国泰君安证券 GUOTAI JUNAN SECURITIES
股票研究 / 2024.09.20

MoE 与思维链助力大模型技术路线破局

计算机

评级: 增持

上次评级: 增持

李博伦(分析师) 伍巍(研究助理)
0755-23976516 021-38031029
libolun@gtjas.com wuwei028683@gtjas.com
登记编号 S0880520020004 S0880123070157

细分行业评级

相关报告

计算机《OpenAI o1 开启大模型应用新范式》2024.09.16
计算机《银行间交易自主可控提升行业景气度》2024.09.03
计算机《示范区落地，交通信息化景气度再获验证》2024.07.25
计算机《自主可控迎内外催化，行业有望超预期》2024.07.22
计算机《萝卜快跑带领自动驾驶进入快速落地期》2024.07.14

本报告导读：

Transformer 架构大模型对算力成本要求高，一定程度阻碍了大模型研发和应用的进一步创新，随着 o1 大模型的发布以及 MoE 架构的成熟，技术路线有望破局。

投资要点：

- **投资建议：**随着 Transformer 架构大模型在算力侧成本攀升，升级迭代遇到瓶颈，技术路线相关探索有望打造性能更优、应用范围更专更准的 AI 大模型。推荐标的：科大讯飞、虹软科技、万兴科技、福昕软件、金山办公、鼎捷软件、紫光股份、浪潮信息，受益标的：昆仑万维、润达医疗。
- **巨额算力投入成为技术和效益优化的瓶颈，技术路径破局迫在眉睫。**从效益端看，基于 Transformer 架构的模型在训练计算量（training FLOPs）达到一定量级时，模型性能才出现向上的“拐点”，因此在大模型训练任务中，算力成为必须的基础性资源。但随着模型越来越大，算力成本越来越高，成本飙升源于模型复杂度和数据量攀升对计算资源的需求。Anthropic 首席执行官表示，三年内 AI 模型的训练成本将上升到 100 亿美元甚至 1000 亿美元。巨额的大模型训练投入一定程度减缓了技术进步和效益提升，因此技术路径破局尤为关键。当前 MoE 以及 OpenAI o1 的“思维链”是重要探索实践。
- **MoE 框架是对 Transformer 架构的优化，关键在于路由策略及微调。**其能在不给训练和推理阶段引入过大计算需求的前提下大幅提升模型能力。在基于 Transformer 的大型语言模型（LLM）中，每个混合专家（MoE）层的组成形式通常是*N*个“专家网络”搭配一个“门控网络”G。门控函数（也被称路由函数）是所有 MoE 架构的基础组件，作用是协调使用专家计算以及组合各专家的输出。根据对每个输入的处理方法，该门控可分为三种类型：稀疏式、密集式和 soft 式。其中稀疏式门控机制是激活部分专家，而密集式是激活所有专家，soft 式则包括完全可微方法，包括输入 token 融合和专家融合。MoE 在 NLP、CV、语音识别以及机器人等领域表现出色，且在更高性能的大模型推理芯片 LPU 加持下，MoE 模型提升效果显著。
- **OpenAI o1 基于“思维链”的创新推理模式，学会人类“慢思考”，专业领域的效果突出。**OpenAI o1 相比之前的 AI 大模型最跨越性的一步在于拥有人类“慢思考”的特质：系统性、逻辑性、批判性、意识性。在响应用户提出的难题之前，OpenAI o1 会产生一个缜密的内部思维链，进行长时间的思考，完善思考过程、意识逻辑错误、优化使用策略、推理正确答案。这种深度思考能力在处理数学、编程、代码、优化等高难度问题时发挥重要作用，能够进行博士级别的科学问答，成为真正的通用推理。推理侧的应用模式创新有望在更为专业的领域创造价值应用，从通用的偏娱乐领域逐步过渡到偏严肃的专业领域场景，AI 大模型的真正实践价值有望进一步释放，因此 o1 模型提供的新应用范式和能力维度在大模型技术路线演绎中，具有里程碑意义。
- **风险提示：**技术迭代不及预期，AI 应用市场拓展节奏不及预期。

目 录

1. 投资建议

投资建议：随着 Transformer 架构大模型在算力侧成本攀升，升级迭代遇到瓶颈，技术路线相关探索有望打造性能更优、应用范围更专更准的 AI 大模型。推荐标的：科大讯飞、虹软科技、万兴科技、福昕软件、金山办公、鼎捷软件、紫光股份、浪潮信息，受益标的：昆仑万维、润达医疗。

表 1：推荐标的盈利预测

股票代码	股票名称	股价（元）	EPS（元/股）			PE（倍）			评级
		2024/9/18	2023A	2024E	2025E	2023A	2024E	2025E	
002230.SZ	科大讯飞	33.82	0.28	0.39	0.43	120.79	86.72	78.65	增持
688088.SH	虹软科技	23.86	0.22	0.36	0.48	108.45	66.28	49.71	增持
300624.SZ	万兴科技	40.42	0.64	0.51	0.57	63.16	79.25	70.91	增持
688095.SH	福昕软件	49.73	-1.03	-0.33	0.18	--	--	276.28	增持
688111.SH	金山办公	181.08	2.86	3.72	4.86	63.31	48.68	37.26	增持
300378.SZ	鼎捷软件	16.33	0.56	0.68	0.85	29.16	24.01	19.21	增持
000938.SZ	紫光股份	18.42	0.74	0.87	1.02	25.06	21.17	18.06	增持
000977.SZ	浪潮信息	31.16	1.1804	1.7	1.98	26.40	18.33	15.74	增持

