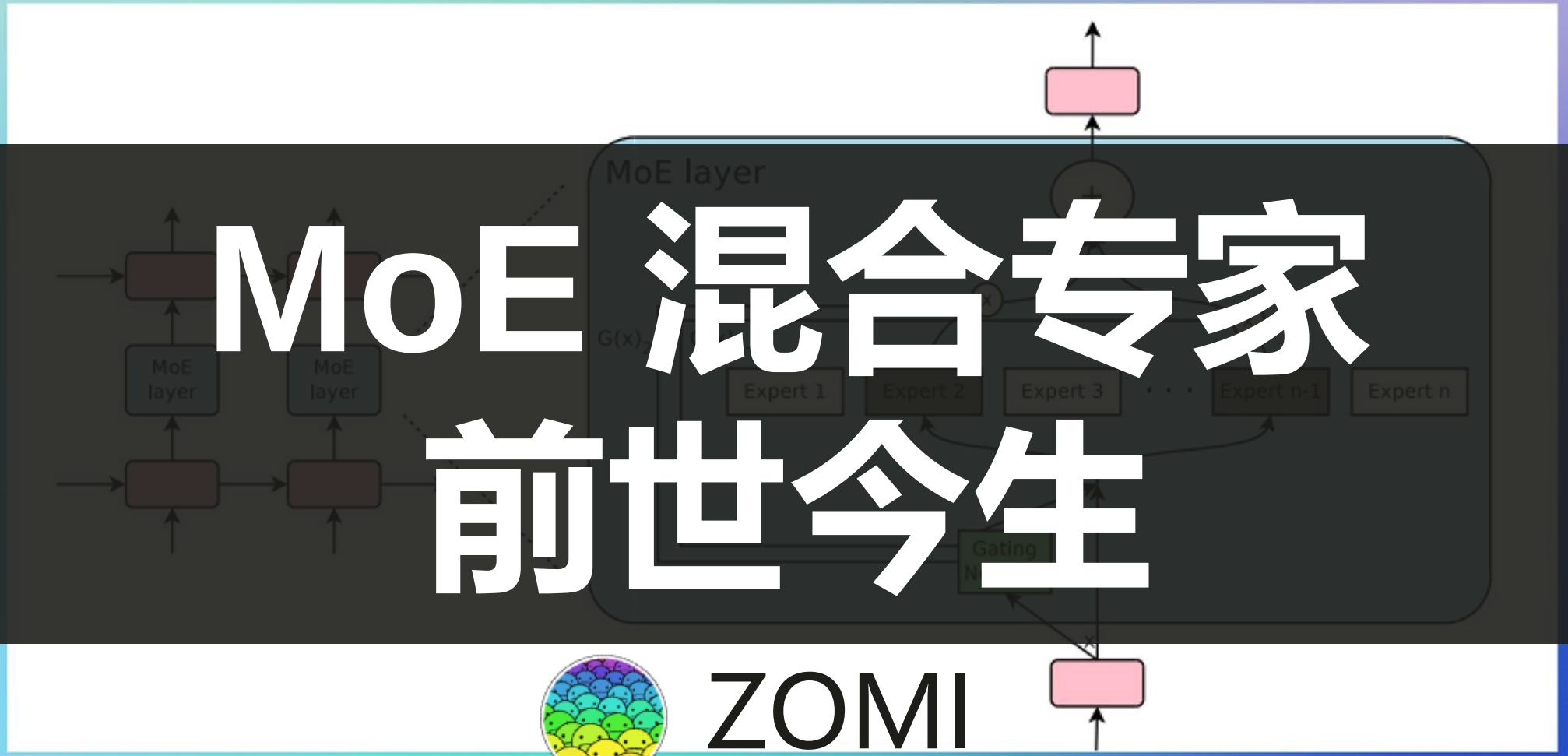
Mixture of Experts (MoE)



MoE 混合专家 前世今生

ZOMI

视频目录大纲

1. 综述：MOE 架构总览
2. 奠基工作：90 年代初期
3. 架构形成：RNN 时代
4. 提升效果：Transformer 时代
5. 智能涌现：GPT 时代
6. 可视化：Mixtral 7x8B MOE

视频目录大纲

- MoE 混合专家
  - MoE 基本介绍 4
  - MoE 前世今生
    - 90 年代初期（奠基工作）
    - RNN 时代（架构形成）
    - Transformer 时代（提升效果）
    - GPT 时代（智能涌现）
    - Mixtral 为例 MoE 可视化
  - MoE 核心工作 6
  - MoE 架构原理 3
  - 大模型遇上 MoE 4
  - 手撕 MoE 代码 5
  - MoE 的未来 2

01

MoE 简史

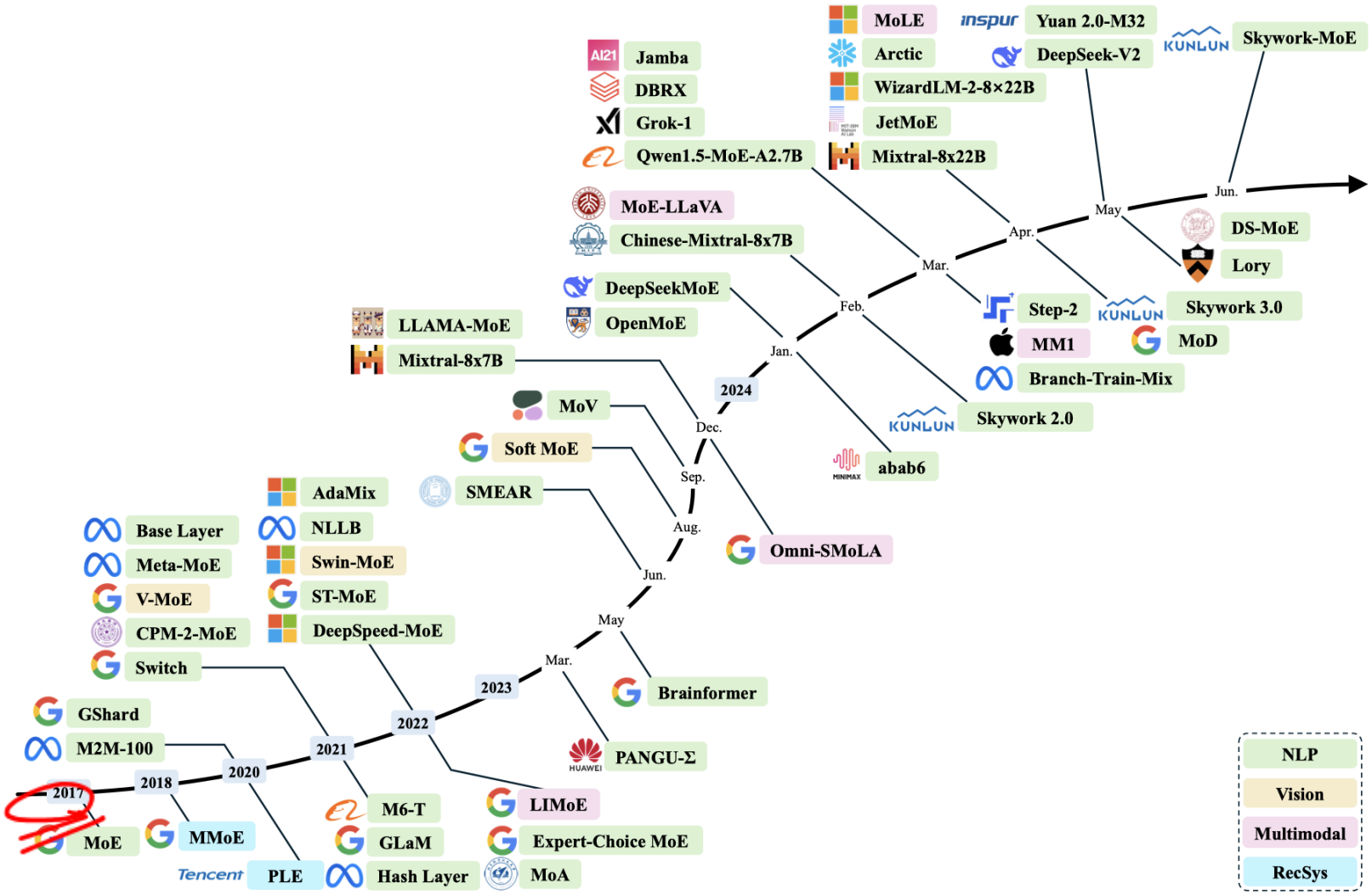
历史对 MoE 模型架构重要文献

文章名称	发布时间	主要贡献
Adaptive Mixtures of Local Experts	1991年	提出了局部 Expert 的概念，通过门控机制选择不同的 Expert 子模型处理不同输入区域，为MoE架构奠定了基础。
Hierarchical Mixtures of Experts and the EM Algorithm	1994年	提出了混合 Expert 模型的基本框架，结合了概率模型和神经网络的思想，使用期望最大化（EM）算法进行训练。这是MoE架构的奠基性工作
Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer	2017年	提出了稀疏门控混合 Expert 层，实现了大规模模型的高效推理。
GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding	2020年	首次将MOE技术引入Transformer架构，提供了高效的分布式并行计算架构。
GLaM: Efficient Scaling of Language Models with Mixture-of-Experts	2021年	利用MoE架构在语言模型中实现了高效的扩展，展示了其在自然语言处理任务中的潜力。
Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity	2021年	提出了稀疏激活的MOE架构，显著提升了模型的预训练速度和推理效率。
PaLM: Scaling Language Modeling with Pathways	2022年	提出了Pathways架构下的PaLM模型，结合了MoE技术，实现了高效的大规模语言模型训练。
Llama-MoE: Scaling Mixture-of-Experts Models for Open Pretraining	2023年	将MoE架构引入开源的Llama系列模型中，展示了MoE在开放预训练中的可行性
DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models	2024年	引入了“细粒度/垂类 Expert”和“共享 Expert”的概念，提升了 Expert 的专业性和模型效率。

近期发布的 MoE 大模型

模型	发布时间	备注
GPT4	2023年3月	23年6月George Hotz爆料GPT4是8×220B模型
Mistral-8×7B	2023年12月	Mistral AI，开源
LLAMA-MoE	2023年12月	Mate，开源
DeepSeek-MoE	2024年1月	幻方量化(深度求索)，国内首个开源 MoE 模型，有技术报告
Step-2	2024年3月	阶跃星辰，无开源，无细节发布
MM1	2024年3月	苹果，多模态MoE，无开源，有技术报告
Grok-1	2024年3月	XAI，开源
Qwen1.5-MoE-A2.7B	2024年3月	阿里巴巴，开源
DBRX	2024年3月	Databricks，开源
Mistral-8×22B	2024年4月	Mistral AI，开源
WizardLM-2-8×22B	2024年4月	微软，开源
Arctic	2024年4月	Snowflake，480B，Dense-MoE Hybrid，开源
Grok-2	2024年8月	XAI，开源
DeepSeek-V3	2025 年 1 月	幻方量化(深度求索)，国内首个开源 MoE 模型，有技术报告
MiniMax-01	2025 年 1 月	MiniMax 发布的MoE架构大模型，参数规模达4560亿，支持长达400万tokens的输入
Qwen2.5-Max	2025 年 1 月	采用超大规模MOE架构，预训练数据量超过20万亿tokens，支持高达100万token的上下文窗口

