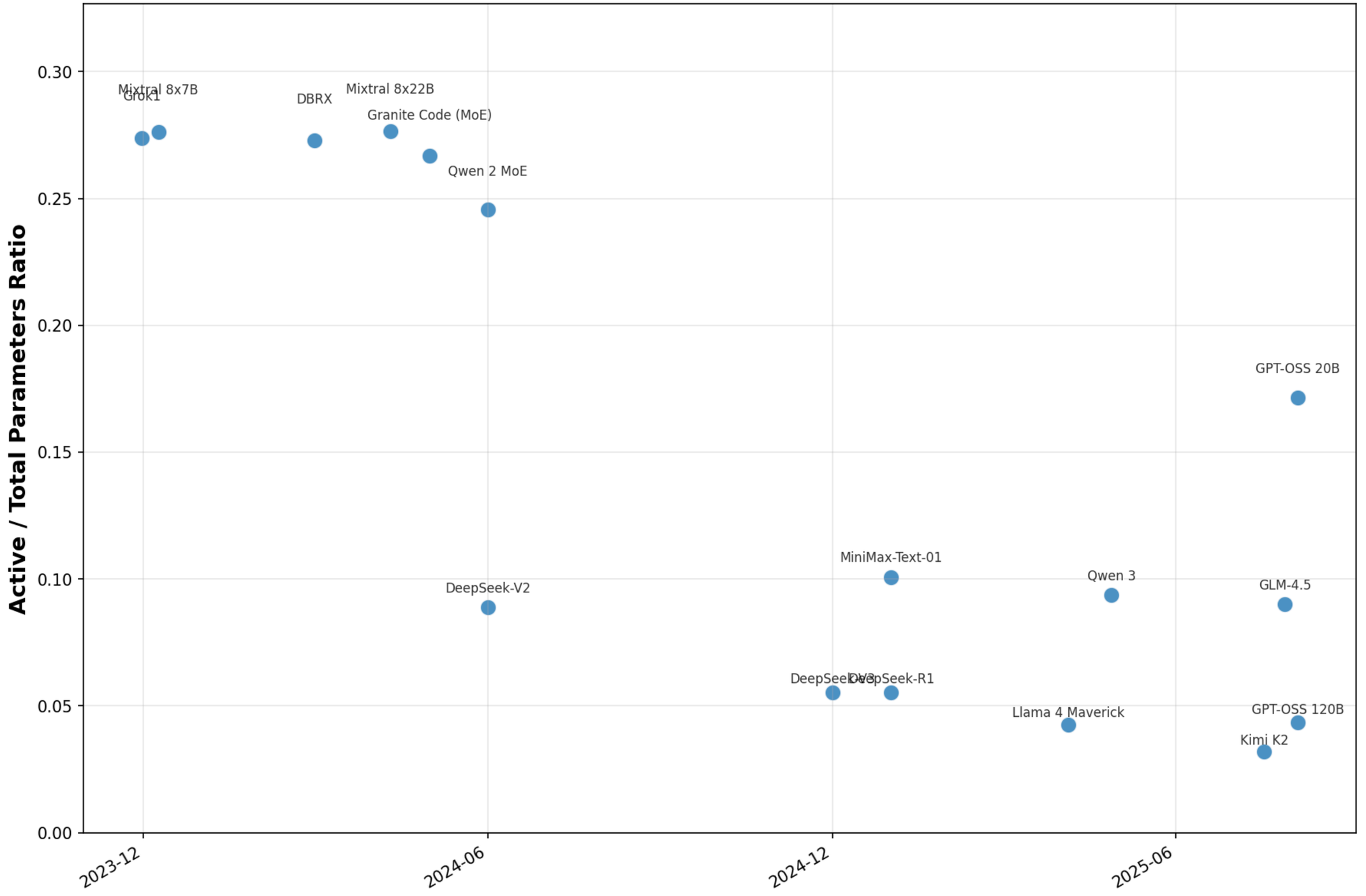
www.nuseum.ai


Total Parameters (Billions)
10^3
10^2
10^1
2023-09
2024-03
2024-09
2025-03
2025-09
Kimi K2
DeepSeek-R1
MiniMax-Text-01
Llama 4 Maverick
GLM-4.5
Llama 3.1
Nemotron-4
Grok1
DeepSeek-V2
Qwen 3
Claude 3
Claude 3.5
Mixtral 8x22B
DBRX
Mistral Large 2
Pixtral Large
GPT-OSS 120B
Llama 2
DeepSeek-LLM
LLama 3 70B
Qwen 2
Qwen 2 MoE
Mixtral 8x7B
Gemma
GPT-OSS 20B
Phi-3
Fugaku-LLM
Granite Code (MoE)
Phi-2


MoE
No
Yes
Total Parameters (Billions)
10^3
10^2
10^1
Kimi K2
DeepSeek-V3
DeepSeek-R1
MiniMax-Text-01
Llama 4 Maverick
GLM-4.5
Llama 3.1
Nemotron-4
Grok1
DeepSeek-V2
Qwen 3
Claude 3
Claude 3.5
Mixtral 8x22B
DBRX
Mistral Large 2
Pixtral Large
GPT-OSS 120B
Llama 2
DeepSeek-LLM
LLama 3 70B
Qwen 2
Qwen 2 MoE
Mixtral 8x7B
Gemma
GPT-OSS 20B
Phi-3
Fugaku-LLM
Granite Code (MoE)
Phi-2
2023-09
2024-03
2024-09
2025-03
2025-09

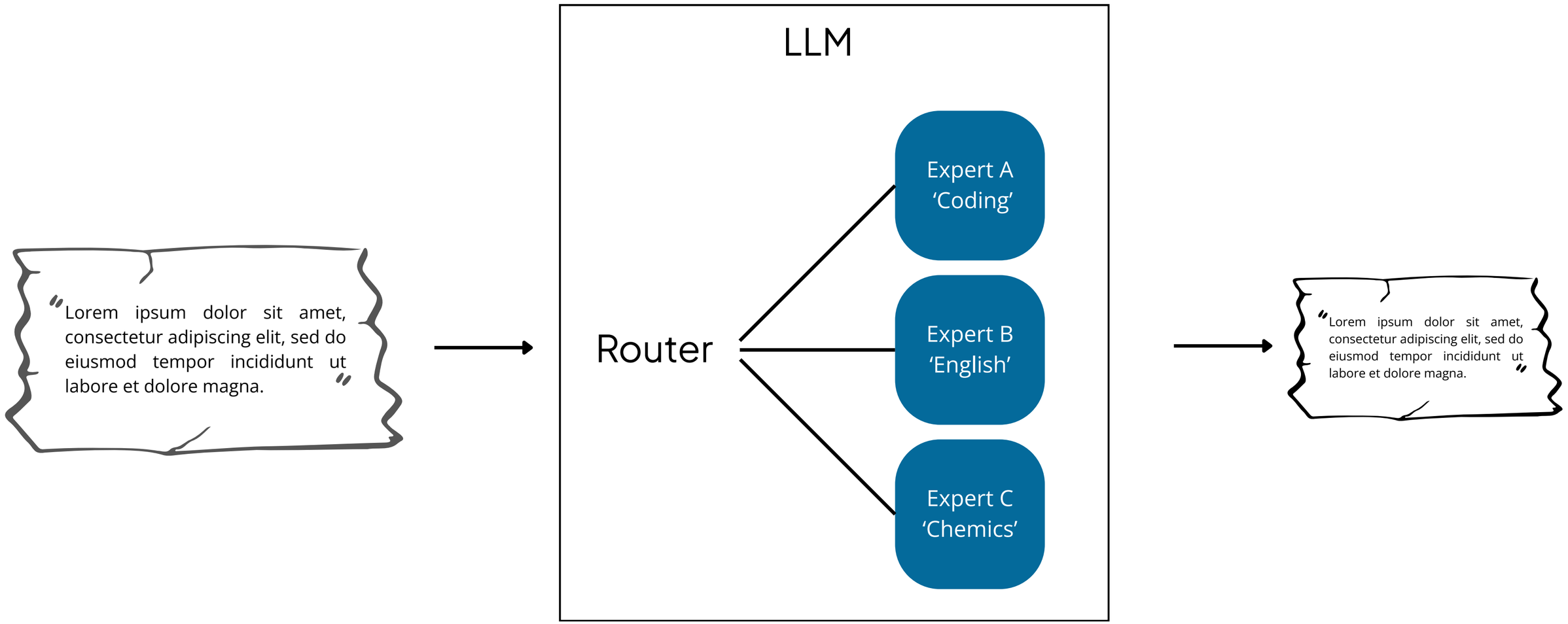
Why MoEs?