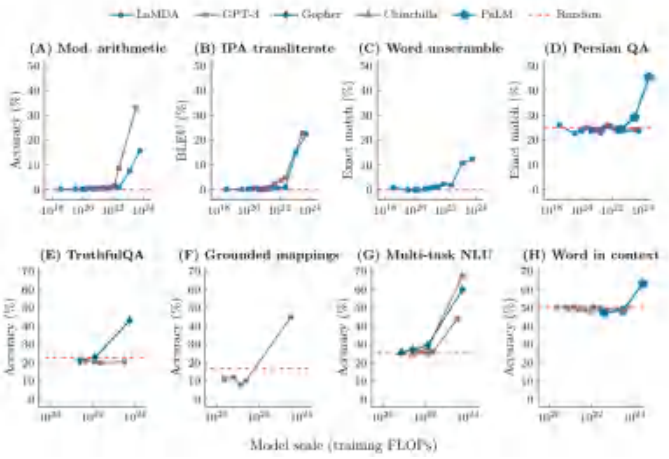
数据来源：国泰君安证券研究（以上公司盈利预测数据均来自国泰君安证券盈利预测）

2. MoE 另辟蹊径，有望破局大模型发展瓶颈

2.1. Transformer 路径下，算力资源成为大模型发展的瓶颈

Transformer 模型通过扩大计算量（training FLOPs）达到更好的性能，算力是重要资源。2020 年 OpenAI 的 GPT-3 开启大模型时代，此后多家公司快速发布基于 Transformer 架构的大模型。从架构机制看，Transformer 架构有以下特点：（1）自注意力机制：需要计算每个词与其他所有词之间的相关性，这种计算复杂度随着输入序列长度的增加而呈平方增长，（2）多头注意力机制：为了捕捉不同方面的语义信息，Transformer 模型通常使用多头注意力机制，这进一步增加了计算复杂度。（3）层数深：Transformer 模型通常具有很多层，每层都需要进行大量的矩阵运算，因此模型需要用到大量算力资源。从效益端看，基于 Transformer 架构的模型在训练计算量（training FLOPs）达到一定量级时，模型性能才出现向上的“拐点”，因此在大模型训练任务中，算力成为必须的基础性资源。