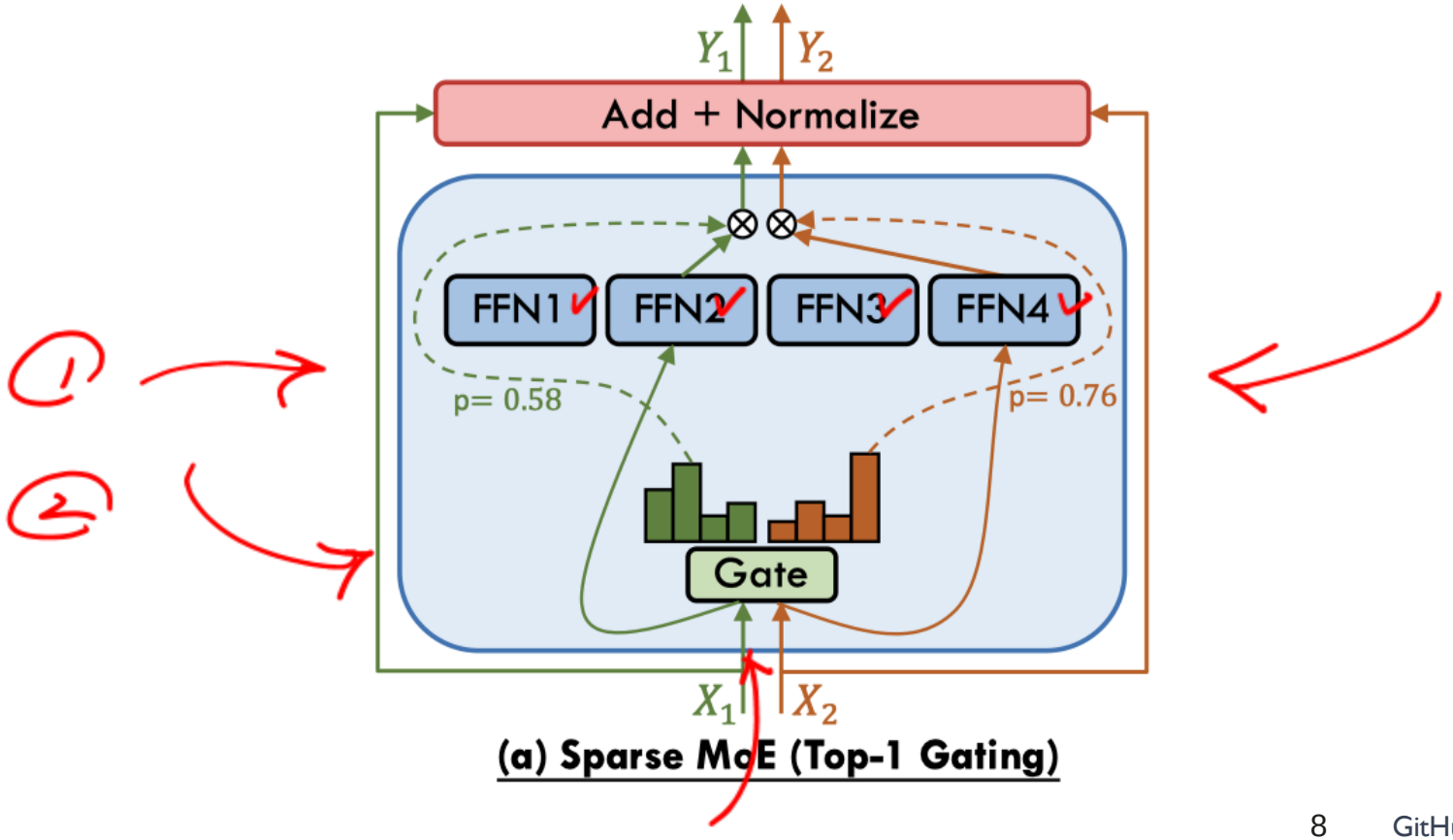
A Survey on Mixture of Experts



- 自然语言处理绿色
- 计算机视觉黄色
- 多模态粉色
- 推荐系统青色

A Survey on Mixture of Experts

- 每个 MoE 层通常由一组 N 个专家网络 $\{f1, \ldots, fN\}$ 和一个“门控网络” G 组成；
- 门控网络通常采用 softmax 函数线性层组成，作用是将输入 Token 引导至适当的专家网络；



Why Replace FFN in Transformer?

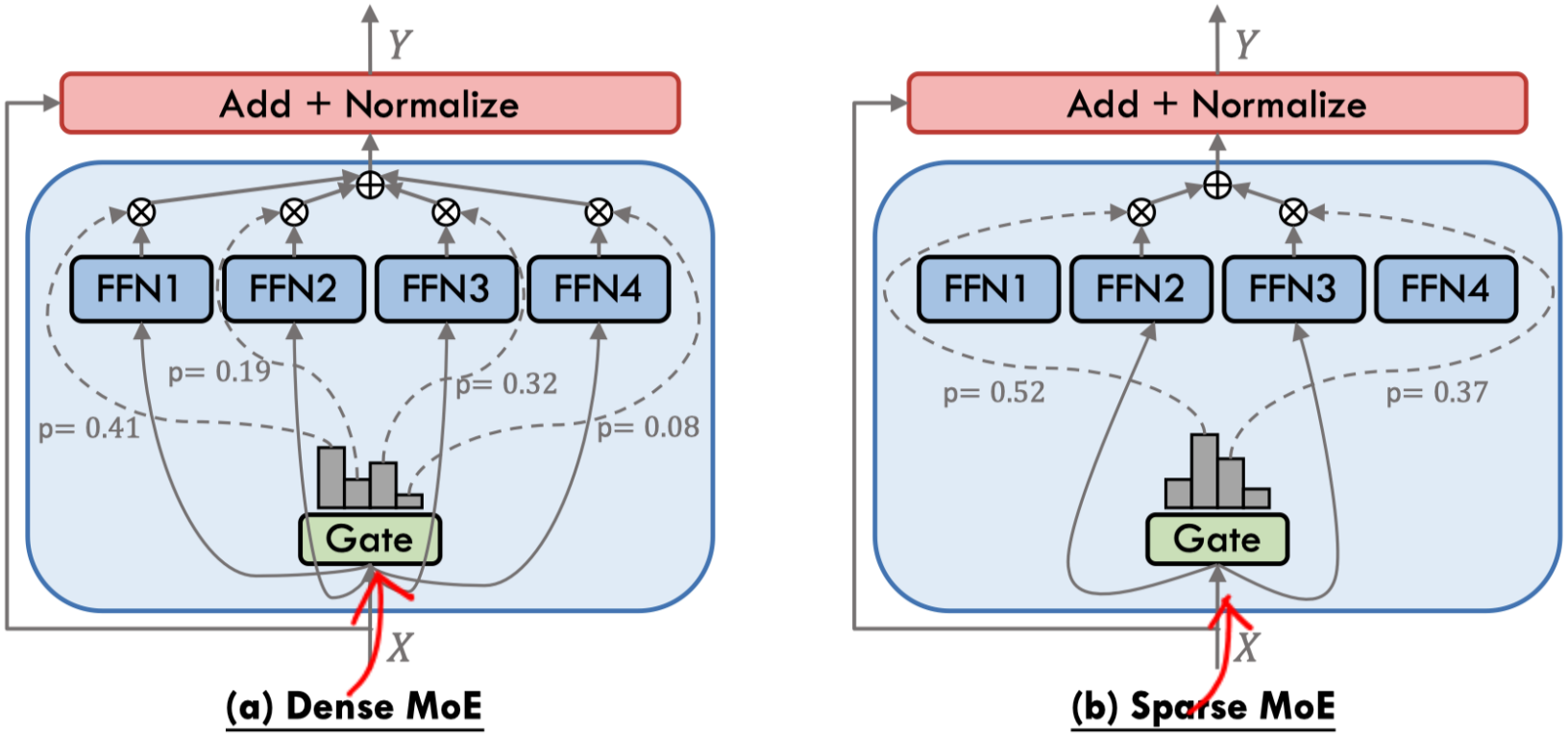
- 为什么 MoE 层的位置选择在每个 Transformer 块内的前馈网络 FFN 层进行替换?

- 因为随着模型的扩大，FFN 的计算需求越来越大，代替 FFN 层能够节省算力，而且提高模型的能力。

MOE 结构的分类：稠密 & 稀疏

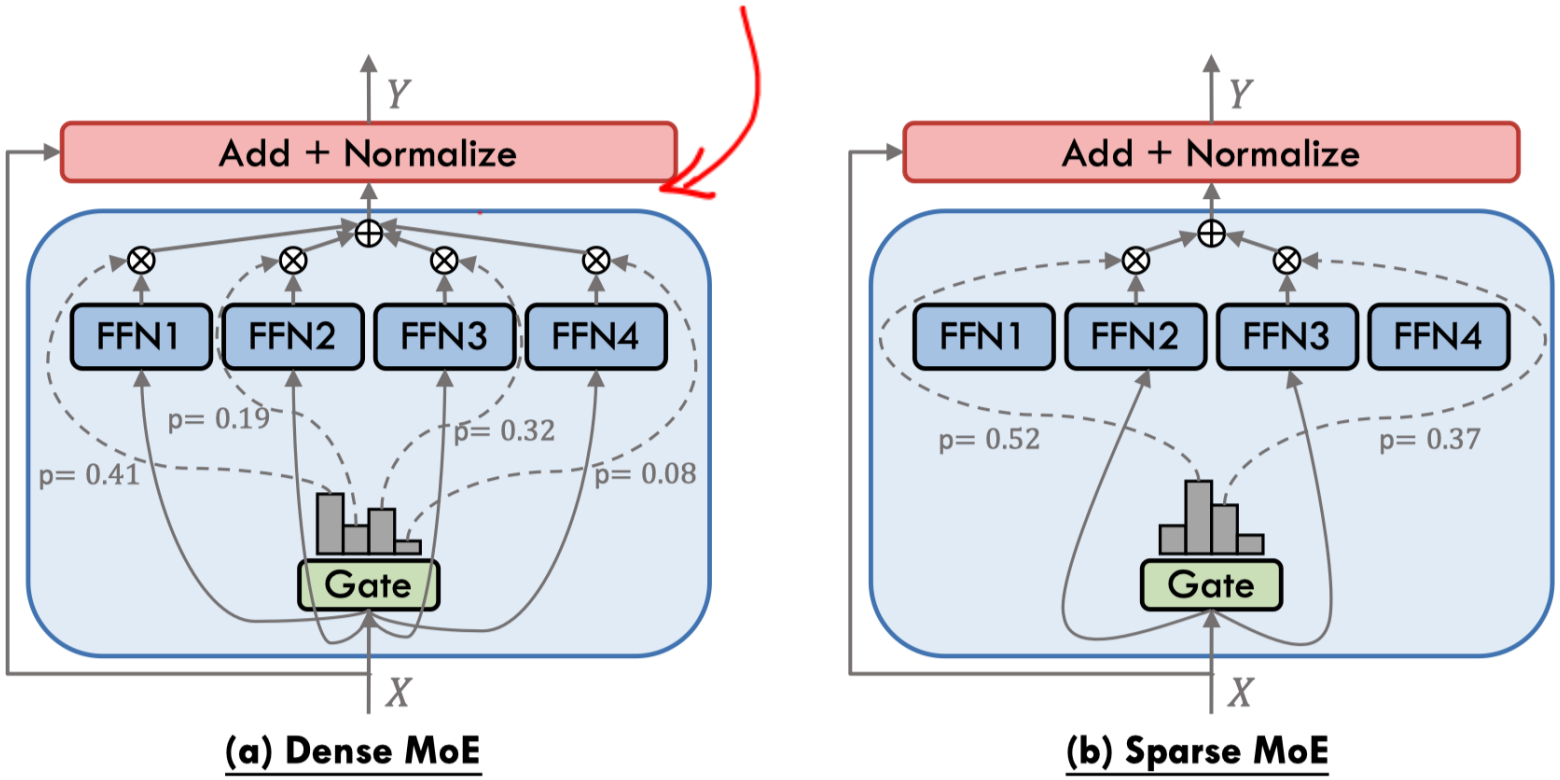
a. **稠密 MOE：**对于每个输入 X，线性 softmax 门控将选择所有专家；

b. **稀疏 MOE：** 选择 top-k 专家来执行条件计算，专家层返回所选专家输出乘以门控函数输出。



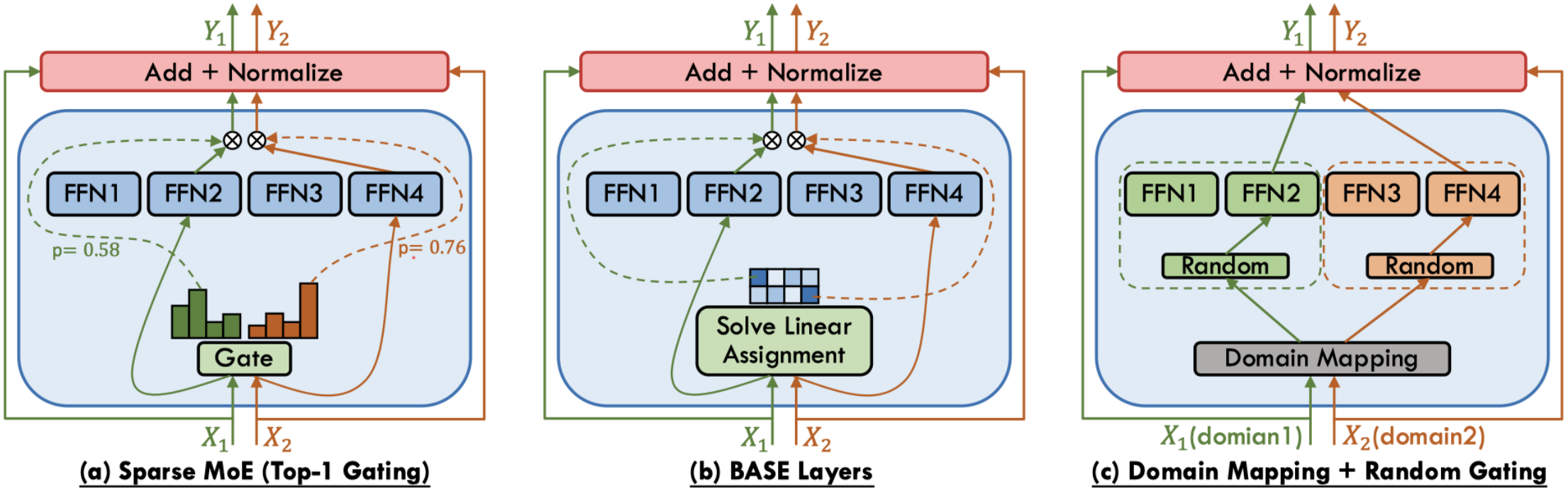
MOE 结构的分类：稠密 & 稀疏

- 稠密混合专家层在每次迭代中激活所有专家网络 $\{f1, \ldots, fN\}$；
- 稠密混合专家 MoE 模型广泛用在 EvoMoE、MoLE、LoRAMoE 和 DS-MoE 等研究。