MoE
No
Yes
Active Parameters (Billions)
10^2
10^1
10^0
2023-09
2024-03
2024-09
2025-03
2025-09
Llama 3.1
Nemotron-4
Claude 3
Claude 3.5
Mistral Large 2
Pixtral Large
Grok1
Llama 2
DeepSeek-LLM
LLama 3 70B
Qwen 2
MiniMax-Text-01
DBRX
Mixtral 8x22B
DeepSeek-R1
Gemma
DeepSeek-V2
Qwen 3
Llama 4 Maverick
Phi-3
Qwen 2 MoE
Mixtral 8x7B
Fugaku-LLM
GPT-OSS 120B
GPT-OSS 20B
Phi-2
Granite Code (MoE)

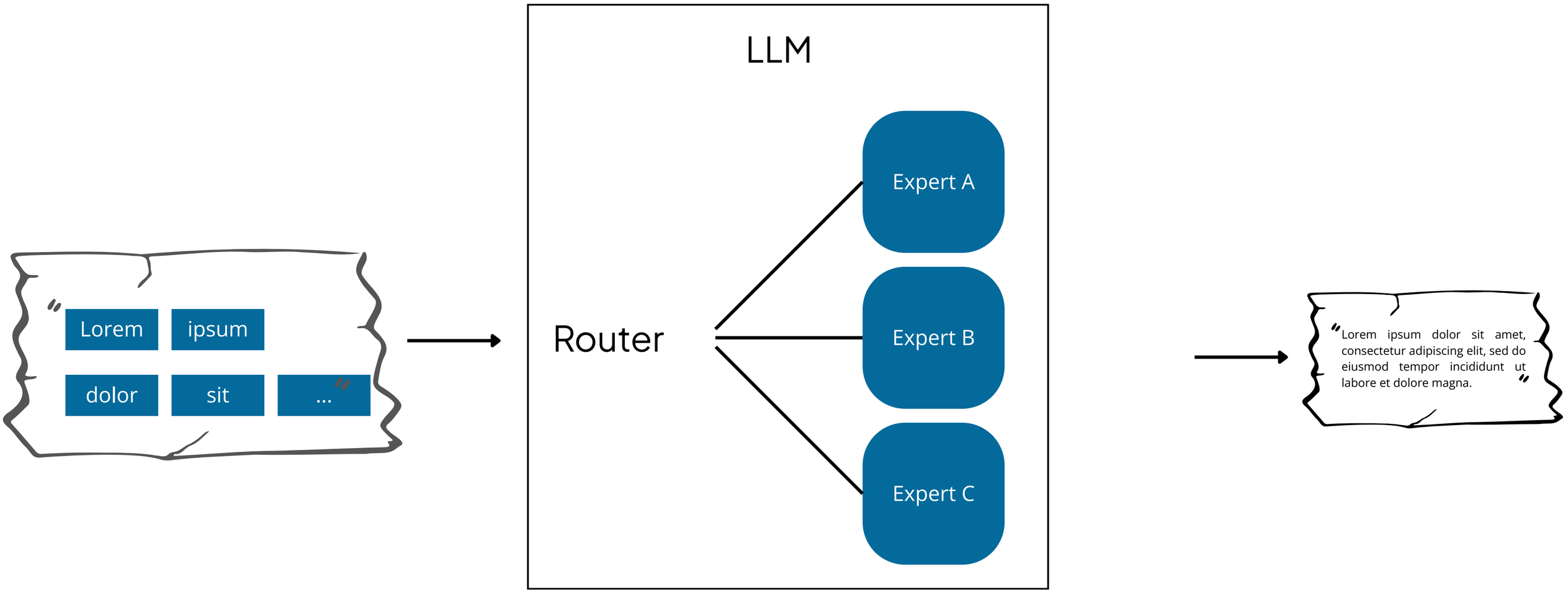

Active / Total Parameters Ratio
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2023-12
2024-06
2024-12
2025-06
Mixtral 8x7B
Grok1
DBRX
Mixtral 8x22B
Granite Code (MoE)
Qwen 2 MoE
DeepSeek-V2
MiniMax-Text-01
Qwen 3
GLM-4.5
DeepSeek-V3
DeepSeek-R1
Llama 4 Maverick
GPT-OSS 120B
Kimi K2
GPT-OSS 20B

How does it work?