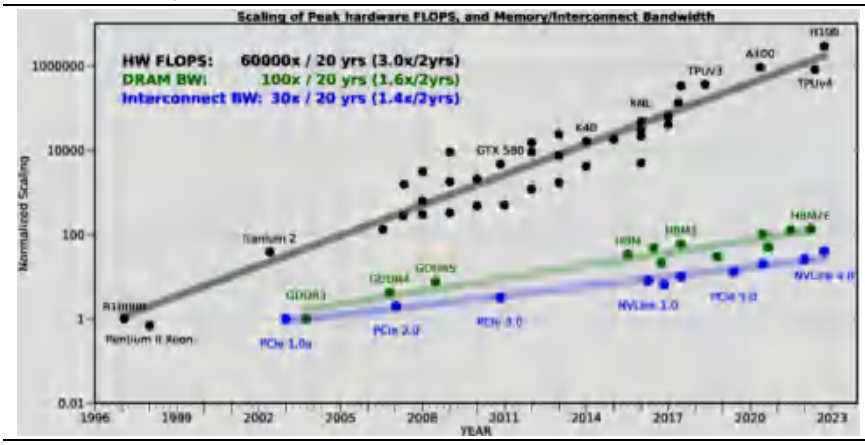
图1：随着规模增加，模型能力呈现“涌现”现象



数据来源：《Emergent Abilities of Large Language Models》

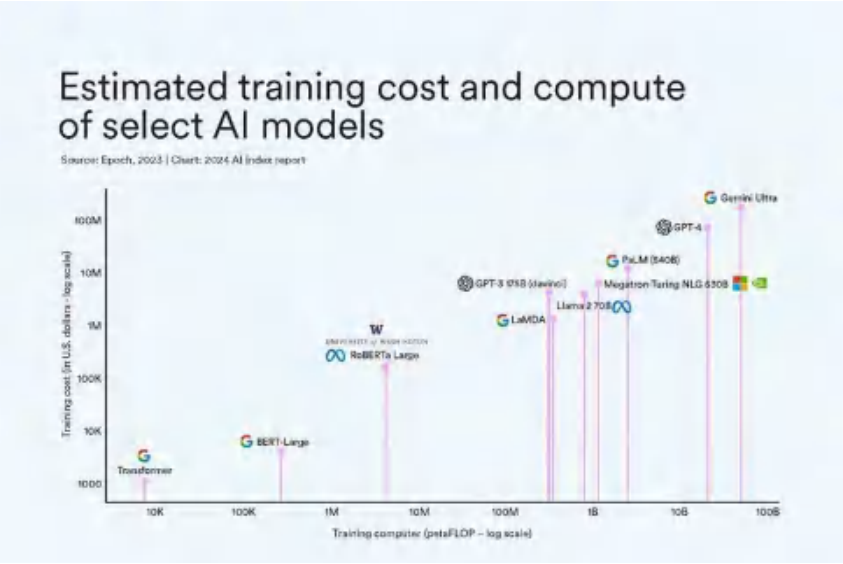
模型越来越大，算力成本越来越高。在算力方面，AI 模型公司通常购买算力卡，再使用不同的计算框架和算法等部署大模型的计算集群。从算力供给看，英伟达算力卡需求较大，当前英伟达 GPU 的拿货能力已经成为大模型公司的核心能力；从需求端看，Transformer 类架构模型运算量每两年约翻 750 倍，远超 CV/NLP/Speech 模型算力需求增长，更远远超过摩尔定律的速度。随着时间增长，模型运算量增长和芯片性能增长之间的巨大差距需要更好的下游集群策略来弥补，但集群策略将会越来越复杂，花费在算力上的成本也将更高。据第一财经透漏，2023.09-2023.11 中贝通信向客户提供 AI 算力技术服务的单价两个月内上涨了 50%。Anthropic 首席执行官 Dario Amodei 在 In Good Company 播客节目中表示，目前正在开发的人工智能模型的训练成本高达 10 亿美元，且预计从现在开始，三年内 AI 模型的训练成本将上升到 100 亿美元甚至 1000 亿美元。

图2：大模型算力需求增长过快



数据来源：《AI and Memory Wall》

图3：大模型训练成本呈指数级攀升



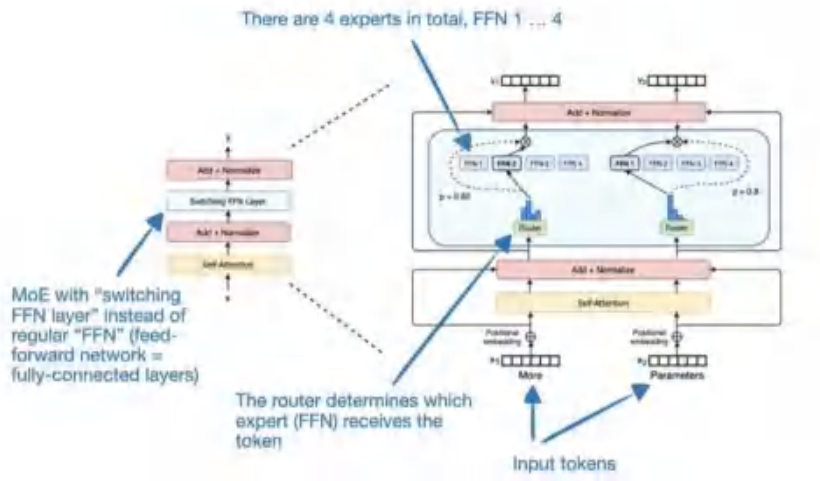
数据来源：Stanford 2024 AI Index Report

成本飙升源于模型复杂度和数据量攀升对计算资源的需求，因此行业应用需求的多样性可以一定程度减弱短期内对基础性大模型的需求。在 AI 领域，这些成本增长带来的影响各不相同，因为并非所有应用场景都需要最新、最强大的大语言模型。随着众多小型大语言模型替代品的涌现，如 Mistral 和 Llama 3，它们有数十亿个参数，不像 GPT-4 可能具有万亿个参数。且微软也发布了自己的小语言模型（SLM）Phi-3，Phi-3 拥有 38 亿个参数，并且基于相对 GPT-4 等大语言模型更小的数据集进行训练。小模型尽管可能无法完全媲美大型模型的效能，但小语言模型凭借其精简的体型和训练数据集，在成本控制方面展现出独特优势。小型、专业化的语言模型如同庞杂系统中的重要组件，为各类细分应用提供关键高效的功能，因此 Moe 路径或是当前大模型训练及应用演化的重要参考路径之一。

2.2. MoE 框架是对 Transformer 架构的优化，而非完全替代

MoE 框架基于一个简单却又强大思想：模型的不同部分（称为专家）专注于不同的任务或数据的不同方面。MoE（Mixture of Experts）类模型使用远少于 Transformer 架构类模型的算力扩大模型规模，性价比更高。MoE 模型架构起源于 1991 年，2017 年 google 提出《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》中，MoE 模型逐渐被关注。2024 年 1 月，OpenAI 团队发布 Mixtral 8x7B 的论文，MoE 模型成为关注焦点。传统的 Transformer 架构主要包括自注意力层和前馈网络层，MoE 模型使用稀疏 MoE 层代替传统 Transformer 架构中的前馈网络(FFN)，相比于 Transfromal 前馈层的全连接，MoE 架构的层连接更加稀疏，因此也被称为稀疏模型。