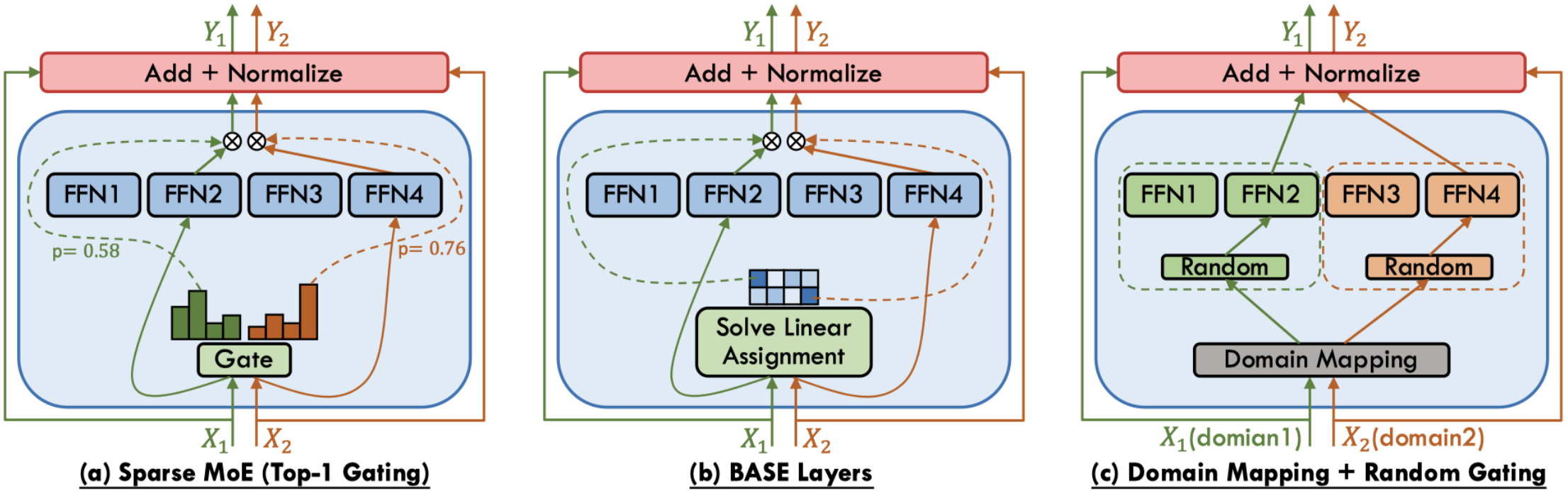
MoE 算法设计

- 门控函数（a.k.a 路由函数或路由器） MoE 基本组件，负责协调专家参与及其各自输出组合。
- 根据每个输入 Token 处理方法，Gating 机制分为三种不同类型：**稀疏门控**，激活专家子集；**稠密门控**，激活所有专家；**软门控**，输入token合并和专家合并 & 可微。



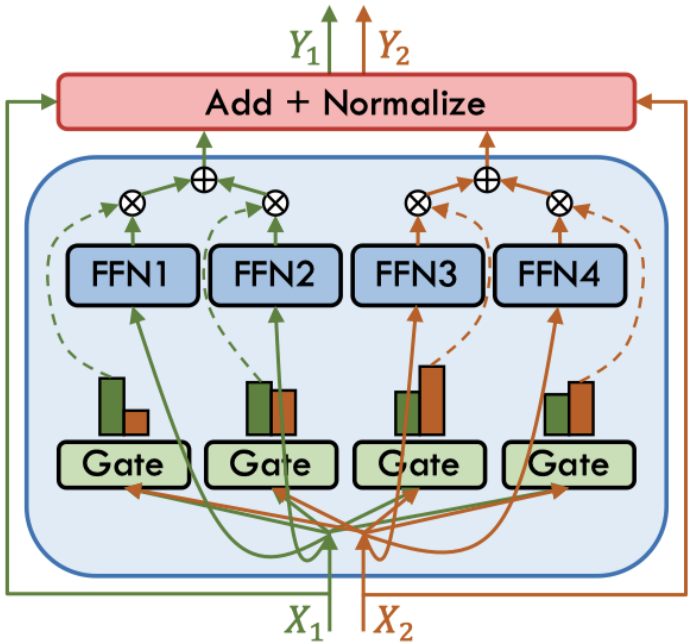
MoE 算法设计

- top-1 门控的稀疏 MoE
- BASE 层
- 分组域映射和随机门控组合

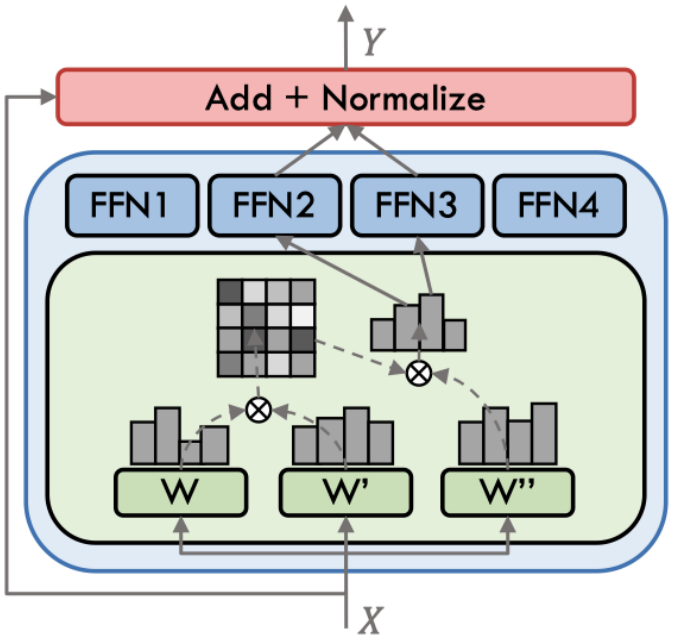


MoE 算法设计

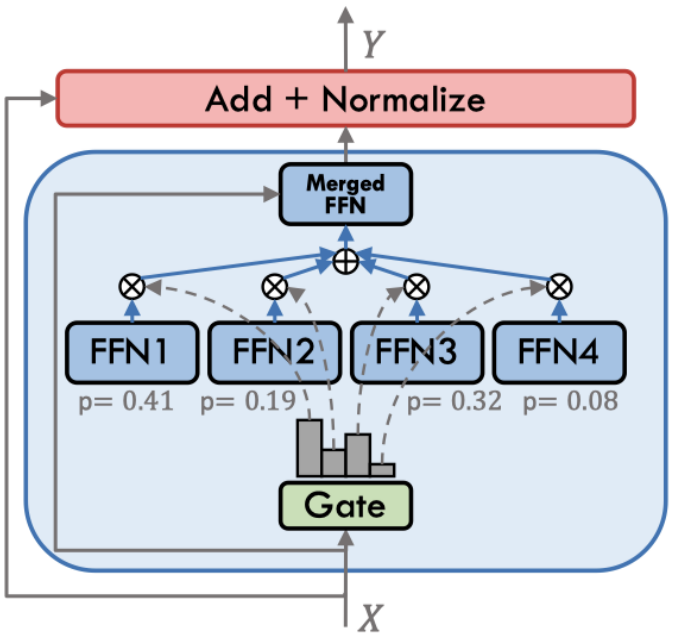
- 专家-选择门控
- 注意力机制的路由器
- 专家合并的软 MoE



(d) Expert-Choice Gating



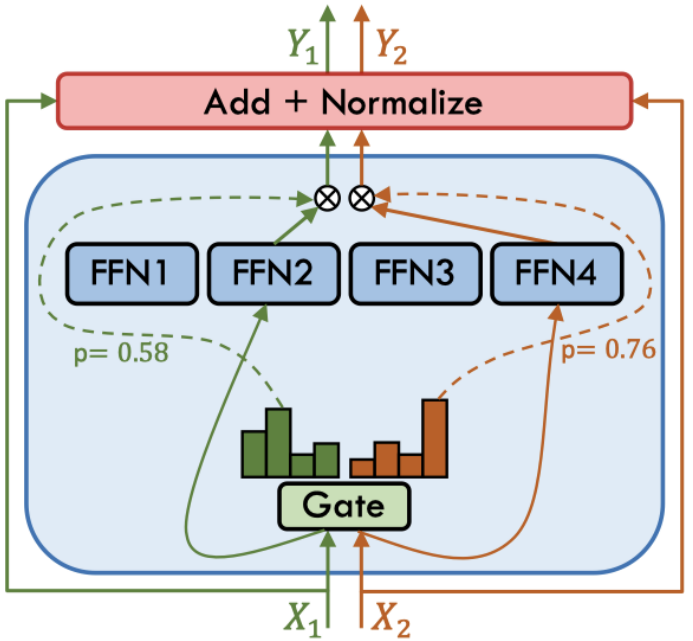
(e) Attention Router



(f) Soft MoE (Expert Merging)

稀疏门控

- 稀疏门控函数选定专家子集来处理每个单独的输入 tokens。
- Shazeer 在 Gshard 提出辅助负载均衡 Loss，专家计算的输出由选择概率加权。
- 由于为每个输入 token 选择专家，因此该方法被视为具有 token 选择的门控函数。



(a) Sparse MoE (Top-1 Gating)

模型大小由总参数量或激活/总参数量表示

- 激活专家数和总专家数在使用时都包含共享专家数量

Reference	Models	Expert Count (Activ./Total)	d_{model}	d_{ffn}	d_{expert}	#L	#H	d_{head}	Placement Frequency	Activation Function	Share Expert Count
GShard [28] (2020)	600B	2/2048	1024	8192	d_{ffn}	36	16	128	1/2	ReLU	0
	200B	2/2048	1024	8192	d_{ffn}	12	16	128	1/2	ReLU	0
	150B	2/512	1024	8192	d_{ffn}	36	16	128	1/2	ReLU	0
	37B	2/128	1024	8192	d_{ffn}	36	16	128	1/2	ReLU	0
Switch [36] (2021)	7B	1/128	768	2048	d_{ffn}	12	12	64	1/2	GEGLU	0
	26B	1/128	1024	2816	d_{ffn}	24	16	64	1/2	GEGLU	0
	395B	1/64	4096	10240	d_{ffn}	24	64	64	1/2	GEGLU	0
	1571B	1/2048	2080	6144	d_{ffn}	15	32	64	1	ReLU	0
GLaM [35] (2021)	0.1B/1.9B	2/64	768	3072	d_{ffn}	12	12	64	1/2	GEGLU	0
	1.7B/27B	2/64	2048	8192	d_{ffn}	24	16	128	1/2	GEGLU	0
	8B/143B	2/64	4096	16384	d_{ffn}	32	32	128	1/2	GEGLU	0
	64B/1.2T	2/64	8192	32768	d_{ffn}	64	128	128	1/2	GEGLU	0
DeepSpeed-MoE [66] (2022)	350M/13B	2/128	1024	$4d_{model}$	d_{ffn}	24	16	64	1/2	GeLU	0
	1.3B/52B	2/128	2048	$4d_{model}$	d_{ffn}	24	16	128	1/2	GeLU	0
	PR-350M/4B	2/32-2/64	1024	$4d_{model}$	d_{ffn}	24	16	64	1/2, 10L-32E, 2L-64E	GeLU	1
	PR-1.3B/31B	2/64-2/128	2048	$4d_{model}$	d_{ffn}	24	16	128	1/2, 10L-64E, 2L-128E	GeLU	1
ST-MoE [37] (2022)	0.8B/4.1B	2/32	1024	2816	d_{ffn}	27	16	64	1/4, add extra FFN	GEGLU	0
	32B/269B	2/64	5120	20480	d_{ffn}	27	64	128	1/4, add extra FFN	GEGLU	0
Mixtral [29] (2023)	13B/47B	2/8	4096	14336	d_{ffn}	32	32	128	1	SwiGLU	0
	39B/141B	2/8	6144	16384	d_{ffn}	56	48	128	1	SwiGLU	0

MOE 模型参数定义

- d_{model} 隐层大小，d_{ffn} 是 FFN 中间层大小，d_{expert} 专家中间层大小
- #L 层数，#H 注意头数量，d_{head} 注意头大小

Reference	Models	Expert Count (Activ./Total)	d_{model}	d_{ffn}	d_{expert}	#L	#H	d_{head}	Placement Frequency	Activation Function	Share Expert Count
LLAMA-MoE [51] (2023)	3.0B/6.7B	2/16	4096	11008	688	32	32	128	1	SwiGLU	0
	3.5B/6.7B	4/16	4096	11008	688	32	32	128	1	SwiGLU	0
	3.5B/6.7B	2/8	4096	11008	1376	32	32	128	1	SwiGLU	0
DeepSeekMoE [69] (2024)	0.24B/1.89B	8/64	1280	-	$\frac{1}{4}d_{ffn}$	9	10	128	1	SwiGLU	1
	2.8B/16.4B	8/66	2048	10944	1408	28	16	128	1, except 1st layer	SwiGLU	2
	22B/145B	16/132	4096	-	$\frac{1}{8}d_{ffn}$	62	32	128	1, except 1st layer	SwiGLU	4
OpenMoE [38] (2024)	339M/650M	2/16	768	3072	d_{ffn}	12	12	64	1/4	SwiGLU	1
	2.6B/8.7B	2/32	2048	8192	d_{ffn}	24	24	128	1/6	SwiGLU	1
	6.8B/34B	2/32	3072	12288	d_{ffn}	32	24	128	1/4	SwiGLU	1
Qwen1.5-MoE [104] (2024)	2.7B/14.3B	8/64	2048	5632	1408	24	16	128	1	SwiGLU	4
DBRX [31] (2024)	36B/132B	4/16	6144	10752	d_{ffn}	40	48	128	1	SwiGLU	0
Jamba [68] (2024)	12B/52B	2/16	4096	14336	d_{ffn}	32	32	128	1/2, 1:7 Attention:Mamba	SwiGLU	0
Skywork-MoE [67] (2024)	22B/146B	2/16	4608	12288	d_{ffn}	52	36	128	1	SwiGLU	0
Yuan 2.0-M32 [84] (2024)	3.7B/40B	2/32	2048	8192	d_{ffn}	24	16	256	1	SwiGLU	0