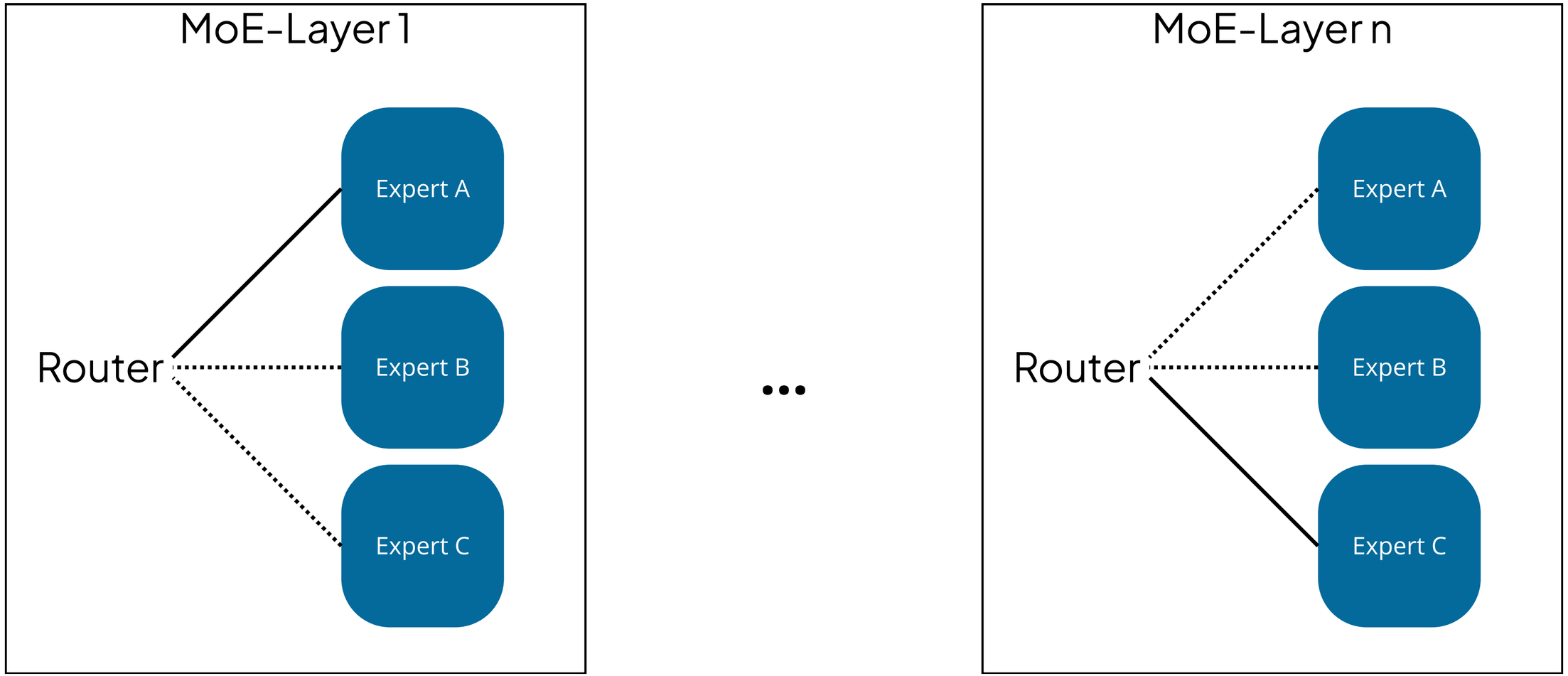
The naive Concept



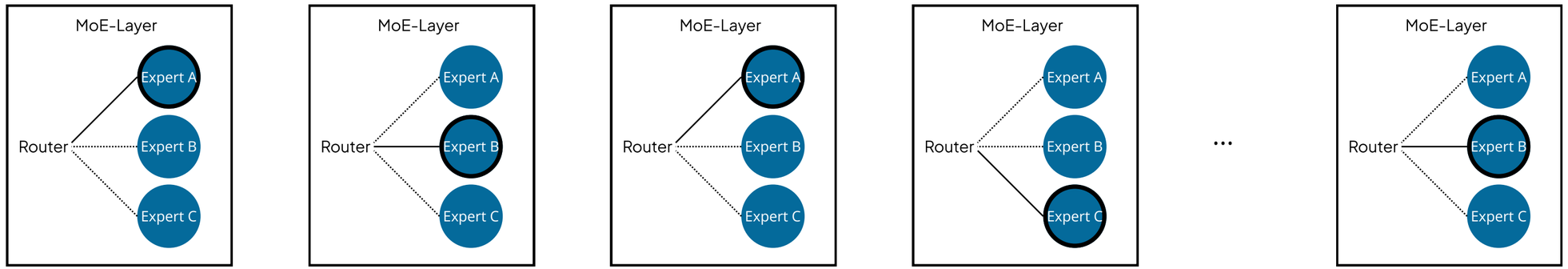
#1 Token-Level Routing

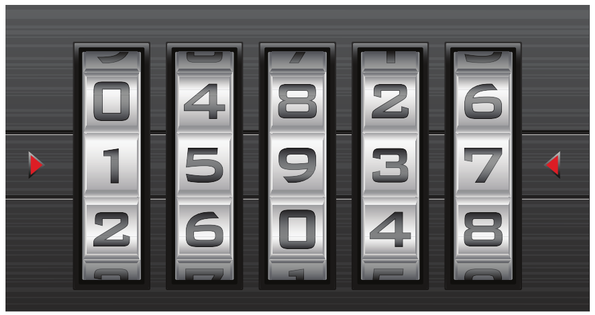


#2 There are many MoE-Layers



#2 There are many MoE-Layers





gpt-oss-120b

128 experts per layer
36 MoE layers

→ 128^{36} paths

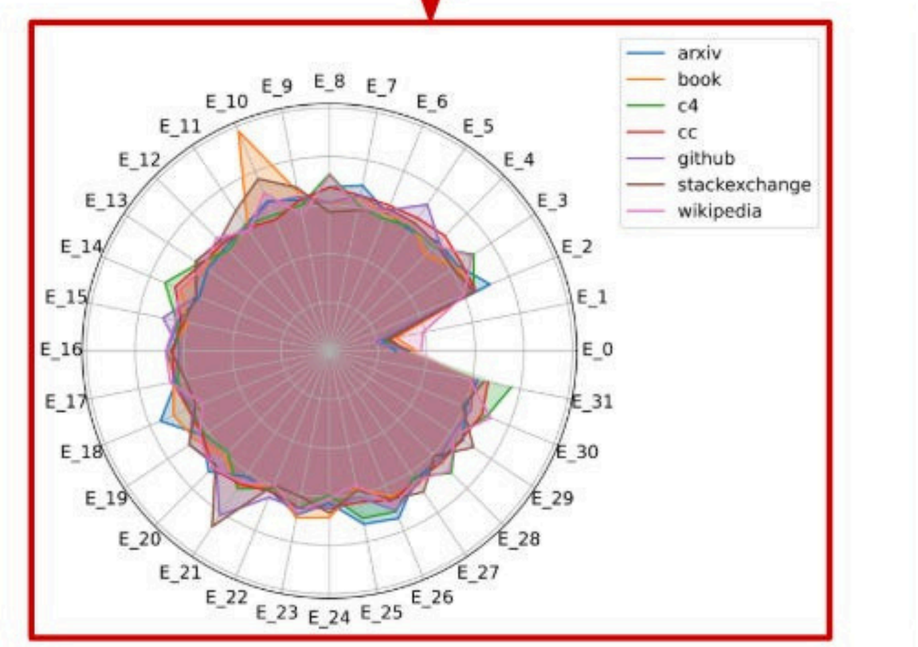
OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models

Fuzhao Xue[1†] Zian Zheng[1] Yao Fu[2] Jinjie Ni[1] Zangwei Zheng[1]
Wangchunshu Zhou[3] Yang You[1]

[1]National University of Singapore
[2]University of Edinburgh
[3]ETH Zurich

#3 There are no domain expert

No indications of expert specialization across domains of data



https://arxiv.org/pdf/2402.01739

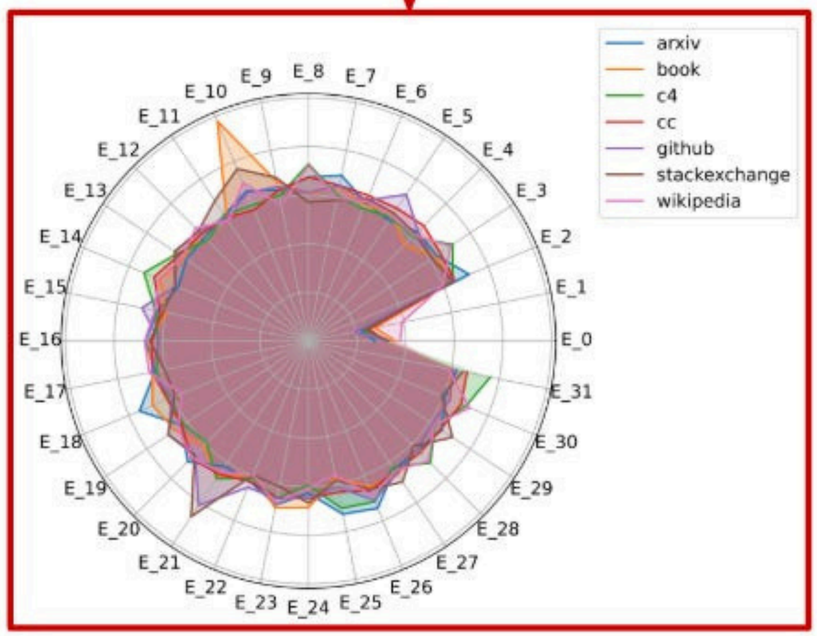
OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models

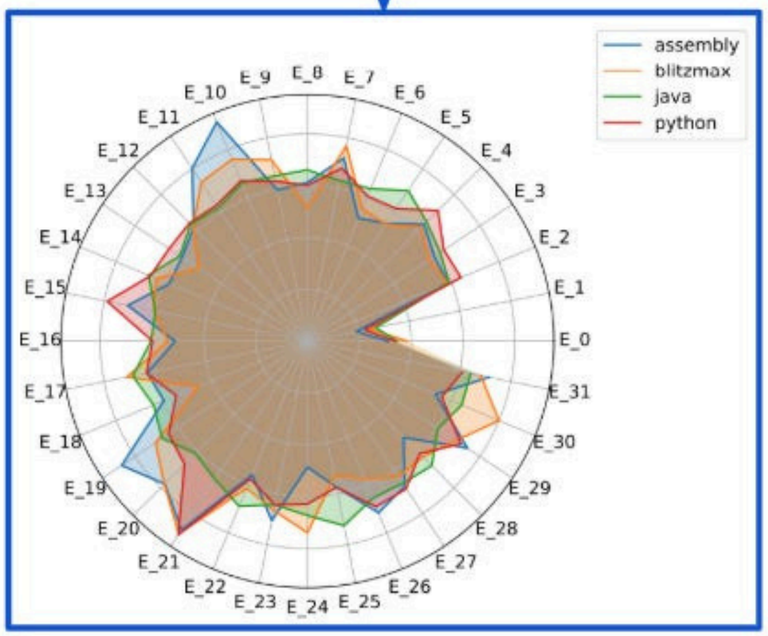
Fuzhao Xue[1†] Zian Zheng[1] Yao Fu[2] Jinjie Ni[1] Zangwei Zheng[1]
Wangchunshu Zhou[3] Yang You[1]

[1]National University of Singapore
[2]University of Edinburgh
[3]ETH Zurich

#3 There are no domain expert

No indications of expert specialization across domains of data or coding languages.





https://arxiv.org/pdf/2402.01739

#3 There are no domain expert

OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models

Fuzhao Xue[1†] Zian Zheng[1] Yao Fu[2] Jinjie Ni[1] Zangwei Zheng[1]
Wangchunshu Zhou[3] Yang You[1]

[1]National University of Singapore
[2]University of Edinburgh
[3]ETH Zurich

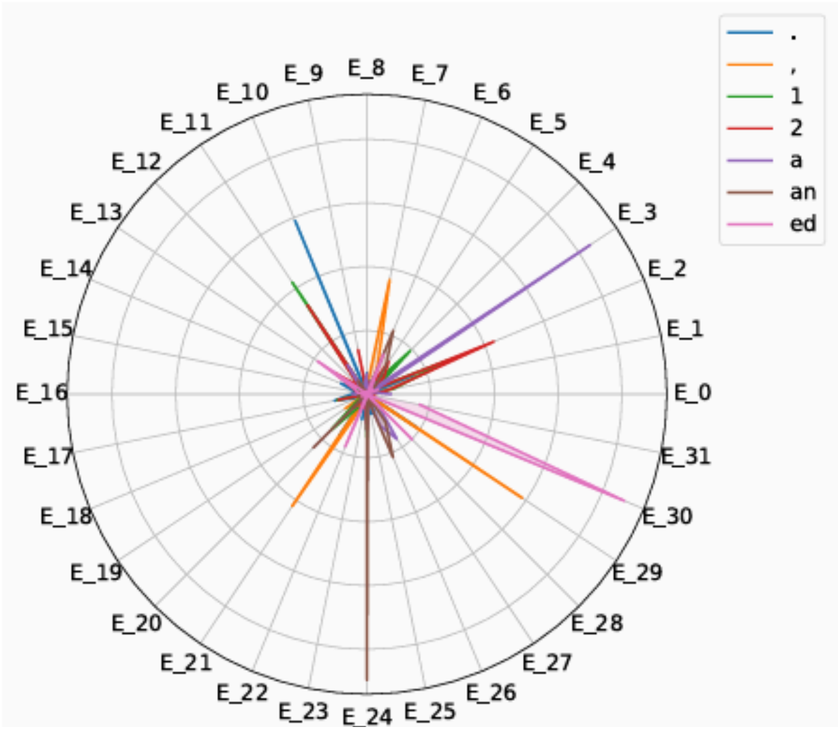


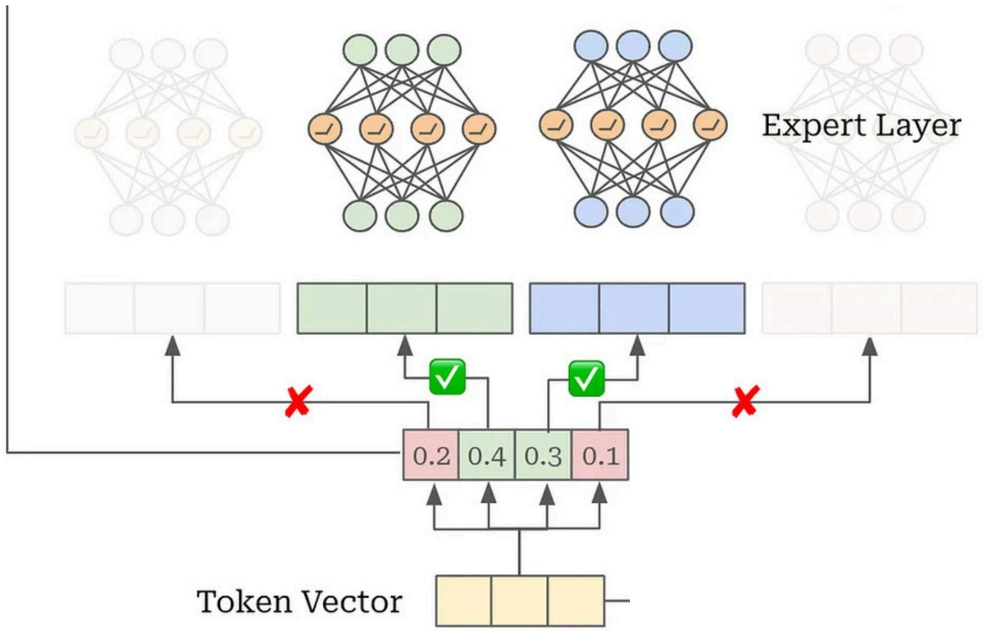
Table 9: Top Tokens selected by each expert.

