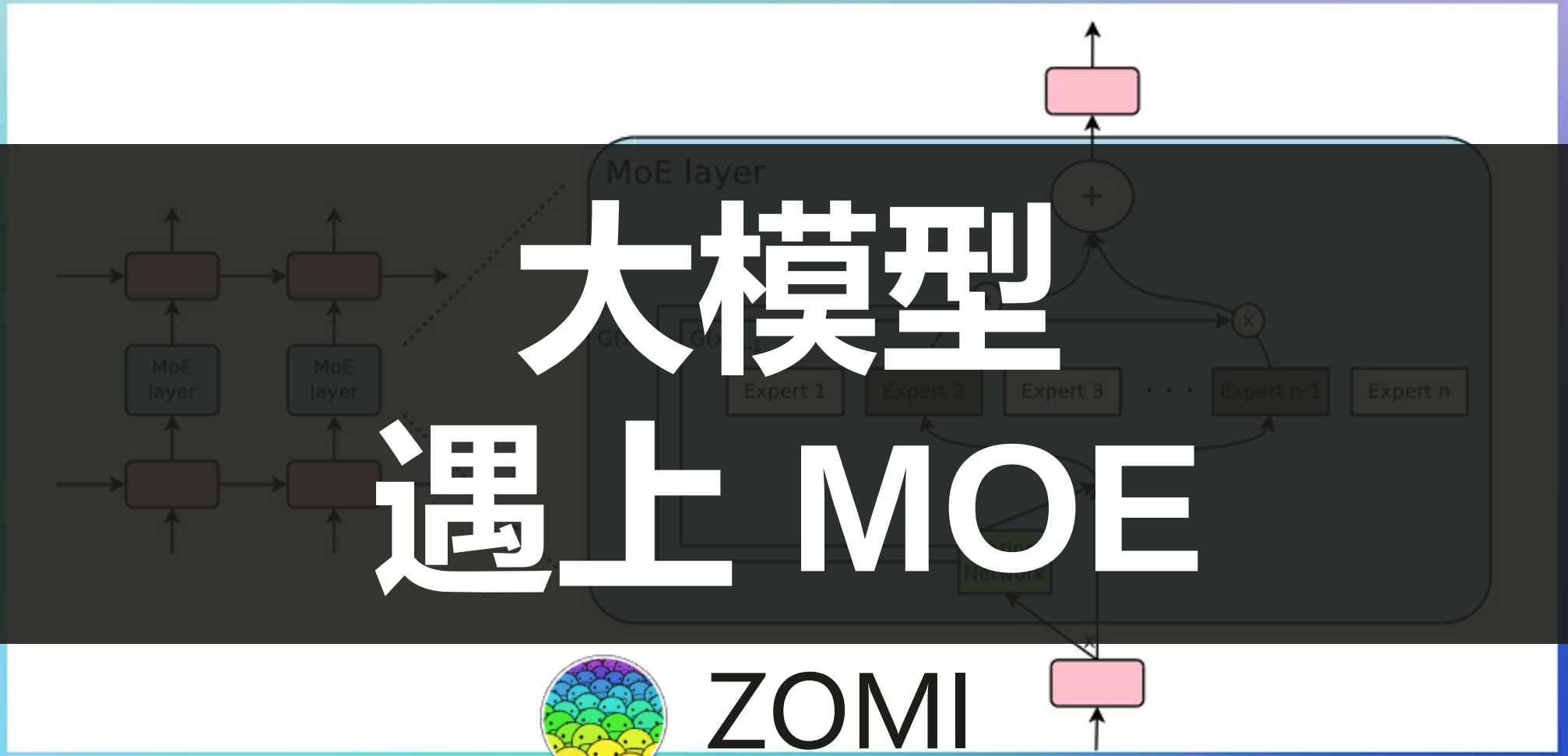
Mixture of Experts (MoE)