02

90 年代初期（奠基工作）

奠基工作

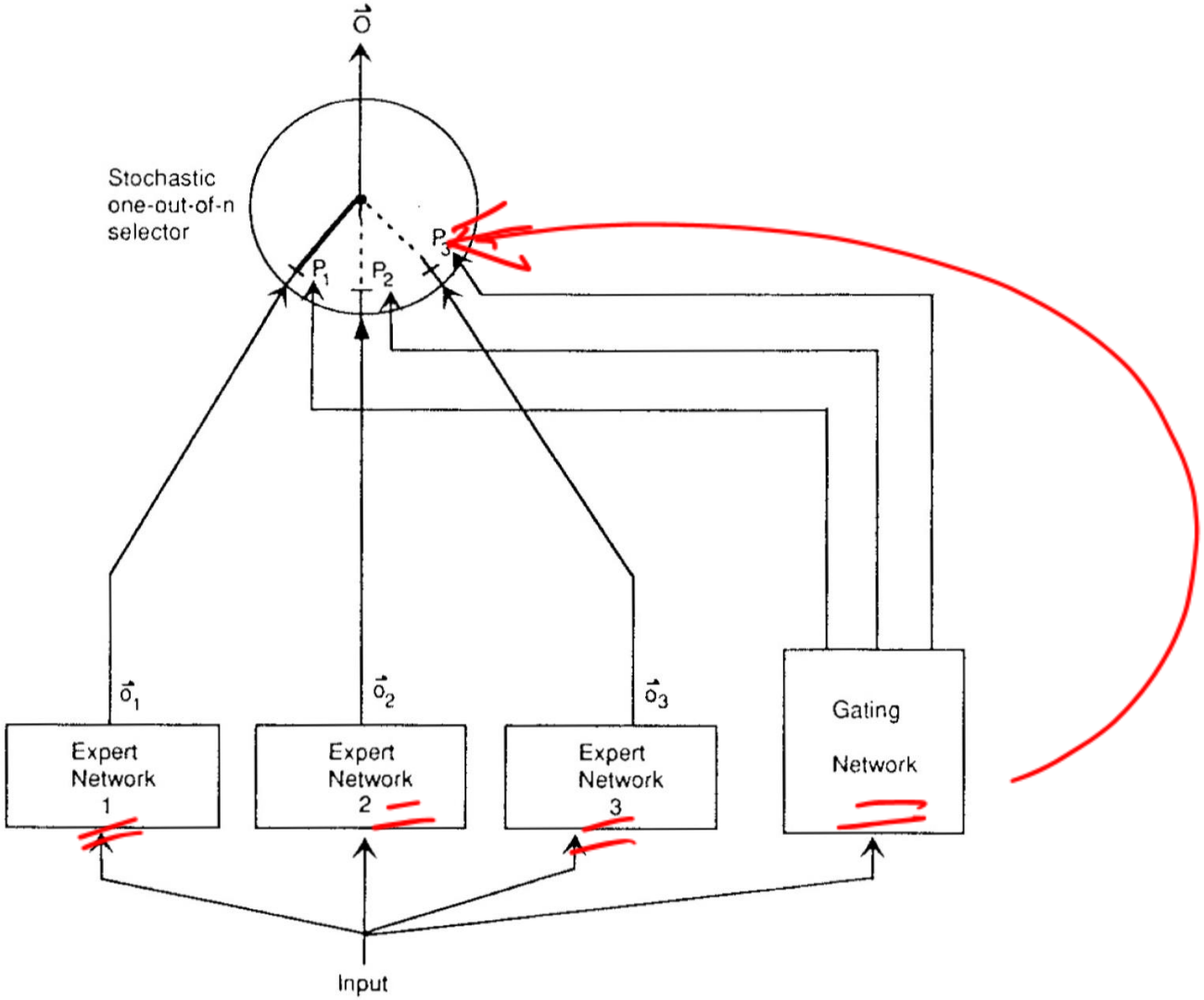
- 1991， Geoffrey Hinton 和 Michael I. Jordan《Adaptive Mixtures of Local Experts》

- **特点：**
  - 最早 MoE 架构，核心思想是通过多个独立 Expert 网络（Experts）处理输入数据不同子集，并由门控网络（Gating Network）动态选择 Expert 。
  - 强调模块化设计，每个 Expert 专注于输入空间特定区域，从而提高模型泛化能力和计算效率。

- **影响：**
  - 奠定模块化神经网络的基础，首次将监督学习与分而治之的思想结合。
  - 大多数 MoE 相关论文都会引用开创性工作。

奠基工作

奠基工作


$\vec{O}$
Stochastic
one-out-of-n
selector
P_1
P_2
P_3
$\vec{o}_1$
$\vec{o}_2$
$\vec{o}_3$
Expert
Network
1
Expert
Network
2
Expert
Network
3
Gating
Network
Input

03

RNN 时代（MOE 架构形成）

MOE 架构形成

- Google，2017年1月《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》，MoE 引入 LSTM，训出最大 137B 参数， Expert 数达到 128k。

- **特点:**
  - 引入路由（Routing Mechanism），使得每个输入只激活少量 Expert，实现计算成本与模型规模分离；
  - 引入 Top-k 门控机制，通过选择前 k 个 Expert 处理输入，进一步优化了计算效率；
  - 提出了负载均衡概念，通过辅助损失函数确保 Expert 间均衡，避免某些 Expert 被过度激活；

MOE 架构形成

- Google，2017年1月《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》，MoE 引入 LSTM，训出最大 137B 参数， Expert 数达到 128k。

- **影响：**
  - 该论文被认为是MoE在大语言模型中的里程碑式研究；
  - 首次将MoE应用于大规模语言模型，提出了稀疏激活的MoE架构。

MOE 架构形成

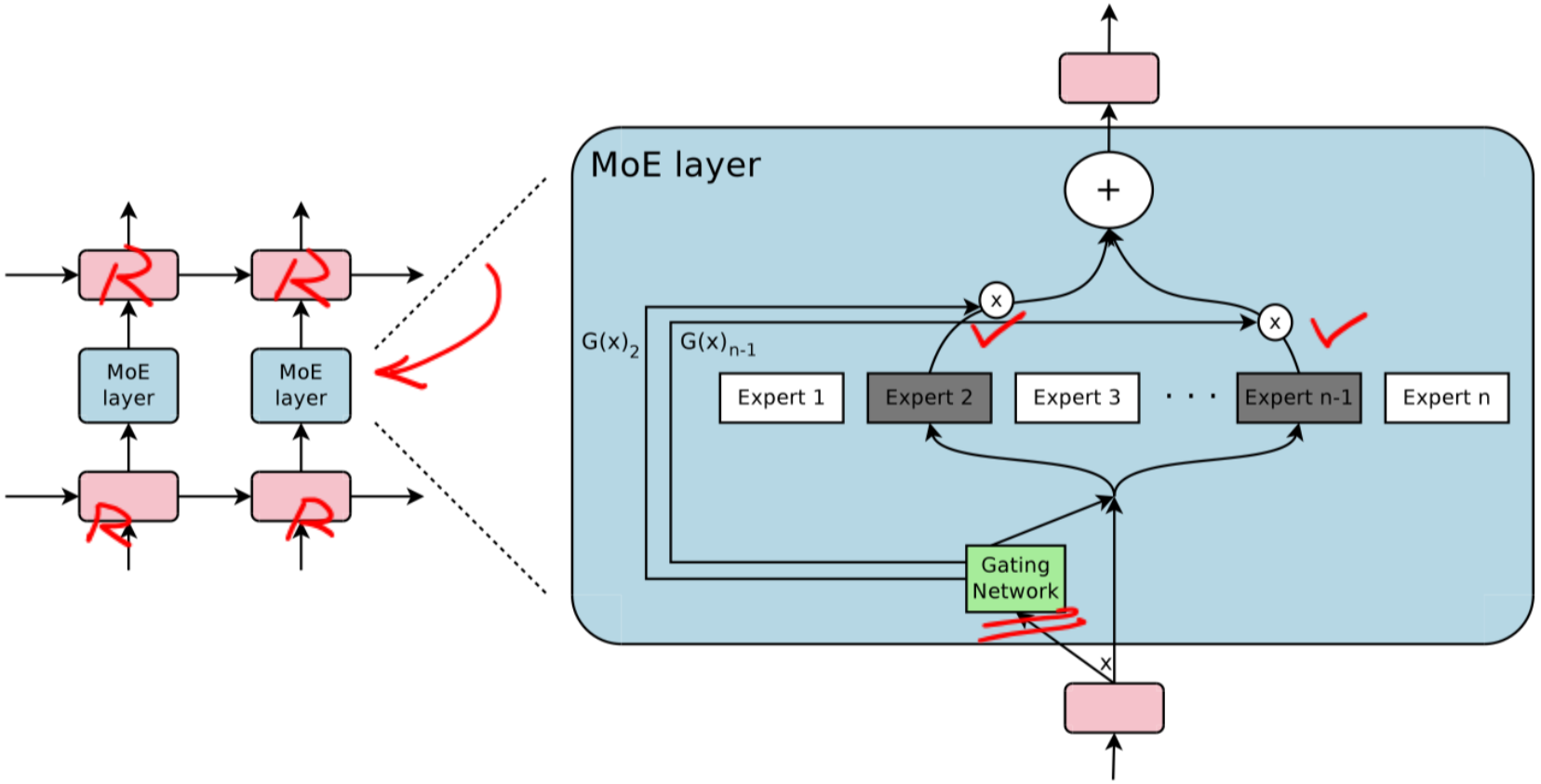


Figure 1: A Mixture of Experts (MoE) layer embedded within a recurrent language model. In this case, the sparse gating function selects two experts to perform computations. Their outputs are modulated by the outputs of the gating network.

04

Transformer 时代（提升效果）

GShard

- Google， 2020年6月《GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding》，把MoE应用在encoder-decoder结构 transformer 模型，每两层将 FFN 替换成 MoE，参数量从 12.5B 到 600B 系列 MoE，每层最大 Expert 数达 2048。

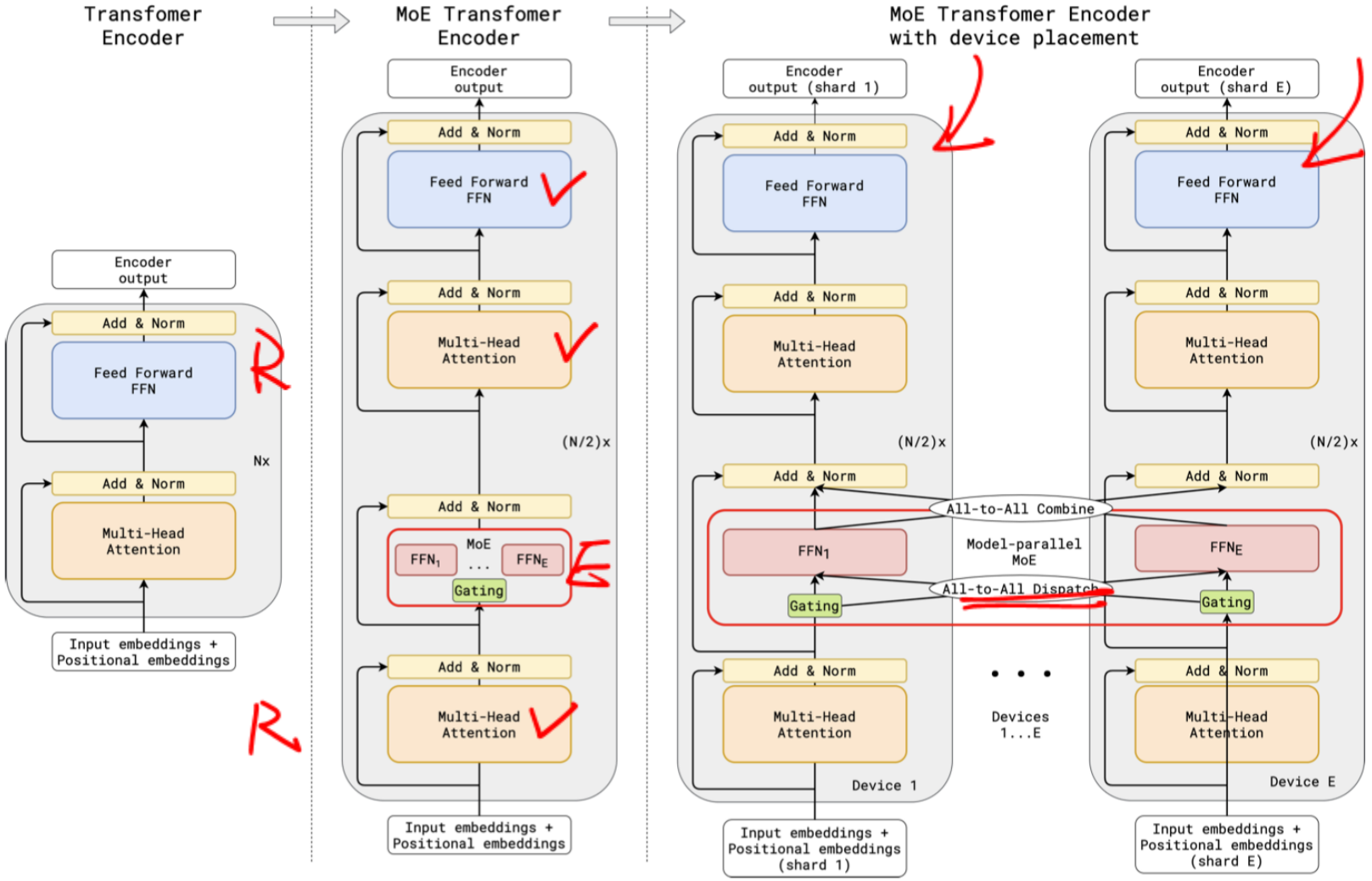
- **特点：**
  - 提出 MoE transformer，通过稀疏激活 MoE 层替代传统的前馈网络（FFN）层。
  - 引入了 随机路由 和 Expert 容量 的概念，优化了大规模分布式训练中的负载均衡问题。
  - 提出了 GShard 框架，支持 6000 亿参数模型训练，展示 MoE 在大模型中的潜力。
  - 通过条件计算和自动分片技术，进一步优化了MoE的扩展性和效率。