Expert ID	Top Tokens
0	\n , ‘ , ’ , s , - , $, y , _ , , 2
1	\n , 1 , , 2 , \\ , S , . , - , C , {
21	, , and , , . , \n , = , \t , the , , n
30	} , ed , d , have , ing , , , has , s , " , had
31	to , can , s , of , ing , will , not , e , ed , would

https://arxiv.org/pdf/2402.01739

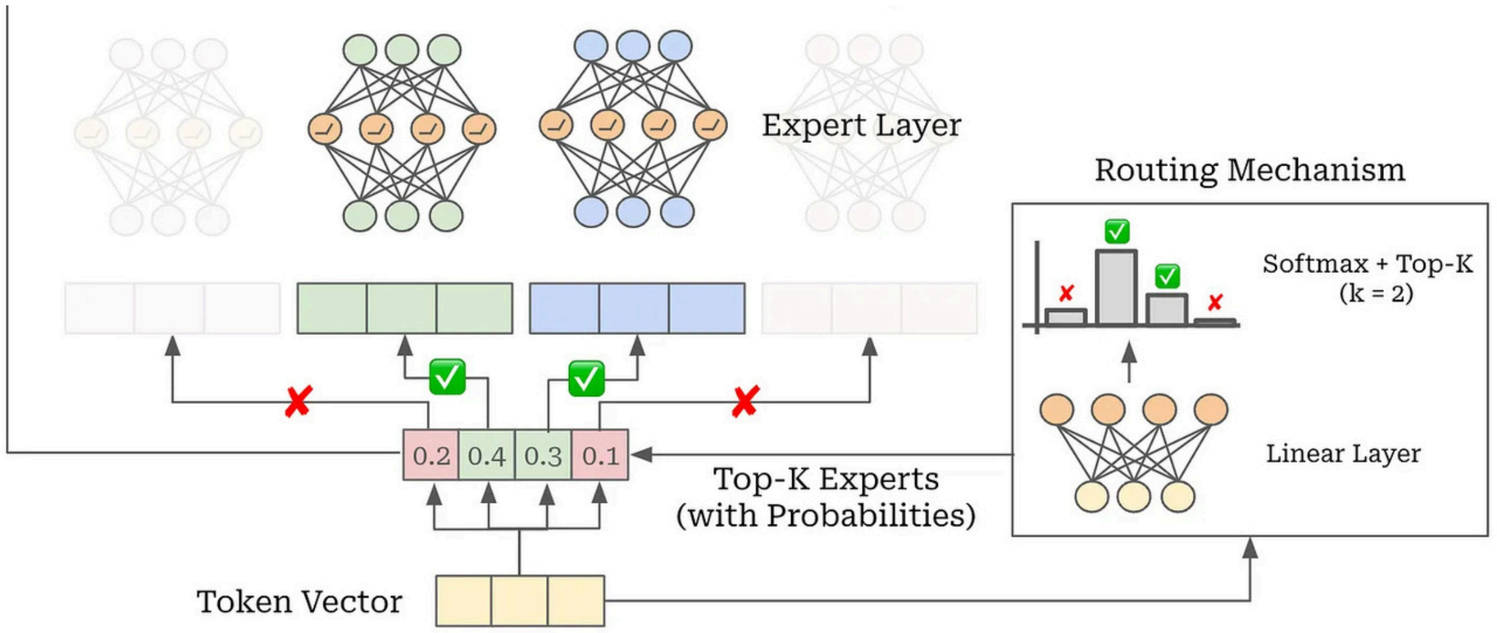
Let’s understand how its trained

MoE Layer



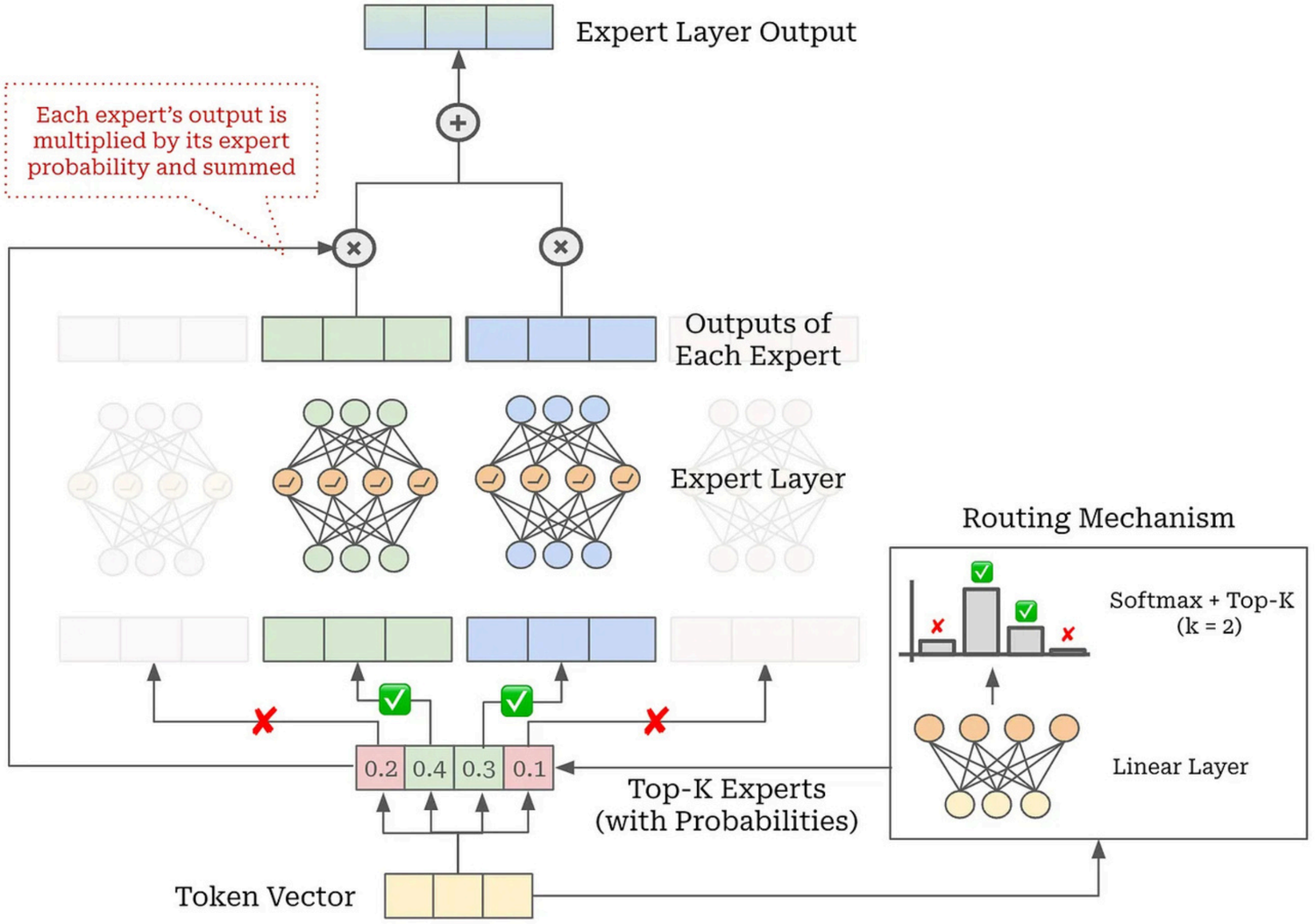
https://cameronrwolfe.substack.com/p/nano-moe