图4：专家层代替 Transformer 模型中的 FFN 层



数据来源：《Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity》

MoE 层主要包括两部分，即门控层和专家层。门控 Gatenet 层用于决定哪些 token 被发送到哪些最相关的专家。门控的输出可以用来解释模型的决策过程，分析哪些专家对特定输入的贡献最大，可解释性更高。门控层的算法决定不同专家的启用情况，不同的算法会带来不同的负载均衡，同时影响模型的稀疏程度。专家 experts 层通常包括多个专家网络，每个专家本身是一个独立的神经网络，可以被独立设计和训练，负责处理来自门控 getenet 层分配的不同数据，针对特定的任务优化参数，更好地处理特定领域的任务。

图5：门控 gatenet 层和专家 experts 层分工流程不同

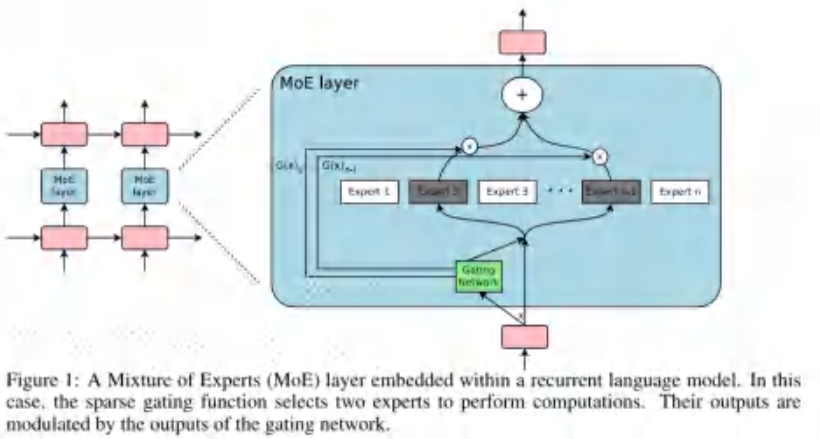


Figure 1: A Mixture of Experts (MoE) layer embedded within a recurrent language model. In this case, the sparse gating function selects two experts to perform computations. Their outputs are modulated by the outputs of the gating network.

数据来源：《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》

MoE 模型测试表现出色，成本更低。较早的 MoE 模型 Mixtral 8x 在多项测试中表现超过 Transformer 架构模型 Llama，而激活参数仅为 13B，为 Llama 2 70B 的五分之一，大幅降低了计算量。OpenAI、谷歌、微软等都发布了 MoE 类自研架构的大模型，选择拥抱 MoE 模型。国内 DeepSeek 在 2024 年 5 月发布的第二代自研架构 MoE 模型 DeepSeek-V2，在中文能力、英文能力、知识等多项测试中位居前列，在发布的测试中，DeekSeek-V2、GPT-4、Genimi1.5Pro 等 MoE 模型表现亮眼。

图6：DeepSeek-V2 发布能力测试

Model	是否开源	中文综合 AlignBench	英文综合 MT-Bench	知识 MMLU	基础算数 GSM8K	数学解题 MATH	逻辑推理 BBH	编程 HumanEval
DeepSeek-V2	✓	7.91	8.97	77.8	92.2	53.9	79.7	81.1
GPT-4-Turbo-1106	✗	8.01	9.32	84.6	93.0	64.1	-	82.2
GPT-4-0613	✗	7.53	8.96	86.4	92.0	52.9	83.1	84.1
GPT-3.5	✗	6.08	8.21	70.0	57.1	34.1	66.6	48.1
Gemini 1.5 Pro	✗	7.33	8.93	81.9	91.7	58.5	84.0	71.9
Claude 3 Opus	✗	7.62	9.00	86.8	95.0	61.0	86.8	84.9
Claude 3 Sonnet	✗	6.70	8.47	79.0	92.3	40.5	82.9	73.0
Claude 3 Haiku	✗	6.42	8.39	75.2	88.9	40.9	73.7	75.9
abab-6.5（MiniMax）	✗	7.97	8.82	79.5	91.7	51.4	82.0	78.0
abab-6.5s（MiniMax）	✗	7.34	8.69	74.6	87.3	42.0	76.8	68.3
ERNIE-4.0（文心一言）	✗	7.89	7.69	-	91.3	52.2	-	72.0
GLM-4（智谱清言）	✗	7.88	8.60	81.5	87.6	47.9	82.3	72.0
Moonshot-v1（月之暗面）	✗	7.22	8.59	-	89.5	44.2	-	82.9
Baichuan 3（百川）	✗	-	6.70	81.7	88.2	49.2	84.5	70.1
Qwen1.5 72B（通义千问）	✓	7.19	8.61	76.2	81.9	40.6	65.9	68.9
LLaMA 3 70B	✓	7.42	8.95	80.3	93.2	48.5	80.1	76.2
Mixtral 8x22B	✓	6.49	8.66	77.8	87.9	49.8	78.4	75.0