GShard

- Google， 2020年6月《GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding》，把MoE应用在encoder-decoder结构 transformer 模型，每两层将 FFN 替换成 MoE，参数量从 12.5B 到 600B 系列 MoE，每层最大 Expert 数达 2048。

- **影响：**
  - 将 MoE 架构与 Transformer 模型结合，展示在实际生产环境中部署 MoE 模型。

GShard


Transfomer Encoder
MoE Transfomer Encoder
MoE Transfomer Encoder with device placement
Encoder output
Add & Norm
Feed Forward FFN
Multi-Head Attention
Nx
(N/2)x
MoE
FFN1
FFNE
Gating
Input embeddings + Positional embeddings
Encoder output (shard 1)
Encoder output (shard E)
All-to-All Combine
All-to-All Dispatch
Model-parallel MoE
Devices 1...E
Device 1
Device E
Input embeddings + Positional embeddings (shard 1)
Input embeddings + Positional embeddings (shard E)

Switch Transformers

- Google， 2021 年 1 月《Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity》，T5（encoder-decoder）基础上，简化 routing 策略，实现 1.6T 参数量 switch transformer。

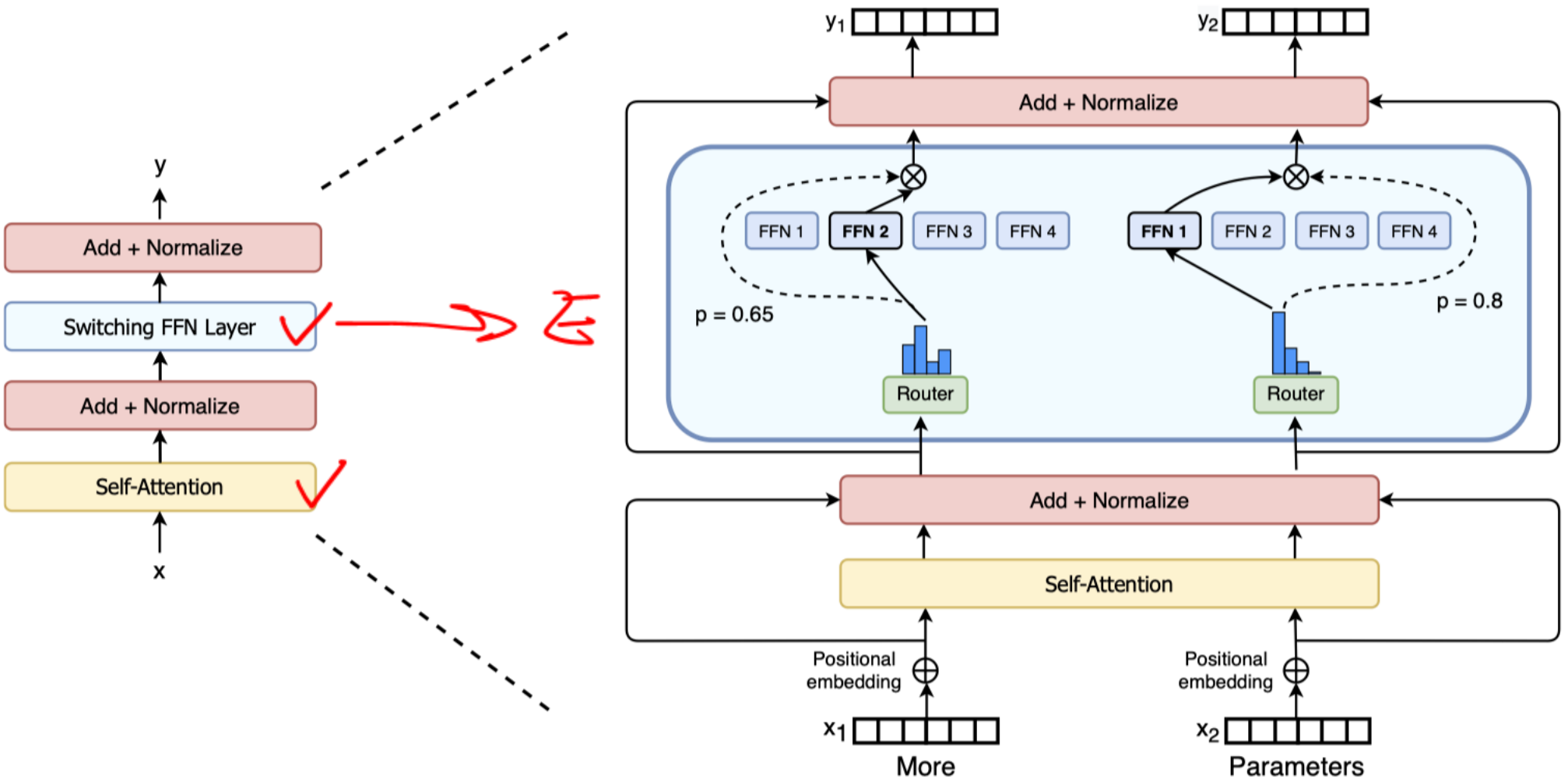
- **特点：**
  - 提出了Switch Transformer架构，简化了MoE的路由机制，仅选择单个 Expert 进行激活。
  - 通过 稀疏门控机制 和 Expert 容量限制，优化计算效率和负载均衡，实现万亿参数级模型规模扩展。

Switch Transformers

- Google， 2021 年 1 月《Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity》，T5（encoder-decoder）基础上，简化 routing 策略，实现 1.6T 参数量 switch transformer。

- **影响：**
  - 展示了 MoE 在大模型中的潜力。
  - 对 scaling law、蒸馏很多详细探索，影响深远，MoE 领域重要工作。

Switch Transformers



ST-MoE

- 2022年2月，Google发布《ST-MoE: Designing Stable and Transferable Sparse Expert Models》，基于 encoder-decoder 结构 MoE，最大 269B，32B 激活参数。解决 MoE 模型在训练和微调中的不稳定性问题，并提升其迁移学习能力。

- **特点：**
  - ST-MoE 通过引入梯度裁剪、噪声注入、路由器限制缓解 MoE 模型的训练不稳定性问题；
  - 优化微调策略，使ST-MoE 提升迁移学习能力，更好地适应下游任务，减少过拟合；

ST-MoE

- 2022年2月，Google发布《ST-MoE: Designing Stable and Transferable Sparse Expert Models》，基于 encoder-decoder 结构 MoE，最大 269B，32B 激活参数。解决 MoE 模型在训练和微调中的不稳定性问题，并提升其迁移学习能力。

- **影响：**
  - 为 MoE 工程实现提供设计指南，推动 MoE 技术在大语言模型中的应用。
  - 设计思路（如稳定性优化和负载均衡机制）为后续 MoE 研究提供了重要借鉴。

05

GPT 时代（智能涌现）

GLaM

- Google， 2021年12月，《GLaM: Efficient Scaling of Language Models with Mixture-of-Experts》，采用稀疏 MoE，包含 1.2 万亿参数，实际激活参数 97B，最大为 1.2T 的 decoder-only 模型。

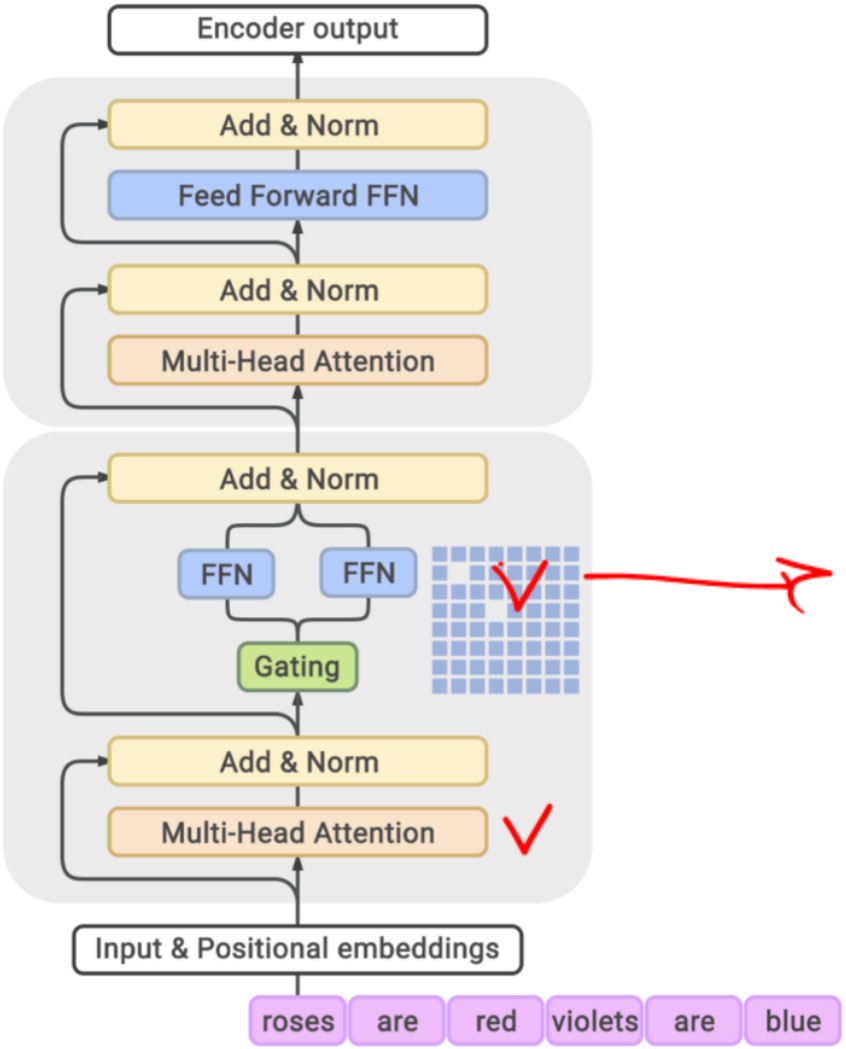
- **特点：**
  - 稀疏激活机制。每个输入 token 通过门控网络动态选择 2 个 Expert ， 仅激活相关 Expert 进行计算。
  - 实现了条件计算，即模型根据输入动态调整计算路径，从而显著提高了计算效率。
  - 每个 MoE 层包含 64 个 Expert ， 可以分布在多个计算设备上，实现跨设备扩展。

GLaM

- Google， 2021年12月，《GLaM: Efficient Scaling of Language Models with Mixture-of-Experts》，采用稀疏 MoE，包含 1.2 万亿参数，实际激活参数 97B，最大为 1.2T 的 decoder-only 模型。

- **影响：**
  - 展示 MoE 在 多任务学习 和 多语言处理 中优势，提升模型的泛化能力和效率。
  - GLaM 稀疏激活和负载均衡机制被应用于 Mistral 8x7B，后续模型设计提供重要参考。

GLaM


Encoder output
Add & Norm
Feed Forward FFN
Add & Norm
Multi-Head Attention
Add & Norm
FFN
FFN
Gating
Add & Norm
Multi-Head Attention
Input & Positional embeddings
roses
are
red
violets
are
blue

DeepSeek MoE

- 幻方量化， 2024年1月《DeepSeek MoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models》，