大模型遇上 MOE

ZOMI

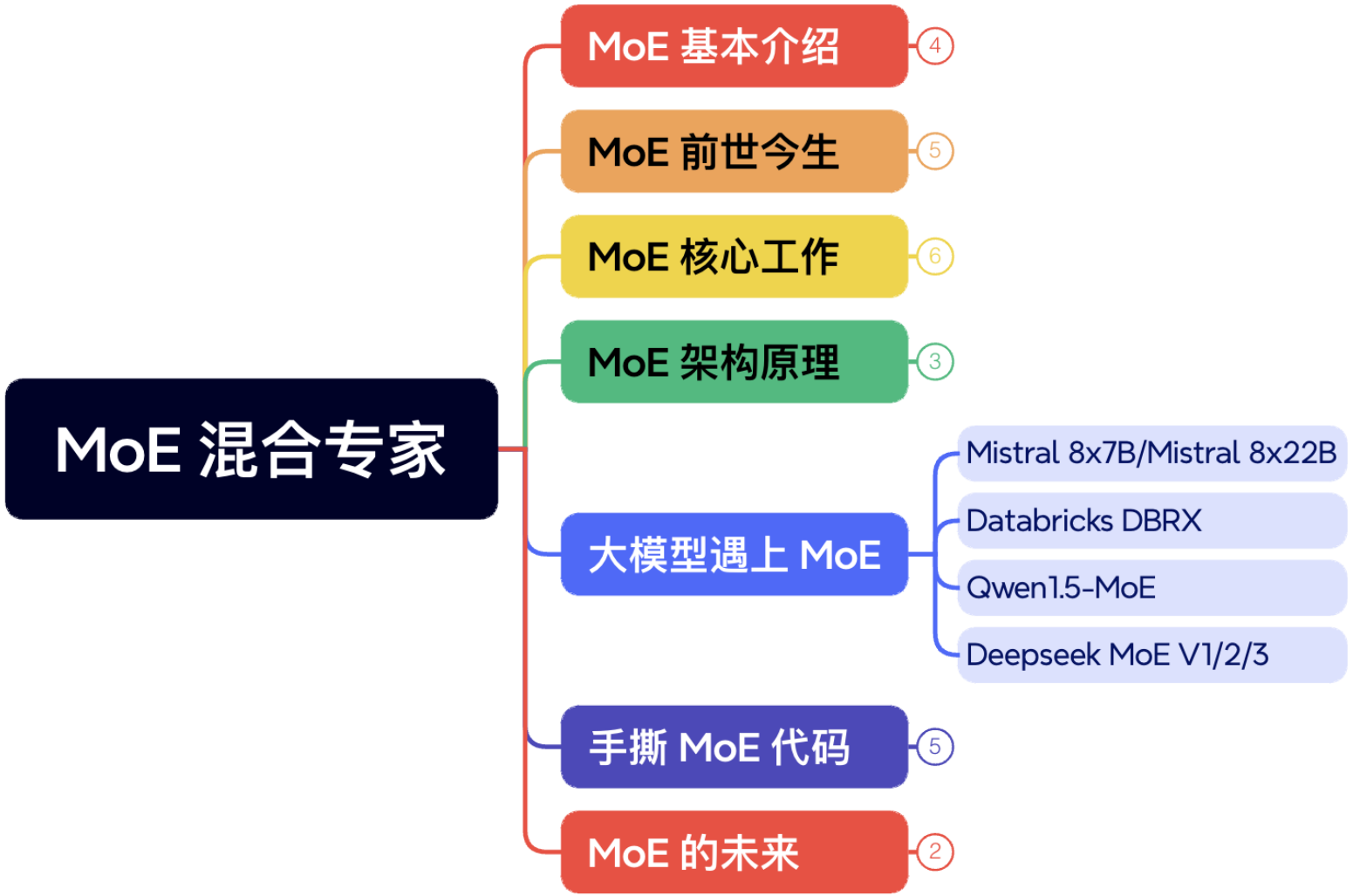
近期发布的 MoE 大模型

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

视频目录大纲

1. Mistral AI
2. Grok-1/2/3
3. DeepSeek 1/2/3
4. 从大参数少专家 to 小参数多专家

视频目录大纲


MoE 混合专家
MoE 基本介绍
4
MoE 前世今生
5
MoE 核心工作
6
MoE 架构原理
3
大模型遇上 MoE
Mistral 8x7B/Mistral 8x22B
Databricks DBRX
Qwen1.5-MoE
Deepseek MoE V1/2/3
手撕 MoE 代码
5
MoE 的未来
2


好好想想

01

Mistral AI

Mistral AI

- Mistral AI是一家成立于2023年初的法国人工智能公司，由前谷歌DeepMind、Meta等科技巨头的研究人员创立，总部位于巴黎。作为欧洲AI领域的代表性企业，Mistral被称为“欧洲版OpenAI”，致力于开发高效、开源的大语言模型（LLMs）和多模态模型，目标是与美国AI巨头竞争。