MoE Layer

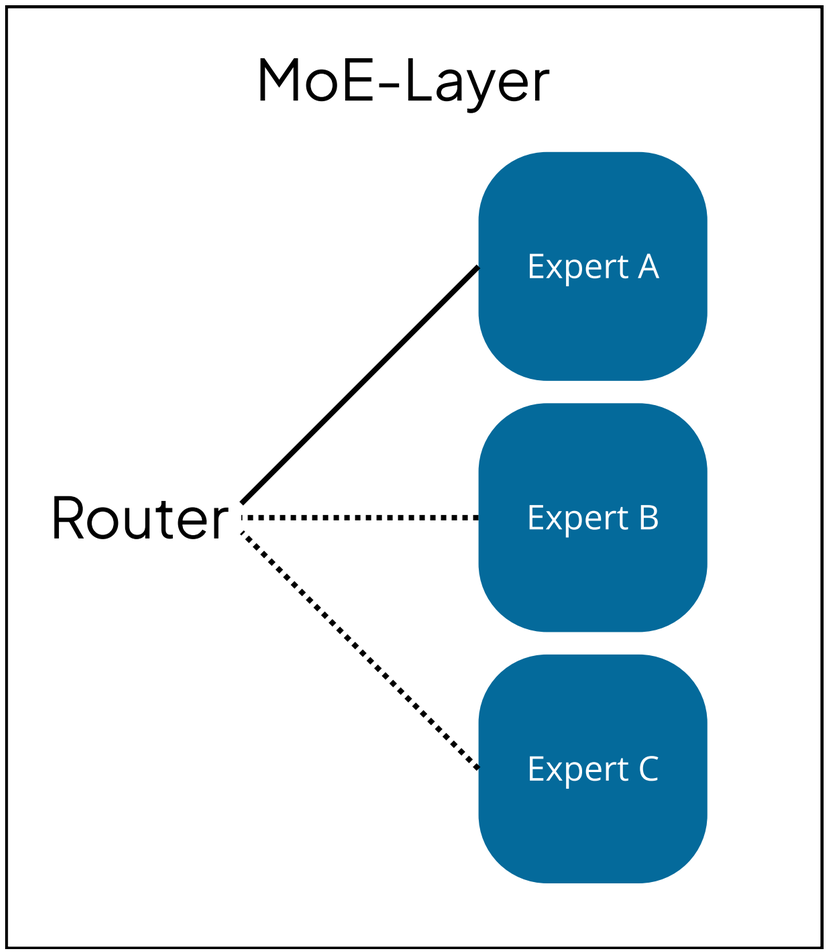


https://cameronrwolfe.substack.com/p/nano-moe

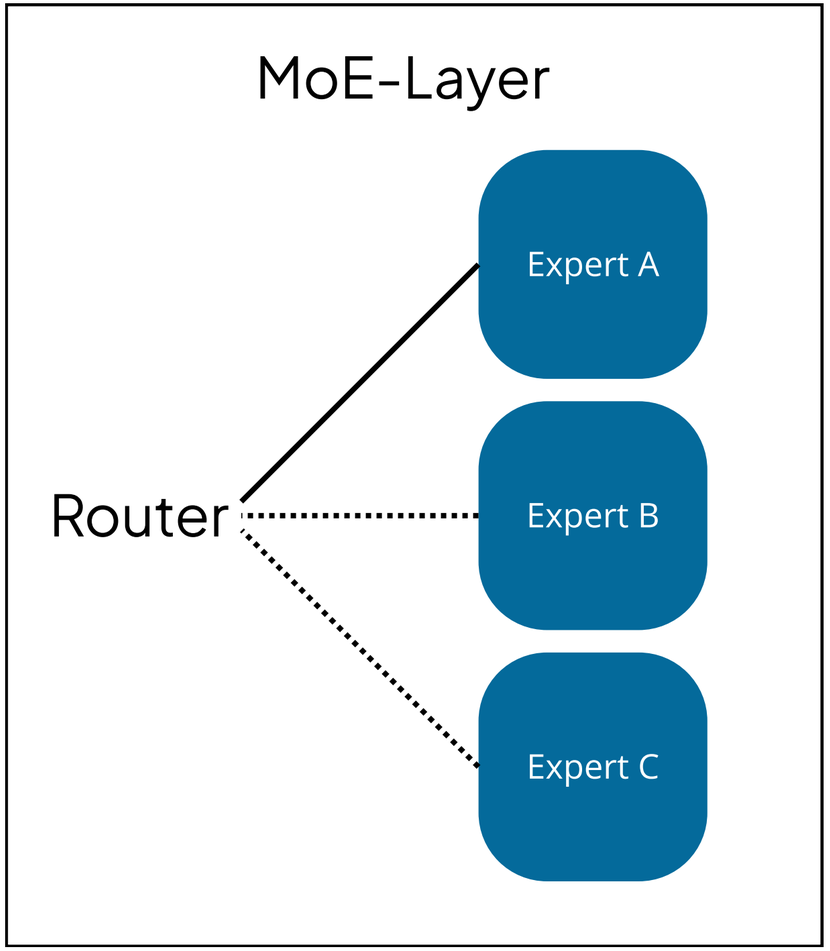
MoE Layer



https://cameronrwolfe.substack.com/p/nano-moe


MoE-Layer
Router
Expert A
Expert B
Expert C

Routing Collapse



Ensures load balancing loss stays fixed with an increasing number of experts

Constant scaling factor for the load balancing loss

$$L_{\texttt{load-balance}}(\mathcal{X}) = \overbrace{w_{\texttt{load-balance}}} \cdot N \cdot \sum_{i=1}^{N} f_i \, P_i$$

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

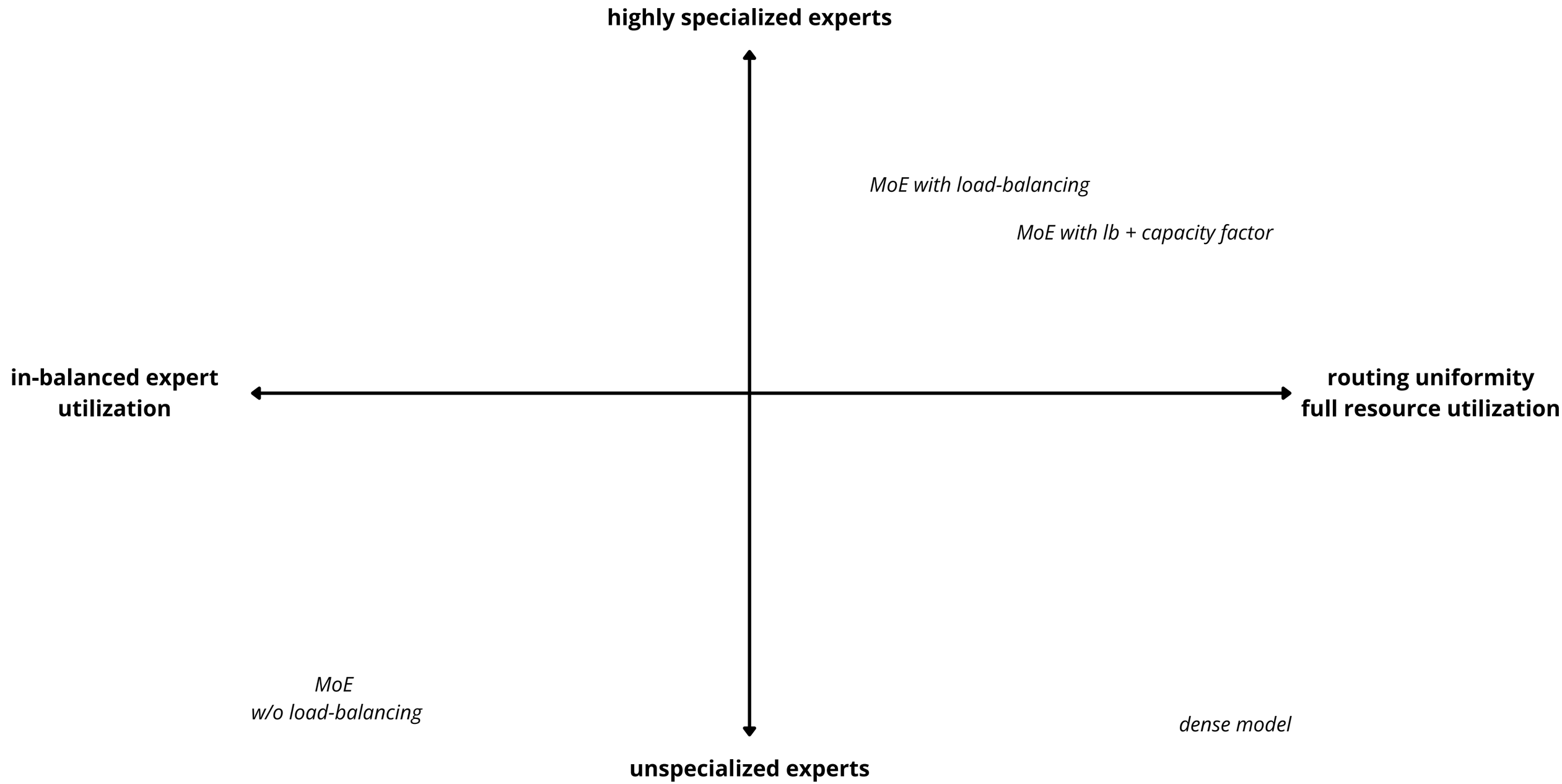
Fraction of tokens sent to expert i (not differentiable)

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

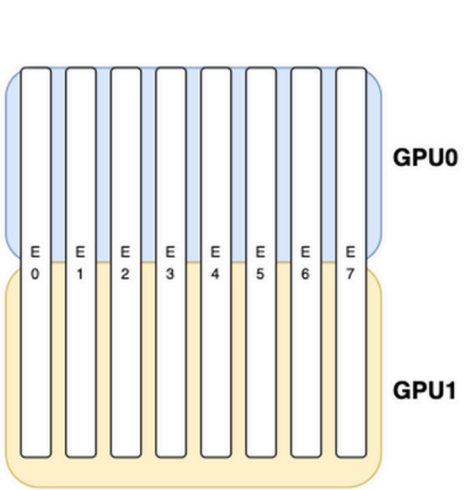
Fraction of probability allocated to expert i (is differentiable)