数据来源：DeepSeek 公众号

2.3. MoE 基于门控函数设计方式可分为多种类型

在基于 Transformer 的大型语言模型（LLM）中，每个混合专家（MoE）层的组成形式通常是*N*个“专家网络”$\{f_1,\ldots,f_N\}$搭配一个“门控网络”G。门控函数（也被称为路由函数或路由器）是所有 MoE 架构的基础组件，其作用是协调使用专家计算以及组合各专家的输出。根据对每个输入的处理方法，该门控可分为三种类型：稀疏式、密集式和 soft 式。其中稀疏式门控机制是激活部分专家，而密集式是激活所有专家，soft 式则包括完全可微方法，包括输入 token 融合和专家融合。这个门控网络的形式通常是一个使用 softmax 激活函数的线性网络，其作用是将输入引导至合适的专家网络。MoE 层的放置位置是在 Transformer 模块内，作用是选取前向网络（FFN），通常位于自注意力（SA）子层之后。这种放置方式很关键，因为随着模型增大，FFN 的计算需求也会增加。例如在参数量达到 5400 亿的 PaLM 模型中，90%的参数都位于其 FFN 层中。

图7：MoE 模型中使用各种门控函数

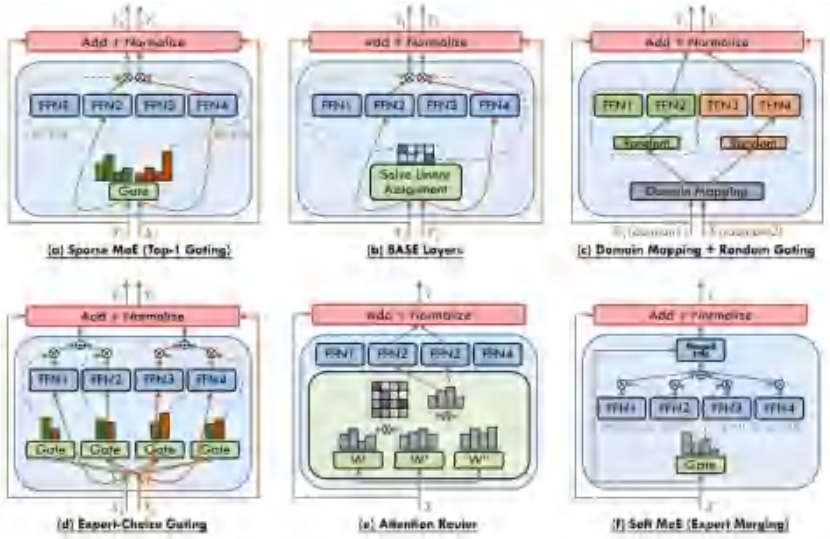


Fig. 4. The illustration of various gating functions employed in MoE models, including (a) sparse MoE with top-1 gating [40], (b) BASE layers [87], (c) the combination of grouped domain mapping and random gating [121], (d) expert-choice gating [193], (e) attention router [166], and (f) soft MoE with expert merging [165].

数据来源：《A Survey on Mixture of Experts》

根据门控函数的设计方式，稠密和稀疏两种方式各有侧重。密集混合专家层是在每次迭代过程中激活所有专家网络 $\{f\_1, \ldots, f\_N\}$。早期的 MoE 研究普遍采用了这一策略。尽管密集混合专家的预测准确度通常更高，但其计算负载也非常高。为了解决这个问题，Shazeer et al.的论文《Outrageously large neural networks: The sparsely-gated mixture-of-experts layer》引入了稀疏门控 MoE 层，开创性地提出了一种使用辅助负载平衡损失的可微分启发式方法，其中可根据选取概率对专家计算的输出进行加权。这为门控过程引入了可微性，由此可通过梯度来引导门控函数的优化。后来，这一范式便成了 MoE 研究领域的主导范式。其能在每次前向通过时仅激活选定的专家子集，该策略实现稀疏性的方式是计算 top-k 个专家的输出的加权和，而非将所有专家的输出聚合到一起。但稀疏门控也有专家负载均衡分布均匀度的问题，例如稀疏门控情况下，某些专家被频繁使用，而另一些专家则很少被调用。为了解决这个问题，每个 MoE 层都要集成一个辅助损失函数，其作用是敦促每批次的 token 被均匀分配给各个专家，比如针对一个包含 T 个 token 的查询批次以及 N 个专家，为了确保该批次在 N 个专家之间均匀分布，应当最小化负载平衡损失函数，当每个专家都被分配了同等数量的 token 和同等的门控概率时，即达到最优条件，此时各专家的负载达到平衡。由于这种方法会针对每个输入 token 选择专家，因此可将其看作是 token 选择式门控函数。

图8：稀疏式一大基本离散优化难题是如何为每个 token 分配合适专家

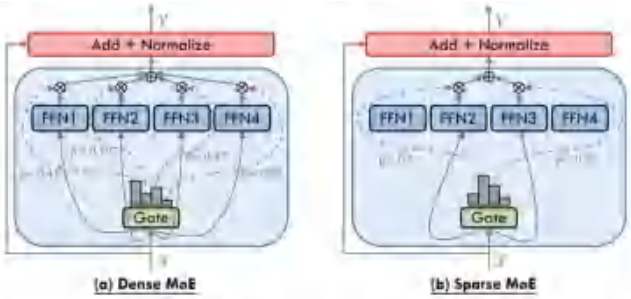


Fig. 2. An illustration of an MoE layer in Transformer-based models. For each input X, the linear-softmax gating will select all experts namely **(a) Dense MoE** or top k experts namely **(b) Sparse MoE** to perform conditional computation. The expert layer returns the output of the selected expert multiplied by the gate value (softmax of the gating function output).