Mistral AI

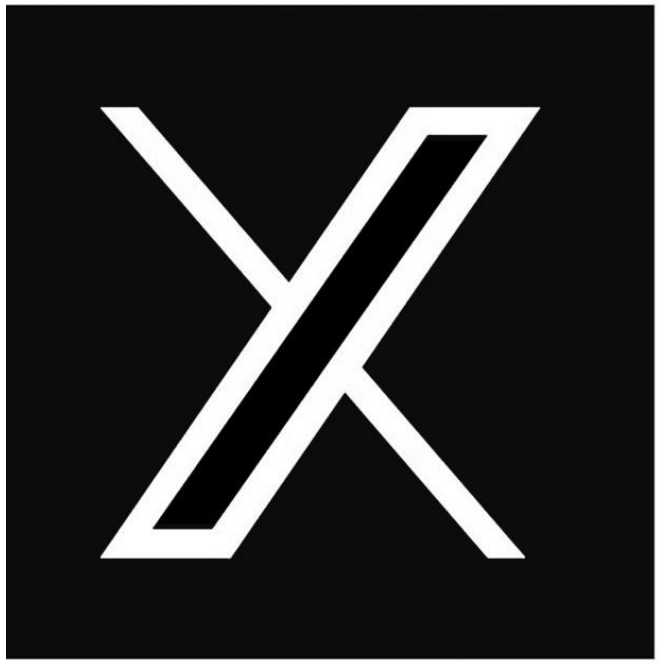
特性	Mixtral-8x22B	Mixtral-8x7B
总参数量	1760 亿参数	46.7 亿参数
激活参数量	约 390 亿参数（稀疏激活）	约 12.9 亿参数（稀疏激活）
专家数量	8 个专家，每个专家 220 亿参数	8 个专家，每个专家 7 亿参数
上下文窗口	64K tokens	32K tokens
多语言支持	英语、法语、意大利语、德语、西班牙语	英语、法语、意大利语、德语、西班牙语
推理效率	比密集 70B 模型更快，成本效率更高	推理速度比 Llama 2 70B 快 6 倍
开源许可	Apache 2.0	Apache 2.0