Fedu et al. Switch Transformers

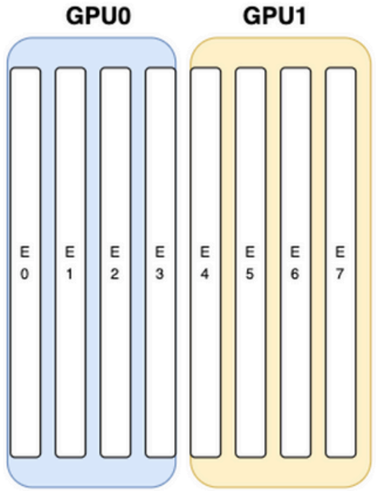
Mixture of Experts trade-off



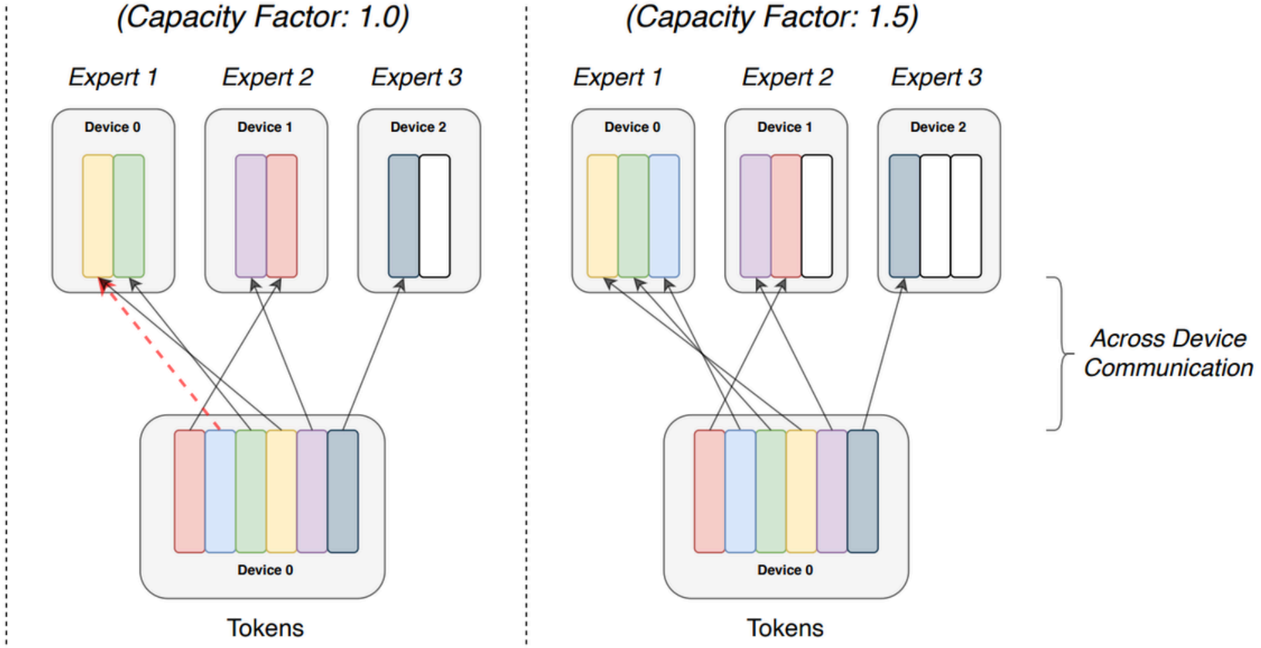
Expert Capacity



Tensor Parallel



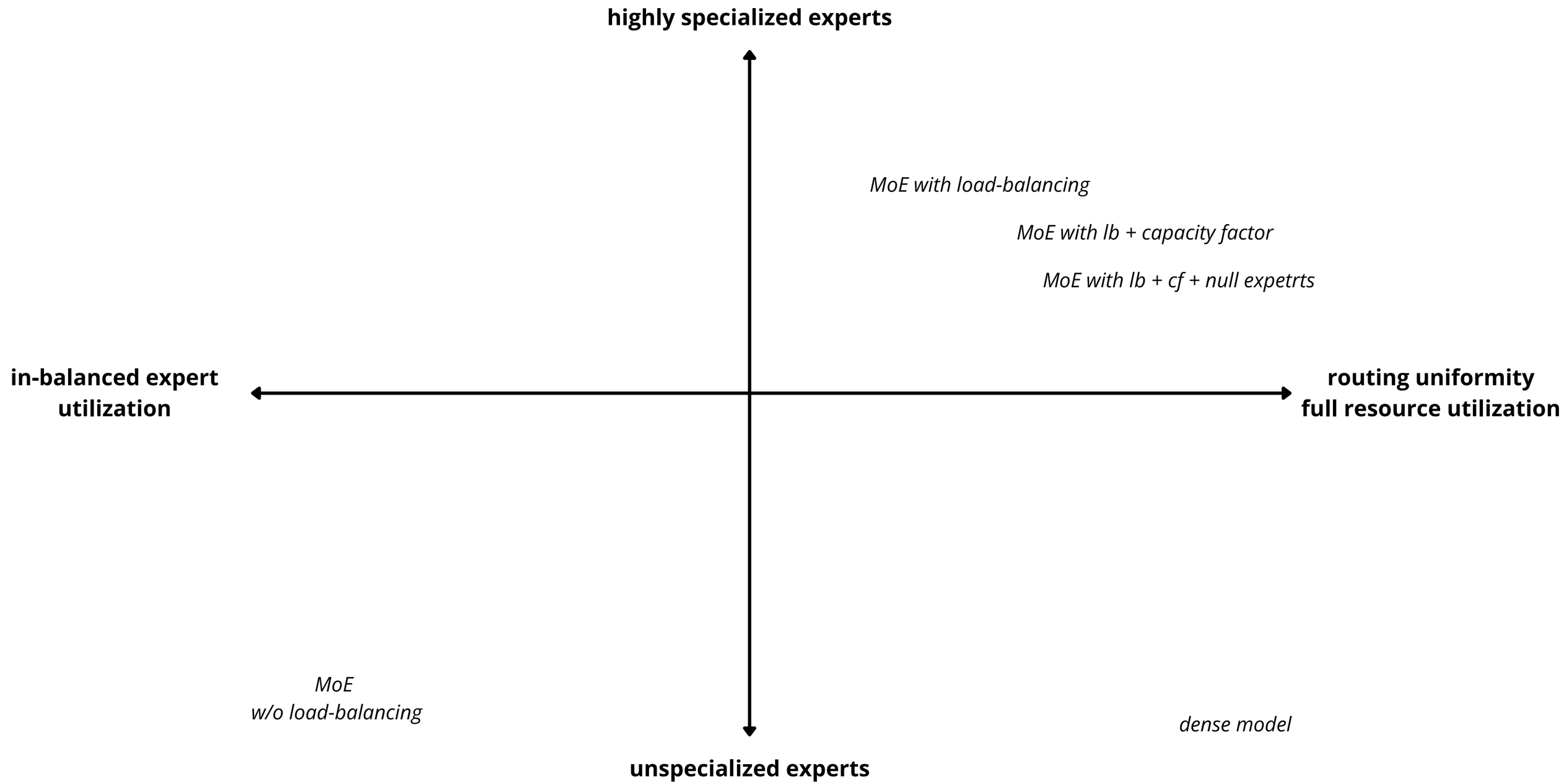
Expert Parallel



https://nvidia.github.io/TensorRT-LLM/advanced/expert-parallelism.html

Fedu et al. Switch Transformers

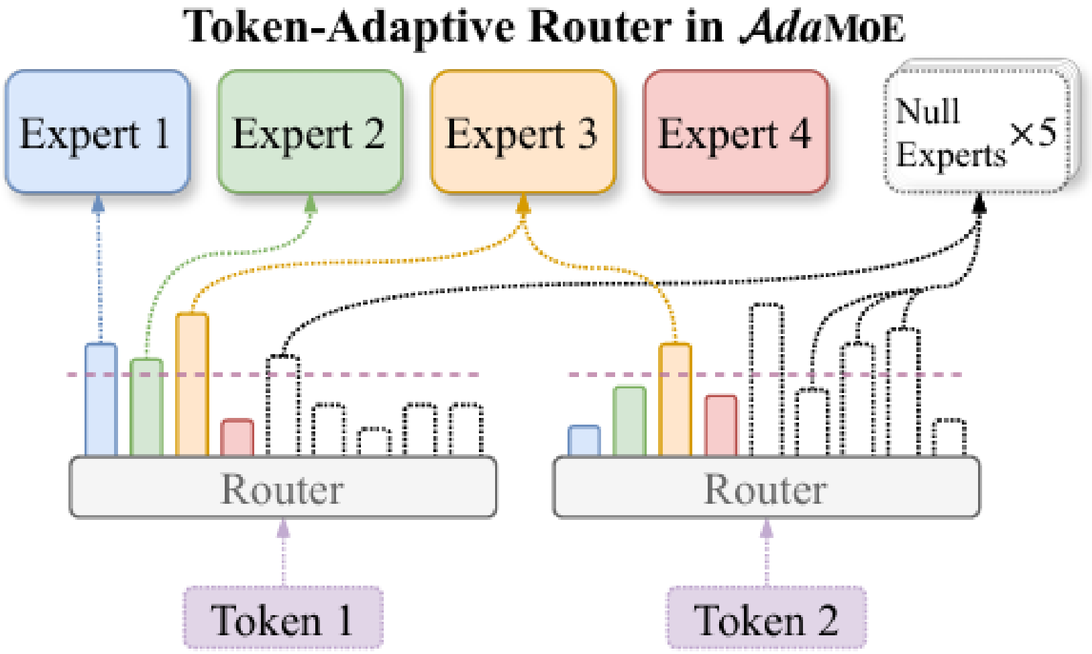
Mixture of Experts trade-off



AdaMoE: Token-Adaptive Routing with Null Experts for Mixture-of-Experts Language Models