数据来源：《A Survey on Mixture of Experts》

虽然稀疏 MoE 有效率方面的优势，但密集 MoE 方向依然在不断迎来创新。比如二元决策、稀疏或连续决策、随机或确定性决策，其都已经得到了深入的研究，可使用各种形式的强化学习和反向传播来训练。密集激活在 LoRA-MoE 微调方面表现很好，并且 LoRA 专家的计算开销相对较低。这种方法能够有效灵活地集成多个 LoRA 以完成各种下游任务。这能保留原始预训练模型的生成能力，同时保留各个 LoRA 针对各个任务的独有特性。类似于密集 MoE，soft MoE 方法在处理每个输入时也会使用所有专家，从而维持完全可微性，进而避免离散专家选择方法的固有问题。soft MoE 与密集 MoE 的不同在于前者会通过对输入 token 或专家进行门控加权的融合来缓解计算需求。

2.4. MoE 模型效益的关键在于路由策略及微调

路由策略是指门控层应用某种机制（学习算法）决定哪些数据分配给哪些专家，理想的分配策略要达到负载均衡状态。负载均衡状态是指不存在某些专家被分配过量数据、专家过度拟合，也不存在某些专家缺乏训练数据，导致模型精度低和计算资源浪费。因此路由策略的改良可以大幅提升整个模型的最终效果，降低模型的算力成本。路由策略正在持续研究更新中，例如清华 SmartMoE 设计了专家放置（Expert Placement）策略，实现了动态负载均衡；OpenAI 在 GPT-4 中选择加载 16 个专家和前向通道 2 个路由的设计；微软 MH-MoE 使用多头机制将 token 拆分为多个子 token，达到了更好的负载均衡，国内 DeepSeek 团队自研 MLA 架构，大幅减少计算量和推理显存，测试效果良好。

适配最佳容量因子和 top-n 的硬件-软件系统提高模型效率，降低成本。在门控路由分配 token 的过程中，容量因子和 top-n 是关键：容量因子(Expert Capacity,CF)衡量每个专家处理 token 的数量，top-n 是指每个 token 最多由 n 个专家处理；增加训练和评估容量因子会提高模型质量，增加 top-n 也可以小幅增益模型质量；同时，增加容量因子和使用 top-n 更高的算法会提高计算量，激活内存和通信成本，因此具体的硬件-软件系统决定最佳的 n 和容量因子。

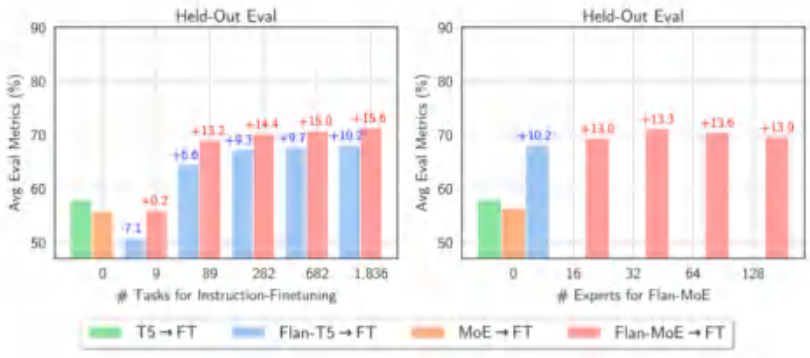
图9：容量因子和 top-n 对模型质量的影响

Algorithm	Train CF	Eval CF	Neg. Log Perp. (↑)
Dense-L	—	—	-1.474
Dense-XL	—	—	-1.384
Top-1	0.75	0.75	-1.428
Top-1	0.75	2.0	-1.404
Top-2	0.75	0.75	-1.424
Top-2	0.75	2.0	-1.402
Top-1	1.0	1.0	-1.397
Top-1	1.0	2.0	-1.384
Top-2	1.0	1.0	-1.392
Top-2	1.0	2.0	-1.378
Top-1	1.25	1.25	-1.378
Top-1	1.25	2.0	-1.373
Top-2	1.25	1.25	-1.375
Top-2	1.25	2.0	-1.369
Top-2	2.0	2.0	-1.360
Top-2	2.0	3.0	-1.359
Top-3	2.0	2.0	-1.360
Top-3	2.0	3.0	-1.356

数据来源：《ST-MoE: Designing Stable and Transferable Sparse Expert Models》

稀疏模型更需要超参数微调，指令式微调效果超过稠密模型。稀疏模型更易于出现过拟合现象，因此需要通过降低学习率、调大批量、权重冻结、更高内部正则化等超参数调整方式微调模型，关于超参数微调的研究动态也在不断变化。另一方面，MoE 模型的指令式微调效果超过稠密模型，《MoEs Meets instructions Tuning》研究发现 Flan-MoE 相比原始 MoE 的性能提升幅度超过了 Flan T5 相对于原始 T5 的提升，MoE 模型将在应用端表现出更高的灵活性。

图10：MoE 模型的指令微调性能提升更高



数据来源：《Mixture-of-Experts Meets Instruction Tuning: A Winning Combination for Large Language Models》