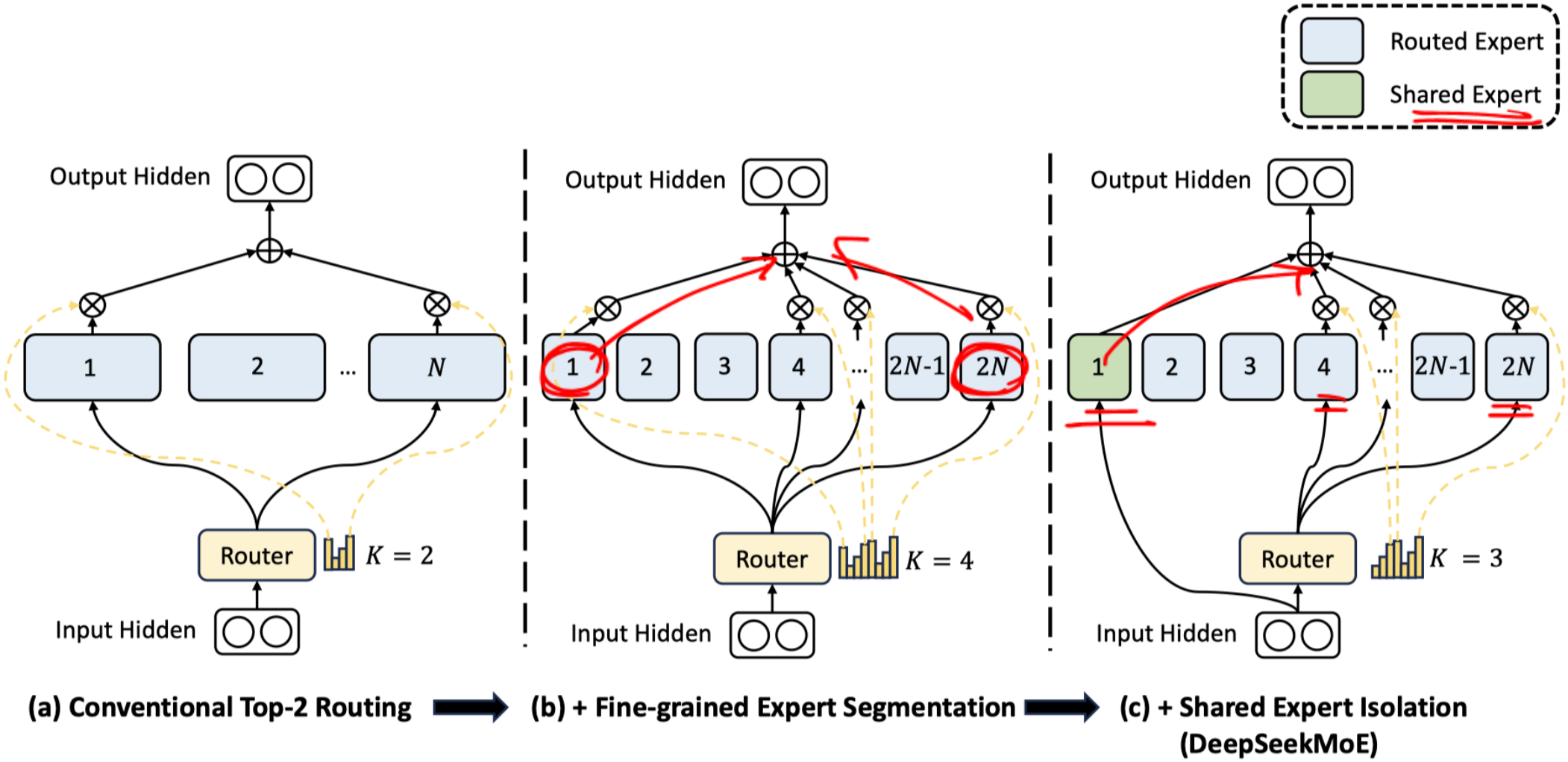
- **特点：**
  - Expert 共享机制：部分 Expert 在不同Tokens或层间共享参数，减少模型冗余，同时提高了参数效率。使得模型在保持高性能同时，计算开销降低 40%。
  - 内存优化：通过多头潜在注意力机制（MLA） 和键值缓存优化，减少生成任务中的浮点运算量，推理延迟降低了 35%。

DeepSeek MoE

- 幻方量化， 2024年1月《DeepSeek MoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models》，

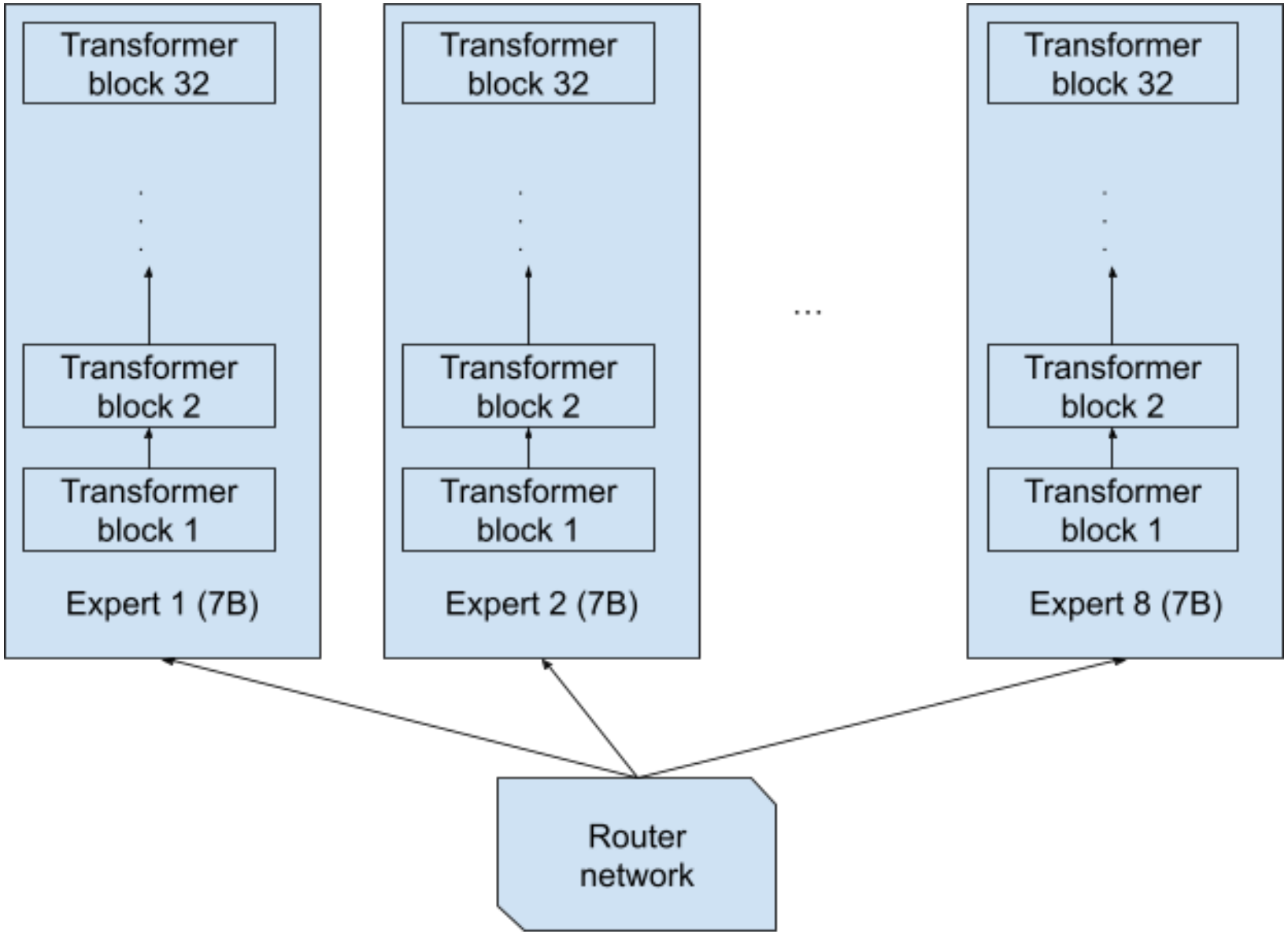
- **影响：**
  - 低成本与高性能：DeepSeek MoE 架构创新和系统优化，实现百倍性价比提升，打破传统大模型依赖算力范式，为资源受限场景下 AI 应用提供新思路。
  - 开源与生态建设：DeepSeek MoE 开源版本在文本生成、代码编写和逻辑推理等任务中表现优异，推动了 MoE 技术普及和应用。

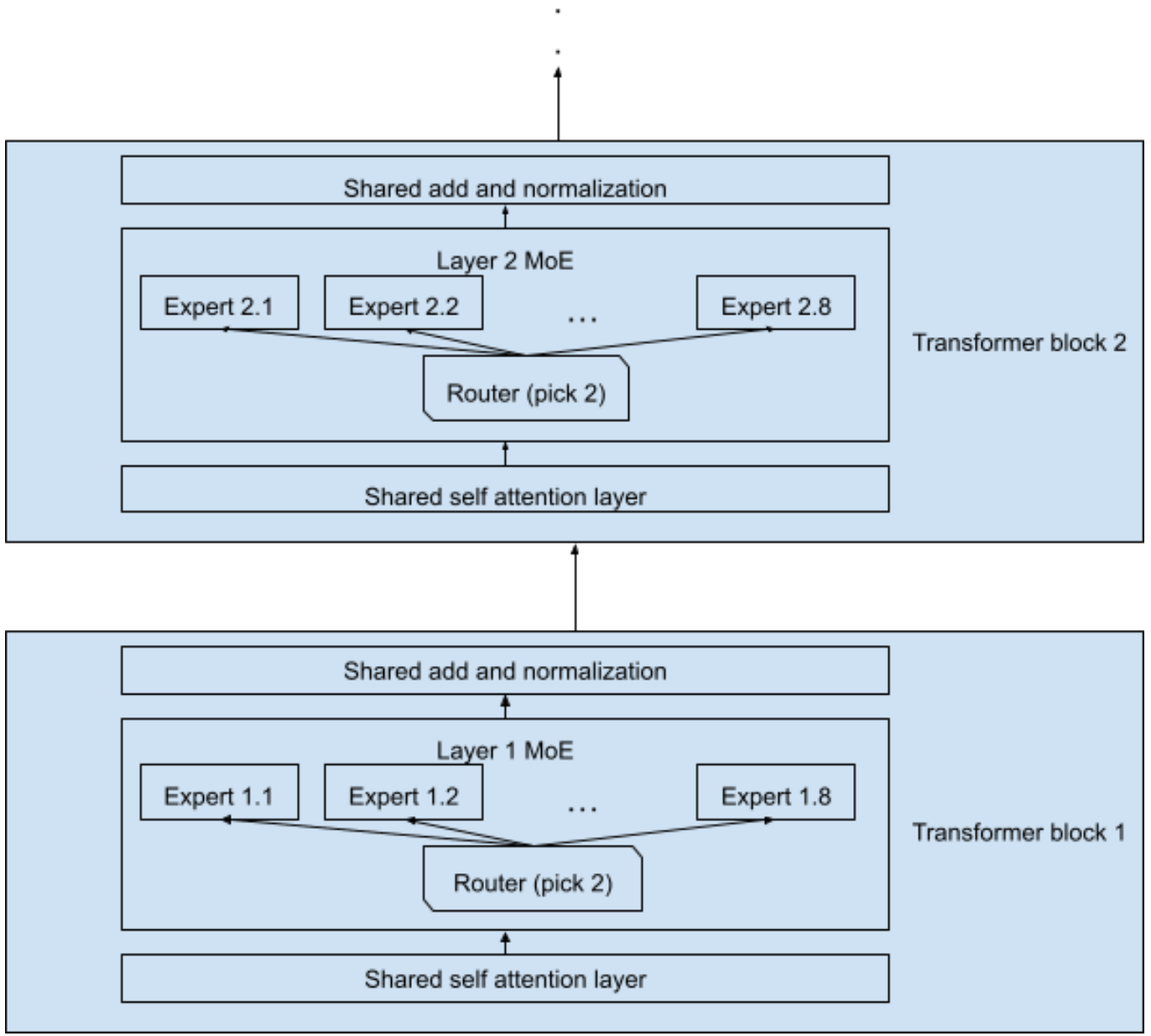
DeepSeek MoE



06

Mixtral-MOE 可视化

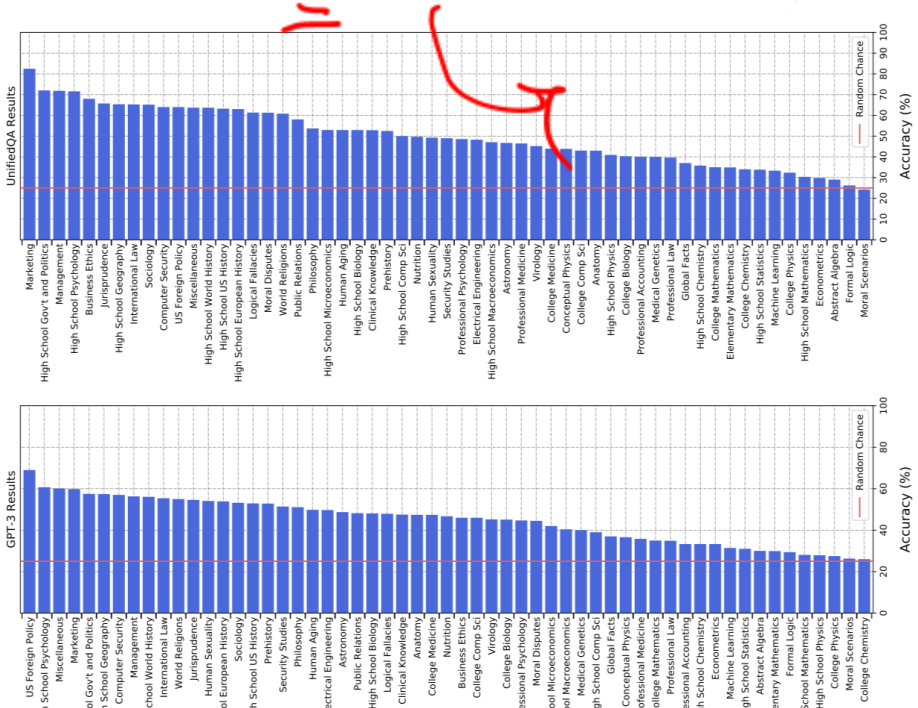

Transformer block 32
Transformer block 2
Transformer block 1
Expert 1 (7B)
Transformer block 32
Transformer block 2
Transformer block 1
Expert 2 (7B)
...
Transformer block 32
Transformer block 2
Transformer block 1
Expert 8 (7B)
Router network