Zihao Zeng[1*], **Yibo Miao**[1*], **Hongcheng Gao**[2], **Hao Zhang**[3], **Zhijie Deng**[1†]
[1]Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University
[2]University of Chinese Academy of Sciences [3]University of California, San Diego
{zengzihao, miaoyibo, zhijied}@sjtu.edu.cn
gaohongcheng23@mails.ucas.ac.cn, haozhang@ucsd.edu

Published as a conference paper at ICLR 2025



MoE++: Accelerating Mixture-of-Experts Methods with Zero-Computation Experts

Peng Jin[1,2], **Bo Zhu**[3], **Li Yuan**[1,2,4] ✉, **Shuicheng Yan**[3,5] ✉
[1]Pengcheng Laboratory
[2]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
[3]Kunlun 2050 Research & Skywork AI
[4]Rabbitpre Intelligence [5]National University of Singapore
jp21@stu.pku.edu.cn, yuanli-ece@pku.edu.cn
Code: https://github.com/SkyworkAI/MoE-plus-plus

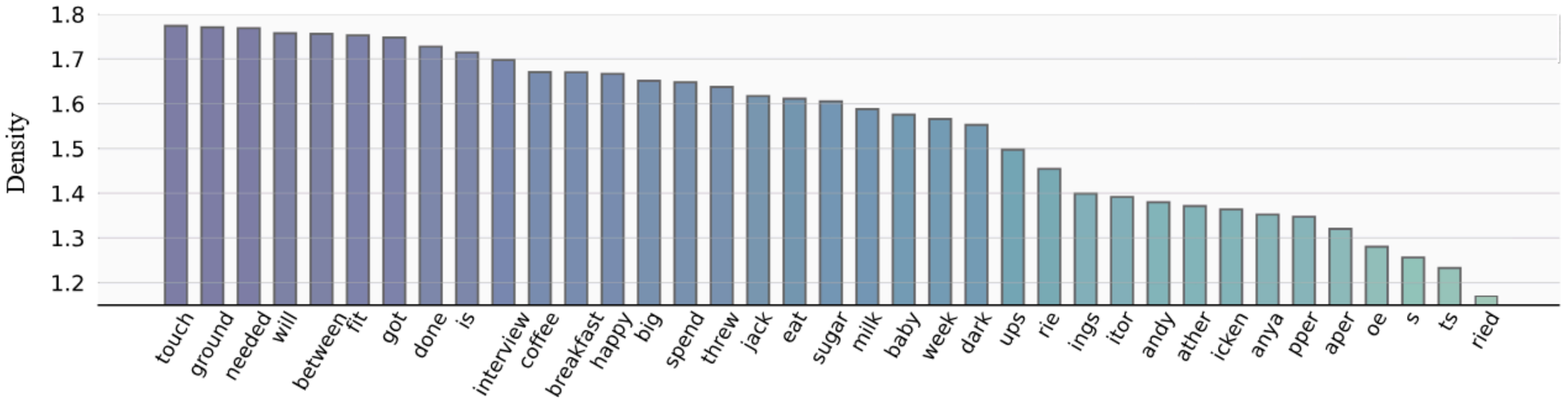
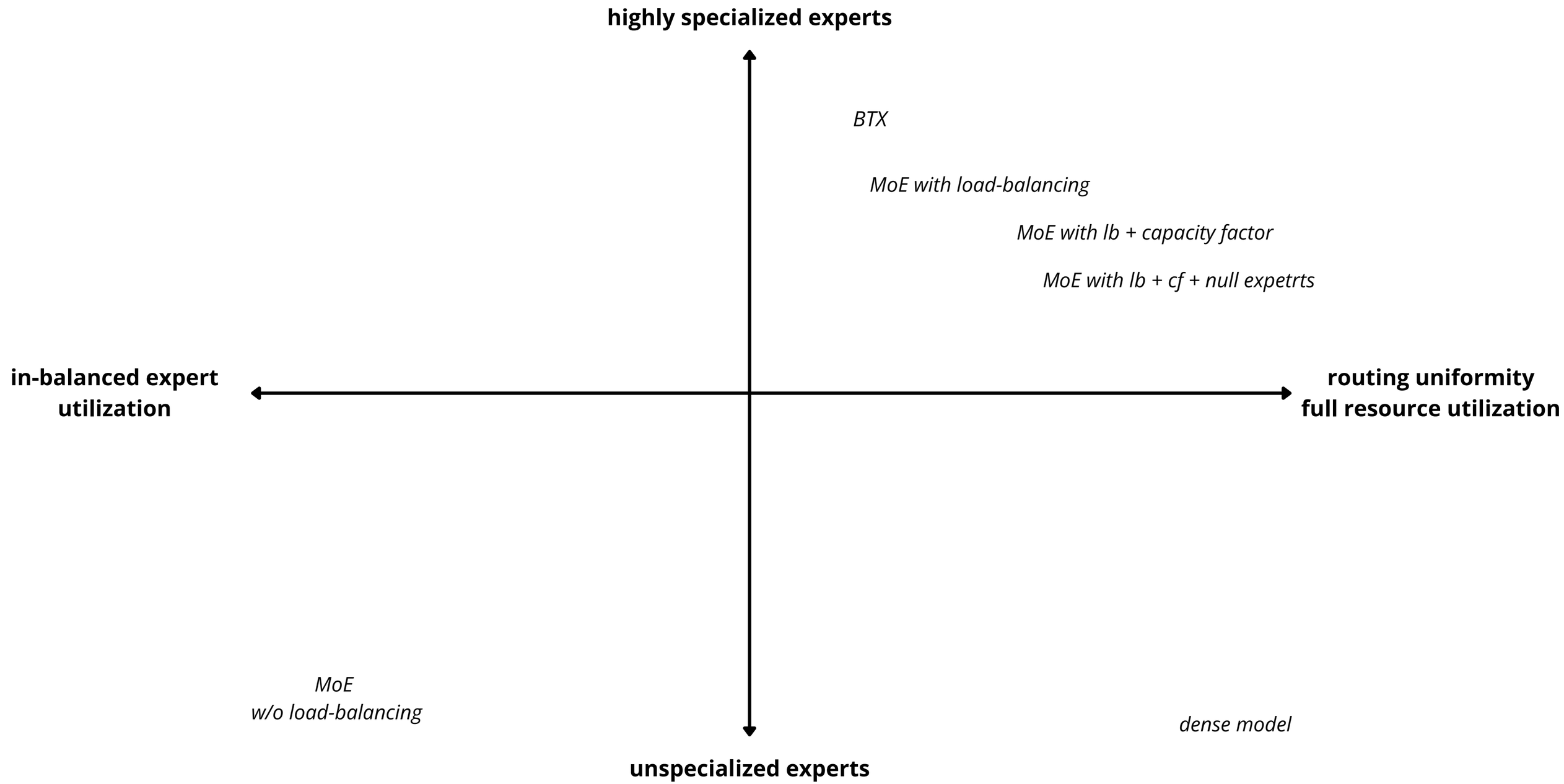


Figure 5: **The visualization of the number of FFN experts activated per token at the token level.**

Mixture of Experts trade-off



Branch-Train-MiX: Mixing Expert LLMs into a Mixture-of-Experts LLM

Sainbayar Sukhbaatar, **Olga Golovneva**, **Vasu Sharma**, **Hu Xu**, **Xi Victoria Lin**, **Baptiste Rozière**, **Jacob Kahn**, **Daniel Li**, **Wen-tau Yih**, **Jason Weston**, **Xian Li**

FAIR at Meta

We investigate efficient methods for training Large Language Models (LLMs) to possess capabilities in multiple specialized domains, such as coding, math reasoning and world knowledge. Our method, named Branch-Train-MiX (BTX), starts from a seed model, which is branched to train experts in embarrassingly parallel fashion with high throughput and reduced communication cost. After individual experts are asynchronously trained, BTX brings together their feedforward parameters as experts in Mixture-of-Expert (MoE) layers and averages the remaining parameters, followed by an MoE-finetuning stage to learn token-level routing. BTX generalizes two special cases, the Branch-Train-Merge metho
sparse upcycling, which omi
approaches, BTX achieves t