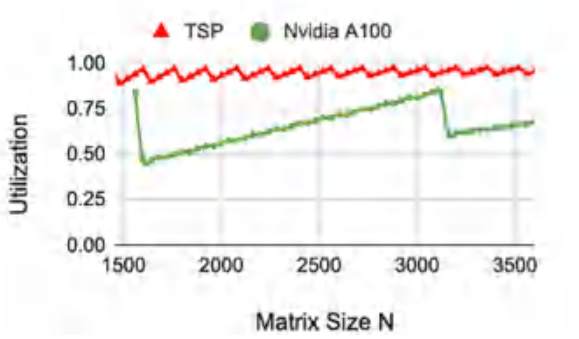
2.5. 针对 MoE 模型训练中的性能问题，LPU 设计厂商有望破局

GPU 在 AI 训练中效率低下。GPU 同时处理多个相同任务，而 MoE 模型每

个专家分支独立处理数据；同时，MoE 中专家需要频繁地交换信息，增加通信成本，降低性能。AI 新智界报道 GPT-4 使用 A100 GPU 和张量并行策略训练模型，算力利用率约为 32%-36%，效率较低，单 GPT-4 一次训练的成本约 6300 万美元，成本极高。英伟达 A100 GPU 目前仍是 AI 训练和推理最广泛应用的 GPU，目前售价和租赁价格仍居高不下。

新兴厂商 Grop 推出更高性能的大模型推理芯片 LPU，MoE 模型效果提升效果显著。Groqd 的 LPU 采用新的设计，与英伟达 A100 相比单元计算利用率更高，且更稳定。Groq 公开 LPU 产品测试中，测试推理的输出速度比谷歌 Gemini 快 10 倍，比 GPT-4 快 18 倍，在 AI 大模型推理中效果提升显著。同时 LPU 已经适配 MoE 模型：Mixtral 8x7B-32k。

图11：LPU 设计更高的单元计算利用率



数据来源：《A Software-defined Tensor Streaming Multiprocessor for Large-scale Machine Learning》

图12：LPU 更快的吞吐速度

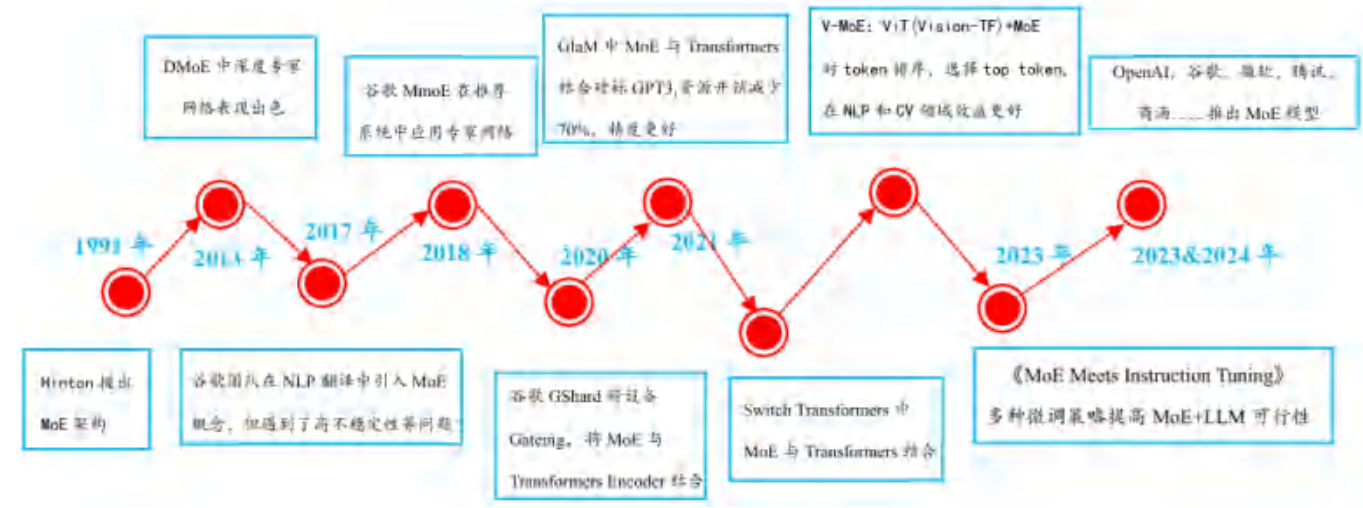
数据来源：groq 官网

3. MoE 模型降本增效，应用广泛

3.1. MoE 模型在多个赛道表现优异

MoE 框架由来已久，目前已经成为大模型赛道焦点之一。1991 年《Adaptive Mixture of Local Experts》中使用多个单独网络（专家）组成的系统建立一个监管机制，首次提出 MoE 概念；2017 年 Google Brain 团队谷歌将 MoE 引入 NLP，在保持模型高规模的同时实现了快速的推理速度，但也面临稀疏模型高通信成本和训练不稳定性等多项挑战；《MoE Meets Instruction Tuning》提出多种微调策略，提高了 MoE+LLM 的可行性；随后 MoE 模型在不到一年的时间内被广泛应用，2023 年 12 月，Mistra AI 在发布了首个开源 MoE 模型，随后 OpenAI、谷歌、微软、字节跳动等大厂都选择拥抱 MoE 框架，推出自研架构的 MoE 模型，国内昆仑万维、幻方量化、新旦智能、元象科技等大模型新宠快速加入，MoE 被市场广泛关注。