02

Grok-1/2/3

Grok

- XAI（由埃隆·马斯克创立的人工智能公司）发布的 Grok 系列模型主要基于混合专家模型（Mixture of Experts, MoE）架构。



Grok

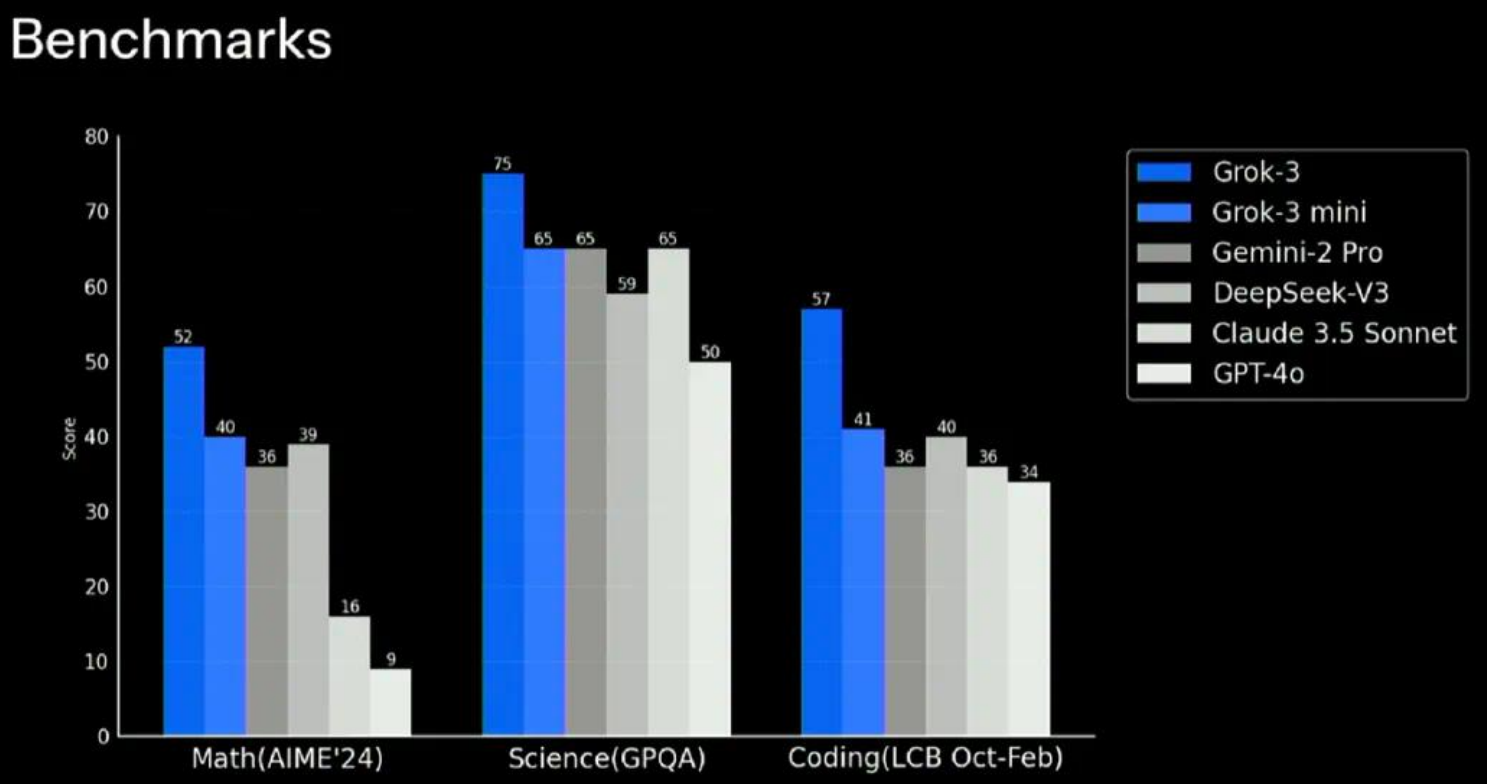
Grok

特性	Grok-1	Grok-2	Grok-3	Grok-3 mini
发布时间	2024年3月	2024年	2025年2月	2025年2月
参数规模	3140 亿参数	未明确，推测更大	未明确，推测更大	未明确，轻量级版本
专家数量	8 个专家，每次激活 2 个	未明确，可能优化路由机制	未明确，可能优化路由机制	未明确，轻量级版本
上下文窗口	8,192 tokens	128,000 tokens	支持多模态任务	未明确，适合实时应用
资源需求	628 GB GPU 显存	未明确，推测更高	20 万块 H100 GPU	未明确，资源需求较低
应用场景	通用语言任务	复杂推理任务	多模态任务（图像、文本）	实时推理任务