Date: March 13, 2024
Correspondence: {sainbar,xian

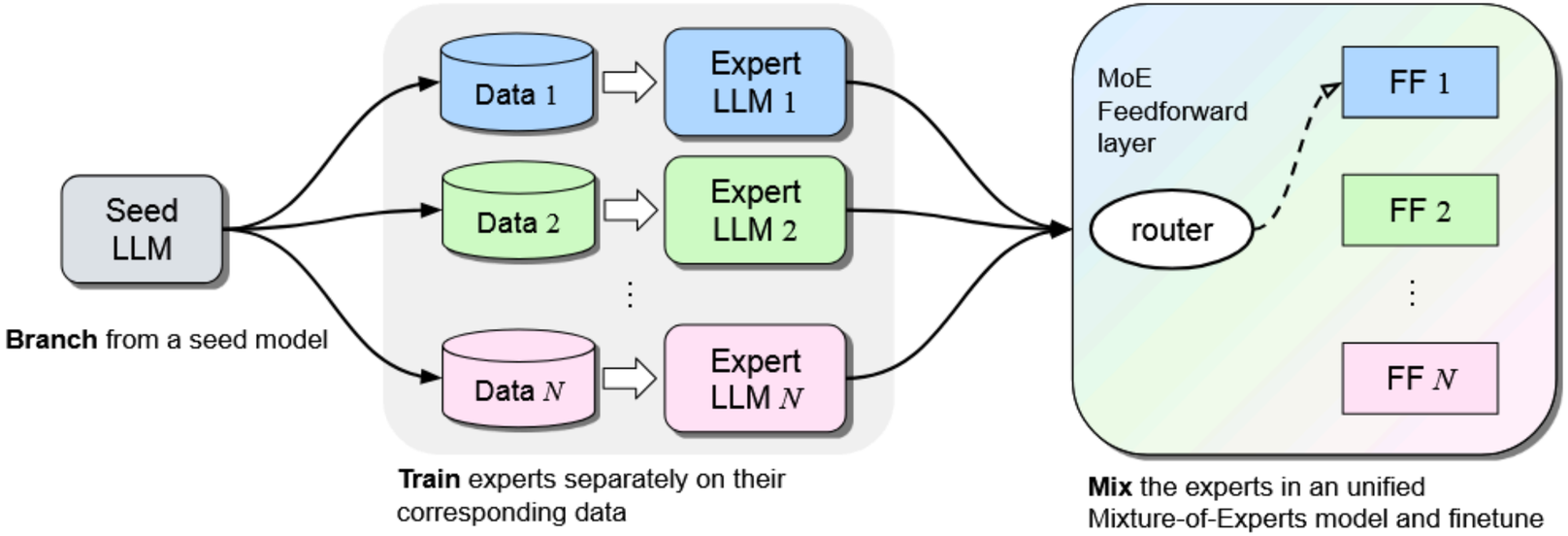


Figure 1 The Branch-Train-MiX (BTX) method has three steps: **1) branch** from a pretrained seed LLM by making multiple copies of it; **2) train** those copies separately on different subsets of data to obtain expert LLMs; **3) mix** those expert LLMs by combining them into a single LLM using mixture-of-experts feedforward (FF) layers, and finetuning the overall unified model.

Self-Selecting Experts

Self-Selecting Experts

Autonomy-of-Experts Models

Ang Lv [1] Ruobing Xie [2] Yining Qian [3] Songhao Wu [1] Xingwu Sun [2 4] Zhanhui Kang [2] Di Wang [2] Rui Yan [1 5 6]

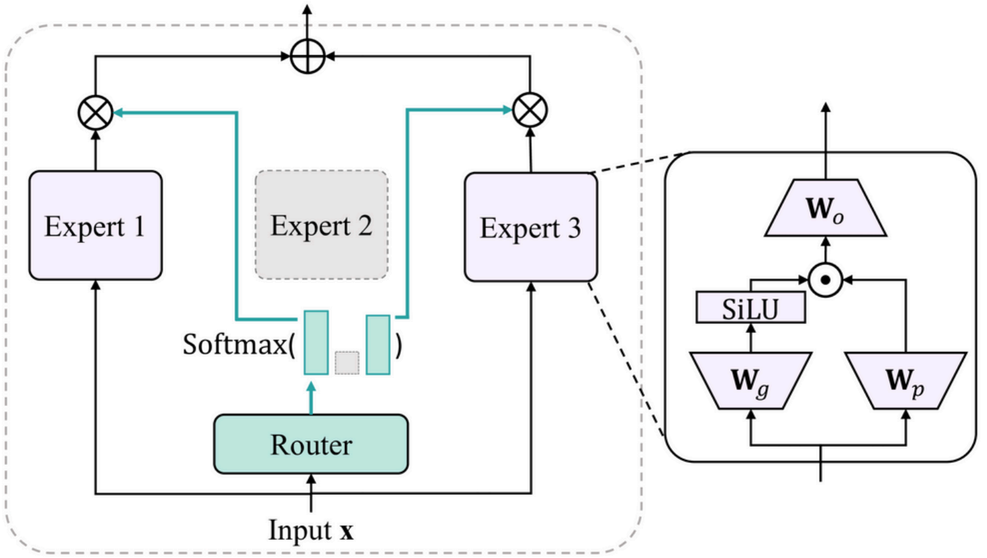
Mixture-of-Experts with Expert Choice Routing

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon

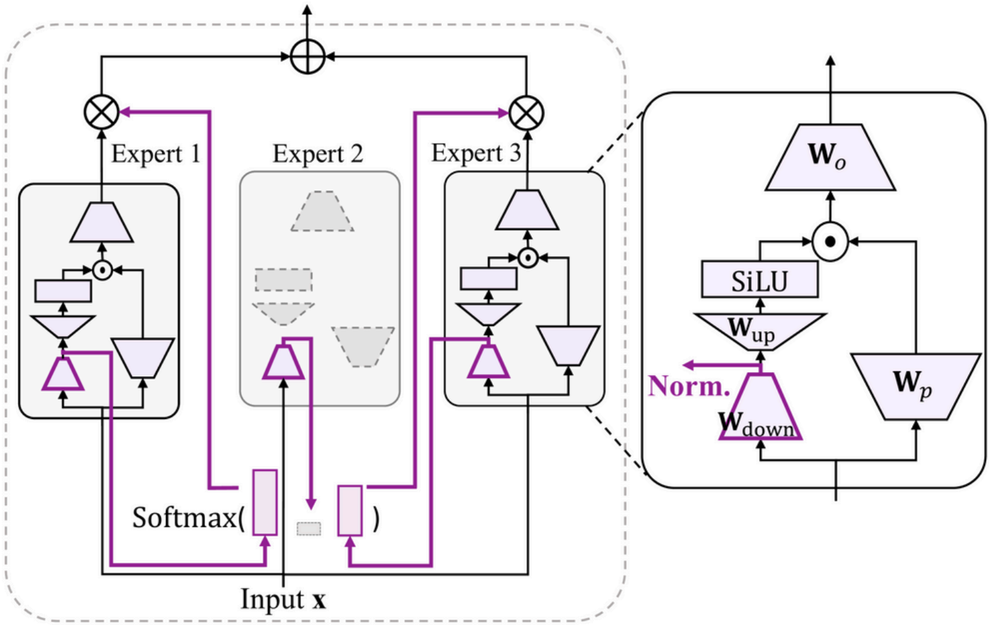
Google, Mountain View, CA, USA
{yanqiz, taole, hanxiaol, dunan, huangyp, vzhao, adai, zhifengc, qvl, jlaudon}@google.com

SELF-MOE: TOWARDS COMPOSITIONAL LARGE LANGUAGE MODELS WITH SELF-SPECIALIZED EXPERTS

Junmo Kang* Georgia Tech
Leonid Karlinsky MIT-IBM Watson AI Lab
Hongyin Luo MIT
Zhen Wang UCSD
Jacob Hansen MIT
James Glass MIT
David Cox MIT-IBM Watson AI Lab
Rameswar Panda MIT-IBM Watson AI Lab
Rogerio Feris MIT-IBM Watson AI Lab
Alan Ritter Georgia Tech



(a) Mixture-of-Experts

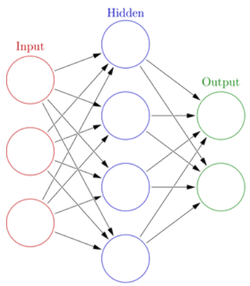


(b) Autonomy-of-Experts

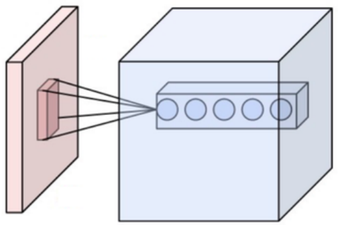
The Bigger Picture

Dense Net → System of Modular Sparse Nets

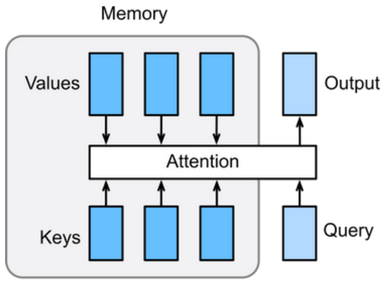
Fully Connected Networks



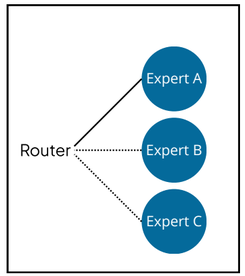
CNNs



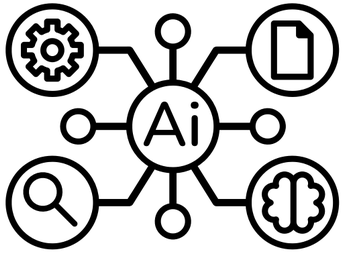
Transformer



MoE



System of Modular Experts



Thank you

Link To The Slides