Shared add and normalization
Layer 2 MoE
Expert 2.1
Expert 2.2
...
Expert 2.8
Router (pick 2)
Shared self attention layer
Transformer block 2
Shared add and normalization
Layer 1 MoE
Expert 1.1
Expert 1.2
...
Expert 1.8
Router (pick 2)
Shared self attention layer
Transformer block 1

FFN

MMLU 实验

- 使用 MMLU（ Massive Multitask Language Understanding）基准测试进行实验。
- MMLU 包括 57 个主题多项选择题，涵盖领域广泛，如抽象代数、信仰、解剖学、天文学等。
- 以Mistral 8x7B 为例记录第 1 层、第 16 层和第 32 层 8 位 Expert 中每个 Expert 的激活情况。

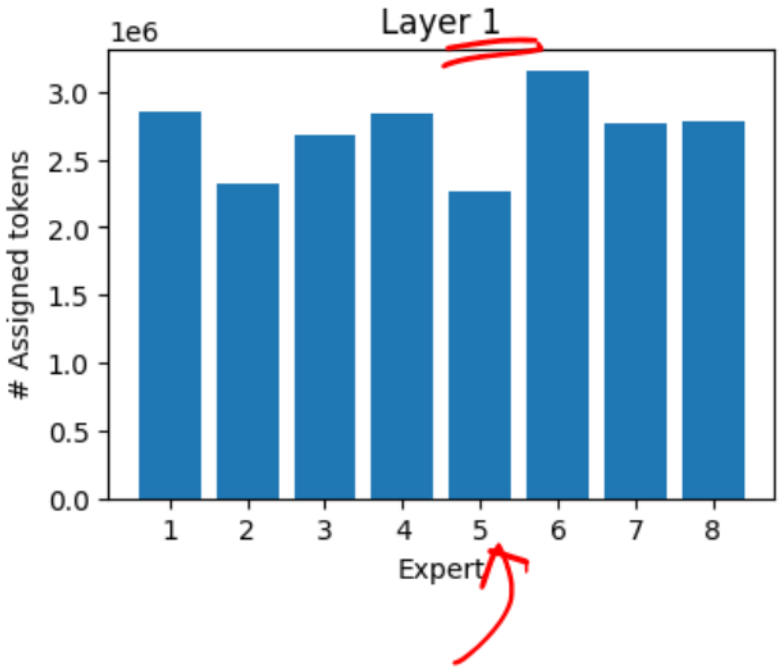


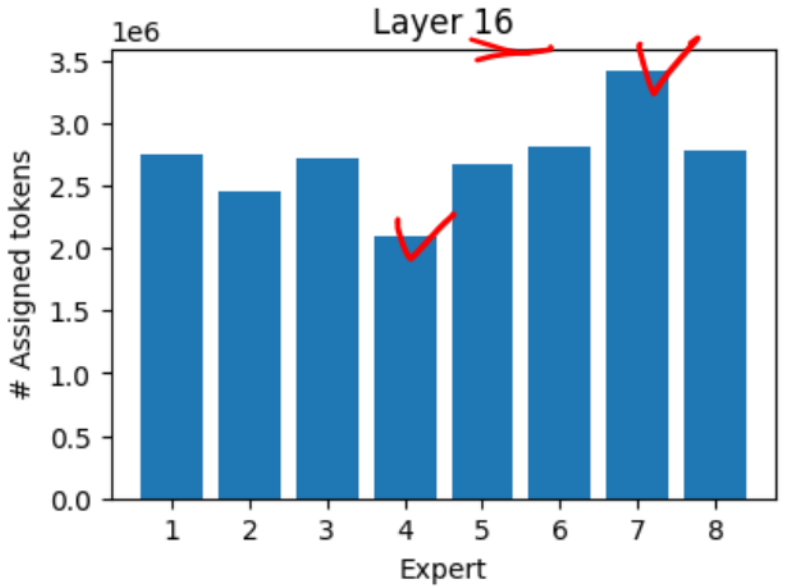
Task	Tested Concepts
Abstract Algebra	Groups, rings, fields, vector spaces, ...
Anatomy	Central nervous system, circulatory system, ...
Astronomy	Solar system, galaxies, asteroids, ...
Business Ethics	Corporate responsibility, stakeholders, regulation, ...
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...
College Biology	Cellular structure, molecular biology, ecology, ...
College Chemistry	Analytical, organic, inorganic, physical, ...
College Computer Science	Algorithms, systems, graphs, recursion, ...
College Mathematics	Differential equations, real analysis, combinatorics, ...
College Medicine	Introductory biochemistry, sociology, reasoning, ...
College Physics	Electromagnetism, thermodynamics, special relativity, ...
Computer Security	Cryptography, malware, side channels, fuzzing, ...
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...
Econometrics	Volatility, long-run relationships, forecasting, ...
Electrical Engineering	Circuits, power systems, electrical drives, ...
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...
Formal Logic	Propositions, predicate logic, first-order logic, ...
Global Facts	Extreme poverty, literacy rates, life expectancy, ...
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...
High School Chemistry	Chemical reactions, ions, acids and bases, ...
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...
High School European History	Renaissance, reformation, industrialization, ...
High School Geography	Population migration, rural land-use, urban processes, ...
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...
High School Macroeconomics	Economic indicators, national income, international trade, ...
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...
High School Microeconomics	Supply and demand, imperfect competition, market failure, ...
High School Physics	Kinematics, energy, torque, fluid pressure, ...
High School Psychology	Behavior, personality, emotions, learning, ...
High School Statistics	Random variables, sampling distributions, chi-square tests, ...
High School US History	Civil War, the Great Depression, The Great Society, ...
High School World History	Ottoman empire, economic imperialism, World War I, ...
Human Aging	Senescence, dementia, longevity, personality changes, ...
Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...
International Law	Human rights, sovereignty, law of the sea, use of force, ...
Jurisprudence	Natural law, classical legal positivism, legal realism, ...
Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...
Machine Learning	SVMs, VC dimension, deep learning architectures, ...
Management	Organizing, communication, organizational structure, ...
Marketing	Segmentation, pricing, market research, ...
Medical Genetics	Genes and cancer, common chromosome disorders, ...
Miscellaneous	Agriculture, Fermi estimation, pop culture, ...
Moral Disputes	Freedom of speech, addiction, the death penalty, ...
Moral Scenarios	Detecting physical violence, stealing, externalities, ...
Nutrition	Metabolism, water-soluble vitamins, diabetes, ...
Philosophy	Skepticism, phronesis, skepticism, Singer's Drowning Child, ...
Prehistory	Neanderthals, Mesoamerica, extinction, stone tools, ...
Professional Accounting	Auditing, reporting, regulation, valuation, ...
Professional Law	Torts, criminal law, contracts, property, evidence, ...
Professional Medicine	Diagnosis, pharmacotherapy, disease prevention, ...
Professional Psychology	Diagnosis, biology and behavior, lifespan development, ...
Public Relations	Media theory, crisis management, intelligence gathering, ...
Security Studies	Environmental security, terrorism, weapons of mass destruction, ...
Sociology	Socialization, cities and community, inequality and wealth, ...
US Foreign Policy	Soft power, Cold War foreign policy, isolationism, ...
Virology	Epidemiology, coronaviruses, retroviruses, herpesviruses, ...
World Religions	Judaism, Christianity, Islam, Buddhism, Jainism, ...

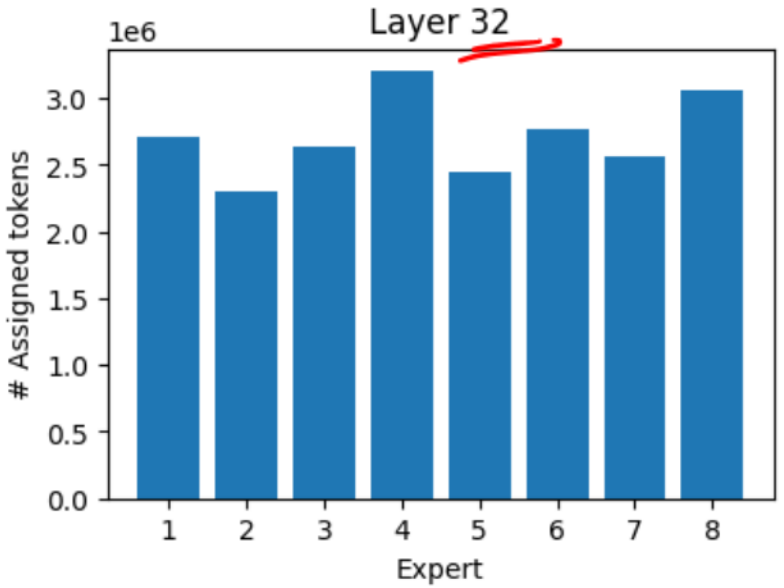
Table 2: Summary of all 57 tasks.