INTRODUCING
GROK - 3

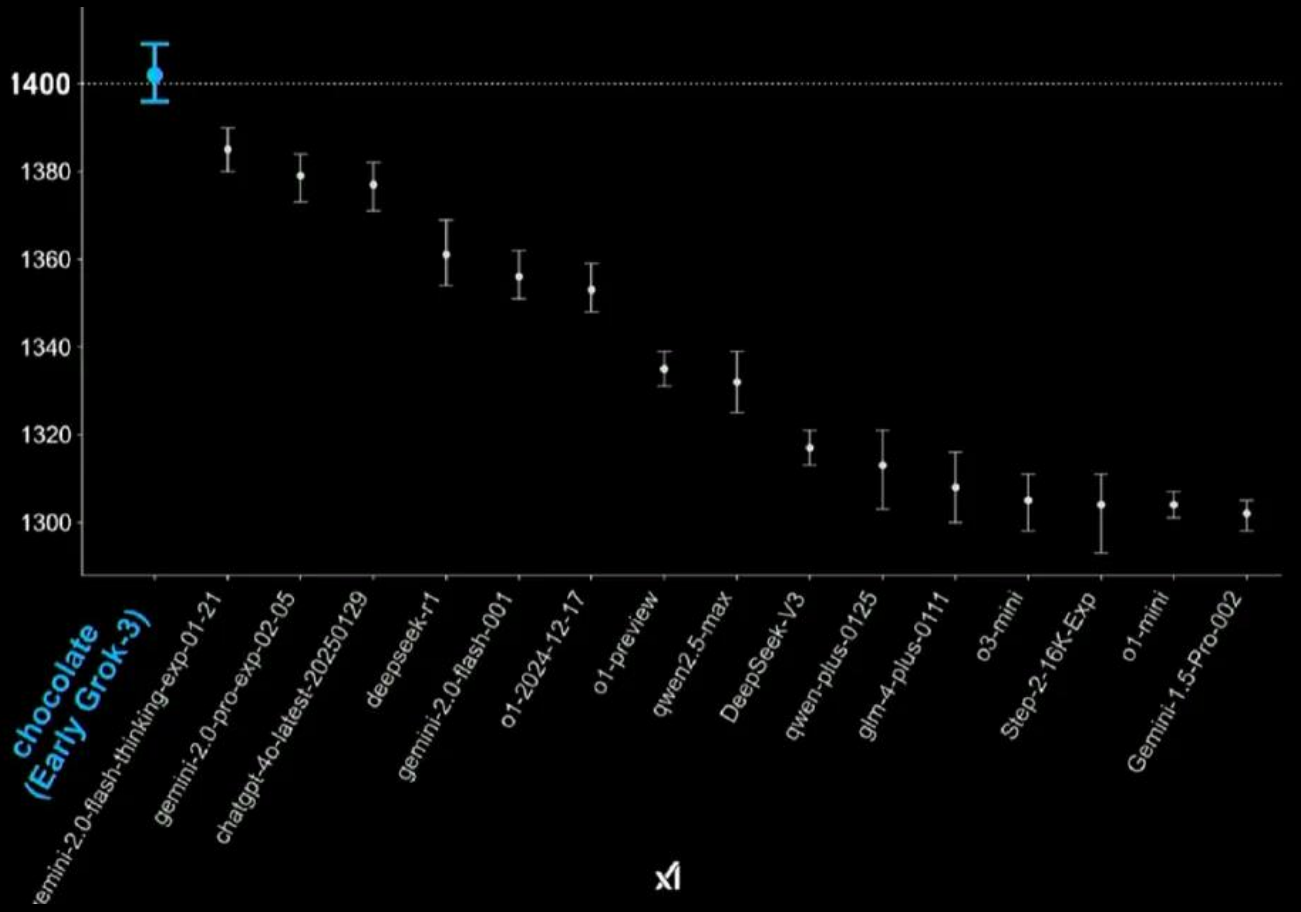
Grok3

- Math（AIME 24）、Science（GPQA）和 Coding（LCB Oct-Feb）三方面，Grok-3 大幅超过 Gemini-2 Pro、DeepSeek-V3、Claude 3.5 Sonnet 和 GPT-4o



Grok3

- Chatbot Arena（LMSYS）中，早期 Grok-3 版本的得分取得了第一，达到 1402 分，超过了包括 DeepSeek-R1 在内的所有其他模型。Grok-3 也成为有史以来首个突破 1400 分的模型



Grok3

- 在编程、数学、创意写作、指令遵循、长查询、多轮对话等场景中的排名情况。

Model	Overall	Overall w/ Style Control	Hard Prompts	Hard Prompts w/ Style Control	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
early-grok-3	1	1	1	1	1	1	1	1	1	1
chatgpt-4o-latest-20250129	2	1	4	3	2	10	1	2	1	1
gemini-2.0-pro-exp-02-05	2	2	1	1	1	2	1	1	1	1
deepseek-r1	5	2	2	1	2	1	4	2	2	1
o1-2024-12-17	5	2	2	1	2	1	5	1	2	5
gemini-2.0-flash-thinking-exp-01-21	2	4	1	1	2	1	1	1	1	1
o1-preview	8	6	5	3	2	1	8	7	6	5
gemini-2.0-flash-001	5	8	4	7	2	1	4	6	4	4
qwen2.5-max	8	8	4	6	5	2	6	7	6	5
claude-3-5-sonnet-20241022	18	8	13	5	11	12	14	12	12	11
deepseek-v3	10	9	13	13	11	12	5	10	6	6
qwen-plus-0125	10	11	10	10	11	10	11	10	6	10
gemini-2.0-flash-lite-preview-02-05	10	11	10	10	10	12	5	10	7	12
o3-mini	11	11	4	3	2	1	17	9	6	11

Question

- 马斯克 的 XA1 使用 20 万张 GPU 训练的 Grok 3：略强于 Deepseek?
- xAI 用了122 天让首批10 万卡集群投入使用，后续又花费 92 天拓展到 20 万卡集群