负载均衡测试

- Expert 可以获得均衡负载，但最忙碌Expert 仍可获得比最闲 Expert 多40%~60% Tokens。



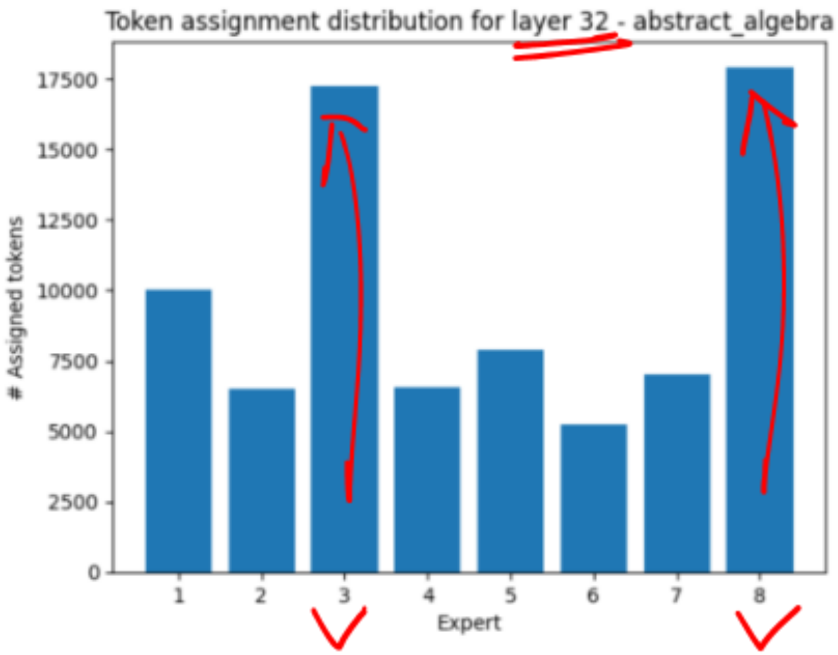




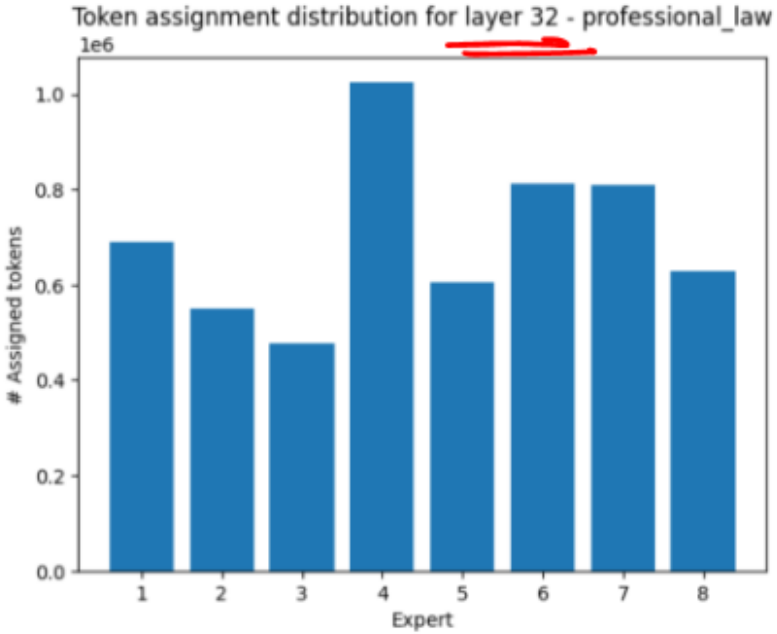
领域 Expert 任务分配

- 某些领域比其他领域更能激活部分 Expert， Expert 能针对领域学习。针对 32 层：

抽象代数 ✓

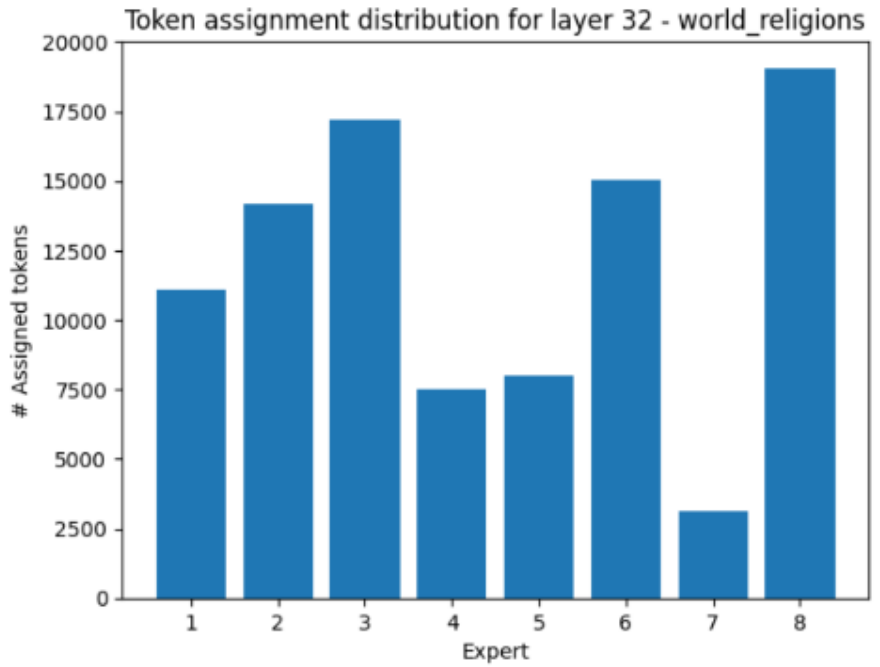


专业法学 ✓



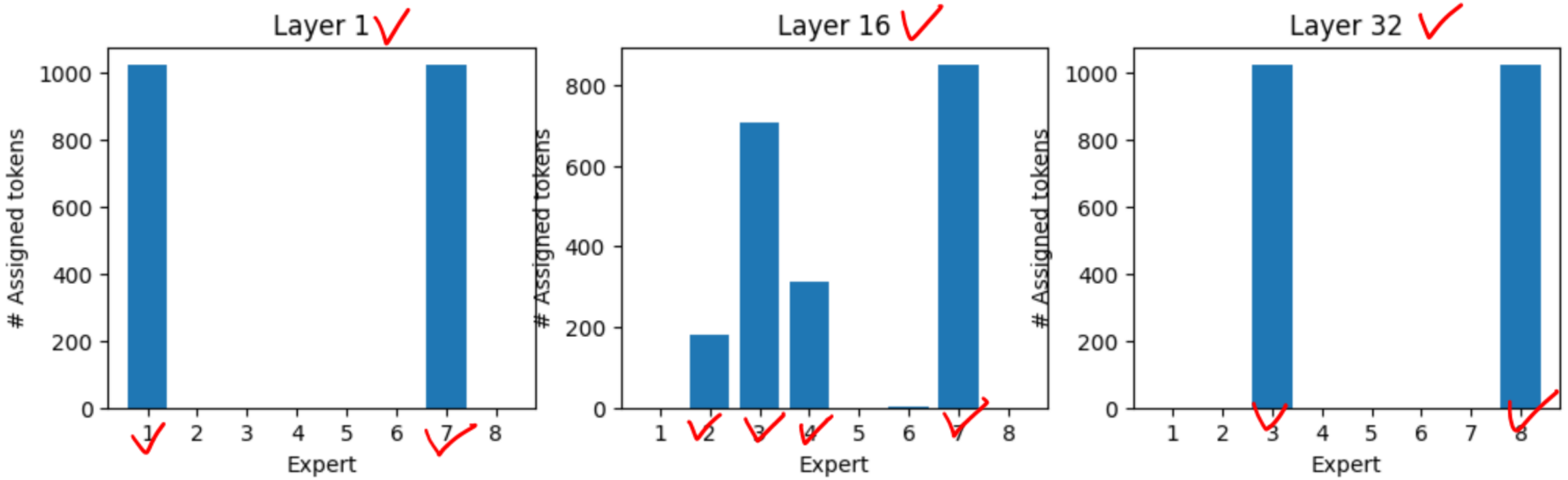
领域 Expert 任务分配

- Expert 的负载分布倾向于在不同的主题范围内保持一致。
- 但当所有样本都完全属于某个主题时，可能会出现很大概率的分布不平衡。



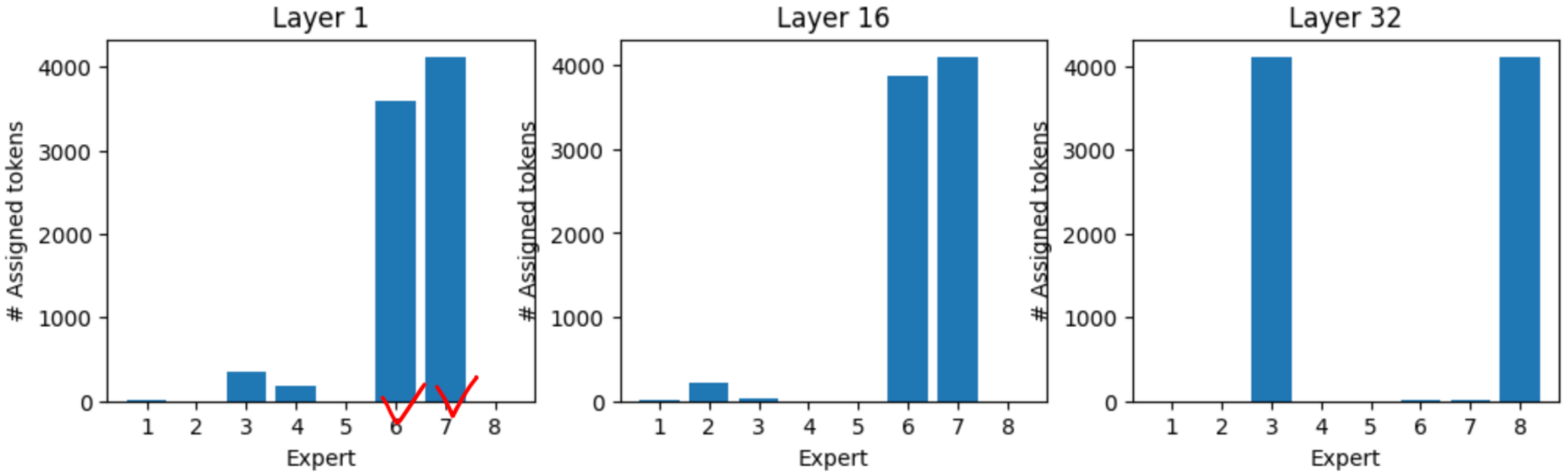
按 Tokens 划分首选 Expert

- 每个Tokens是否都有首选 Expert？ Tokens “：” 的 Expert 分配



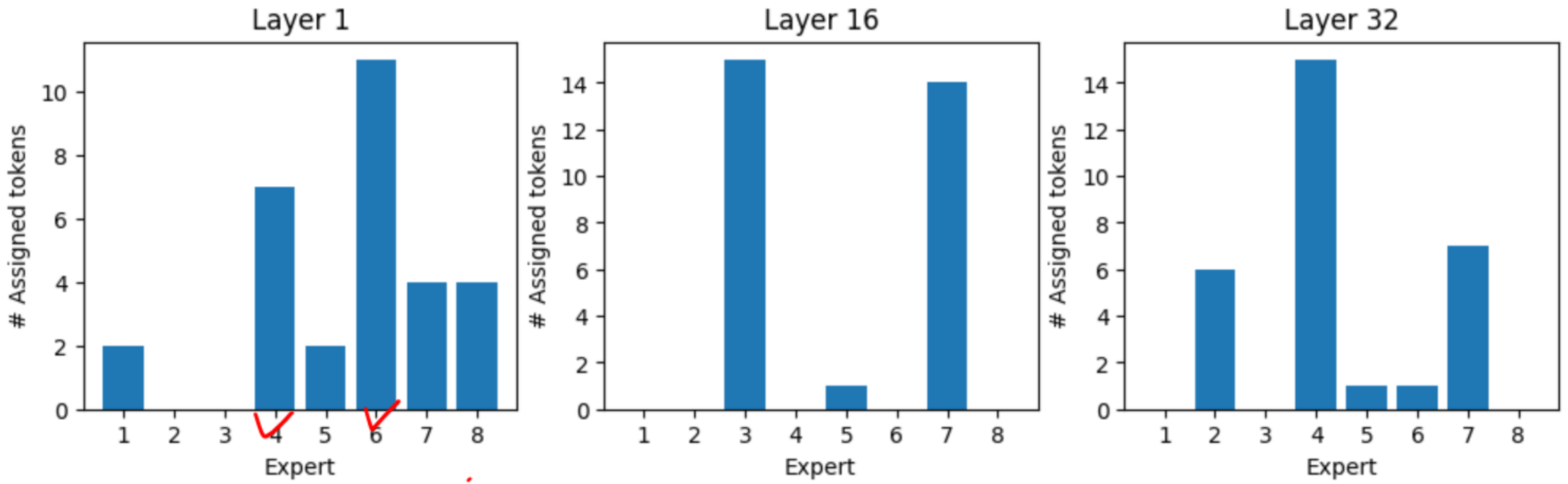
按 Tokens 划分首选 Expert

- Expert 分配 Tokens “。”



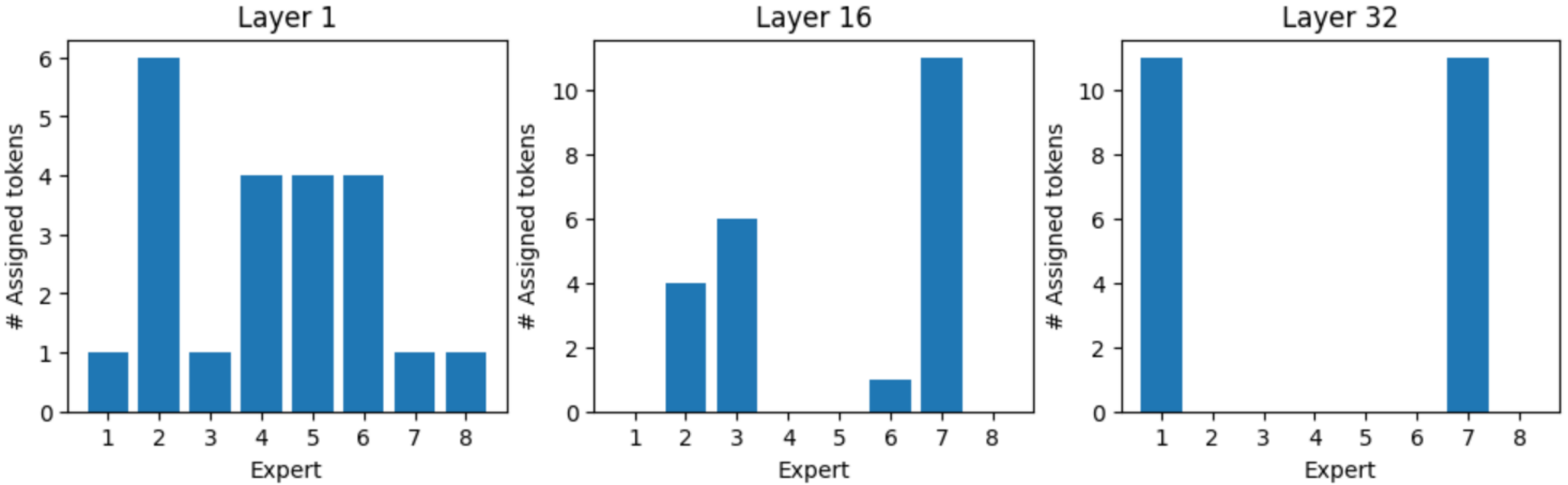
按 Tokens 划分首选 Expert

- Expert 分配 Tokens “what”



按 Tokens 划分的首选 Expert

- Expert 分配 Tokens “who”



思考与小结

总结与思考

1. MoE 架构核心优势在于通过稀疏激活 和 条件计算，显著提升大规模模型训练和推理效率。
2. 从 1991 年首次提出 MoE 架构到近年来的万亿 MoE 参数，成为大模型领域重要方向。
3. MoE 架构不断演进，随着模块化设计和分布式发展，MoE 架构在更多场景中发挥重要作用。

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub https://github.com/chenzomi12/AIFoundation

引用与参考

- https://mp.weixin.qq.com/s/6kzCMsJuavkZPG0YCKgeig
- https://www.zhihu.com/tardis/zm/art/677638939?source_id=1003
- https://huggingface.co/blog/zh/moe
- https://mp.weixin.qq.com/s/mOrAYo3qEACjSwcRPG7fWw
- https://mp.weixin.qq.com/s/x39hqf8xn1cUlnxEIM0_ww
- https://mp.weixin.qq.com/s/ZXjwnO103e-wXJGmmKi-Pw
- https://mp.weixin.qq.com/s/8Y281VYaLu5jHoAvQVvVJg
- https://blog.csdn.net/weixin_43013480/article/details/139301000
- https://developer.nvidia.com/zh-cn/blog/applying-mixture-of-experts-in-llm-architectures/
- https://www.zair.top/post/mixture-of-experts/
- https://my.oschina.net/IDP/blog/16513157

- PPT 开源在：
- https://github.com/chenzomi12/AIInfra
- 夸克链接：https://pan.quark.cn/s/74fb24be8eff