03

DeepSeek 1/2/3

DeepSeek

- DeepSeek（深度求索）是一家成立于2023年的中国人工智能公司，由量化私募巨头幻方量化创立。公司专注于开发高性能、低成本的大语言模型（LLM），致力于推动通用人工智能（AGI）的发展。
- DeepSeek 的独特之处在于其“基础研究-技术转化-产业应用”三位一体的发展模式，以及通过量化投资业务为AI研发提供持续资金支持的“以战养战”策略。公司不仅在技术上实现了多项突破，还通过开源策略和低成本训练模式，推动了AI技术的普惠化。

DeepSeek

特性	DeepSeek MoE	DeepSeek V2	DeepSeek V3	DeepSeek R1
发布时间	2024 年1 月	2024 年 5 月	2024 年 12 月	2025 年 1 月
参数量	2B/16B/145B	236B	671B	671B
激活参数量	/	21B	37B	37B
架构特点	整合了专家混合系统（MoE）、改进的注意力机制和优化的归一化策略，采用动态路由机制和专家共享机制	基于 MoE 架构，采用多头潜在注意力（MLA）和 DeepSeekMoE 架构，每个 token 激活 21B 参数	基于 MoE 架构，采用多头潜在注意力（MLA）和 DeepSeekMoE 架构，每个 token 激活 37B 参数	/
训练数据量	/	/	14.8 万亿个 token	/
训练成本	/	/	278.8 万 H800 GPU 小时	/
层数	/	60	61	61
隐藏维度	/	5120	7168	7168
中间维度	/	12288	18432	18432
MoE 中间维度	/	1536	2048	2048
共享专家数量	/	2	1	1
路由专家数量	/	160	256	256
token 激活专家	/	6	8	8
词汇表大小	/	102400	129280	129280

思考与小结

早期 MoE 架构：大参数少专家

- 早期 MoE 模型采用少量专家（如 8-16），每个专家参数量较大（如数十亿参数）。
- 专家模型设计复杂，基于稠密 Transformer FFN 层，每个专家需要处理广泛的输入数据。
- 门控网络（Gating Network）设计相对简单，主要通过 softmax 函数分配权重。

- **代表模型：** Mixtral 8x7B： Mistral AI 发布的 MoE 模型，包含 8 个专家，每个专家 7B 参数，通过稀疏激活实现高效推理。

早期 MoE 架构：大参数少专家 Pro

- **提高计算效率：**通过增加专家数量，每个专家的参数量减少，降低单专家的计算复杂度，提高了整体的计算效率。
- **容易训练：**小参数量专家更容易训练，不同专家可以更好地利用计算资源，降低了训练成本和训练的难度，特别在均衡负载。

早期 MoE 架构：大参数少专家 Con

- **计算成本高：**稀疏激活减少了计算量，但每个专家的参数量大，导致内存和显存需求高。
- **专家负载不均衡：**部分专家可能被过度使用，其他专家未被充分利用，影响模型性能。
- **专家利用率低：**少专家使得每个专家需要处理更多任务，导致专家专精化程度不够。

当前趋势：小参数多专家

- 现代 MoE 模型倾向于采用更多专家（128/256），每个专家参数量较小。
- 专家模型更加轻量化，基于细粒度划分的任务或数据特征，不同专家专注于特定领域。
- 门控网络的设计更加复杂，引入了动态路由、负载均衡等技术，优化专家的激活和计算效率。

- **代表模型：**DeepSeek-V3：采用 细粒度专家划分 和 动态路由机制，每个专家参数量较小，但专家数量更多，显著降低了计算成本

当前趋势：小参数多专家 Pro

- **计算效率更高：**小参数多专家减少每次推理计算量，同时通过动态路由优化了专家负载均衡。
- **扩展性更强：**更多专家可以覆盖更广泛任务和数据特征，提升模型的泛化能力和专业化水平。
- **部署成本更低：**小参数专家设计降低了内存和显存需求，适合在资源有限的环境中部署。

大参数少专家 to 小参数多专家

- 本质模型规模与计算效率间寻求平衡， 通过 细粒度专家分工 和 稀疏激活机制， 实现：
  1. **更高参数量：** 提升模型容量上限
  2. **更低推理成本：** 仅激活必要参数，降低单次计算开销
  3. **更强任务适配性：** 覆盖多样化场景，逼近“专家即服务” （Expert-as-a-Service）的理想架构

- 未来，随着路由算法和硬件优化，MoE 模型可能进一步向超大规模小专家集群发展，成为 AGI 的关键路径之一。

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