图13：MoE 架构历史沿革



数据来源：国泰君安证券研究

MoE 在 NLP 领域表现出色，已经在 NLP 领域广泛使用。2017 年《Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer》将 MoE 的概念引入 LSTM，模型在保持极高规模的同时实现了快速的推理速度，在翻译工作中表现出色；2021 年《Scaling Vision with Sparse Mixture of Experts》中 V-MoE 在 NLP 领域表现出出色,大幅度降低了推理成本。在目前主流的大模型如 GPT-4、Genimi 1.5 pro、天工 3.0 语言大模型等都使用了 MoE 框架，MoE 已经成为大语言模型中重要的方法论。从终端 AI Agent 看，萨曼莎 AI 应用了 MoE 技术，终端 Agent 应用于机器人客服，已经开始提供正式服务；医者 AI 也应用了 MoE 架构，目前终端 Agent 在体检和家庭医生两个场景提供服务。

图14：萨曼莎 AI 应用 MoE 技术

数据来源：萨曼莎官网

MoE 在 CV 领域表现出色，在研究和应用中潜力巨大。2021 年《Scaling

Vision with Sparse Mixture of Experts》的精度和算力成本测试中，在相同的算力成本下 MoE 架构具有更好的表现，另外提到的 BPR 算法优化后的模型表现更好。2023 年，《Mod-Squad: Designing Mixture of Experts As Modular Multi-Task Learners》中的 Mod-Squad 将 MoE 引入 Vision Transformer (ViT)，模型在 13 个视觉任务的 Taskonomy 大数据集和 PASCALContext 数据集上取得了最佳效果。

图15：图像识别：相同算力资源下 MoE 模型更加精准

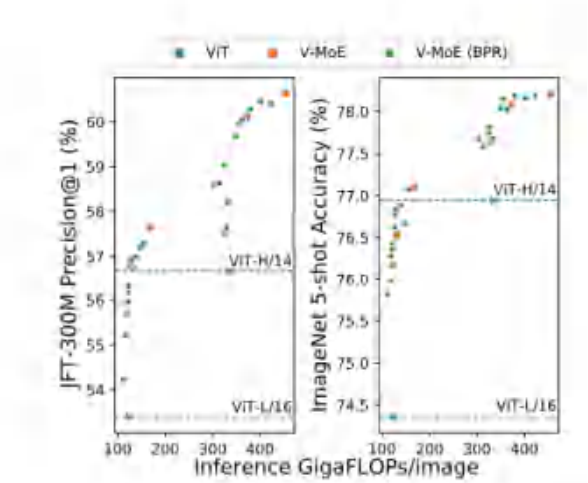


Figure 5: **Reducing compute with priority routing.** Performance vs. *inference* FLOPs for large models. V-MoEs with the original vanilla routing are represented by •, while ■ shows V-MoEs where BPR and a mix of $C \in \{0.6, 0.7, 0.8\}$ and $k \in \{1, 2\}$ are used to reduce compute. ViT models shown as x.

数据来源：《Scaling Vision with Sparse Mixture of Experts》

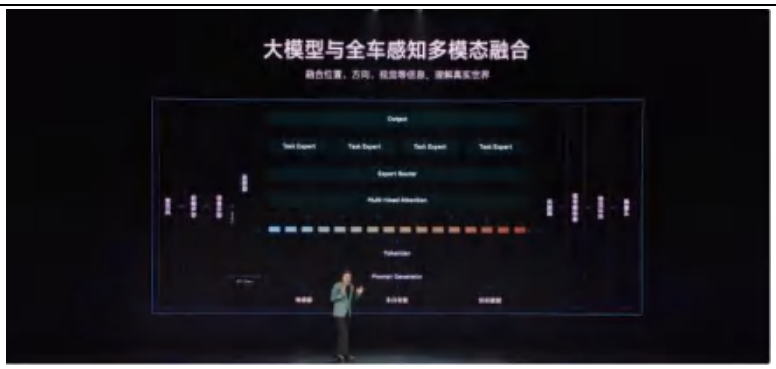
MoE 在语音识别领域表现出色。《BA-MoE: Boundary-Aware Mixture-of-Experts Adapter for Code-Switching Speech Recognition》设计了 BA-MoE 框架，最终将混合错误率（MER）降低到 8.08%，在混合语音识别中表现更好。天工 AI 智能助手应用 MoE 框架，在语音识别上表现出色，小米 SU7 小爱同学的多模态工作使用的商汤模型也应用了 MoE 框架。可见 MoE 在商用及市场拓展中进展迅速。

图16：语音识别：MoE 架构降低 MER

Model	Params(M)	MER(%)	CER(%)	WER(%)
LAE-Conformer [19]	138	8.9	7.3	27.7
Attention Module [14]	112	8.57	6.68	24.11
BA-MoE	125	8.30	6.46	23.26
+ LM	-	**8.08**	**6.28**	**22.78**

数据来源：《BA-MoE: Boundary-Aware Mixture-of-Experts Adapter for Code-Switching Speech Recognition》

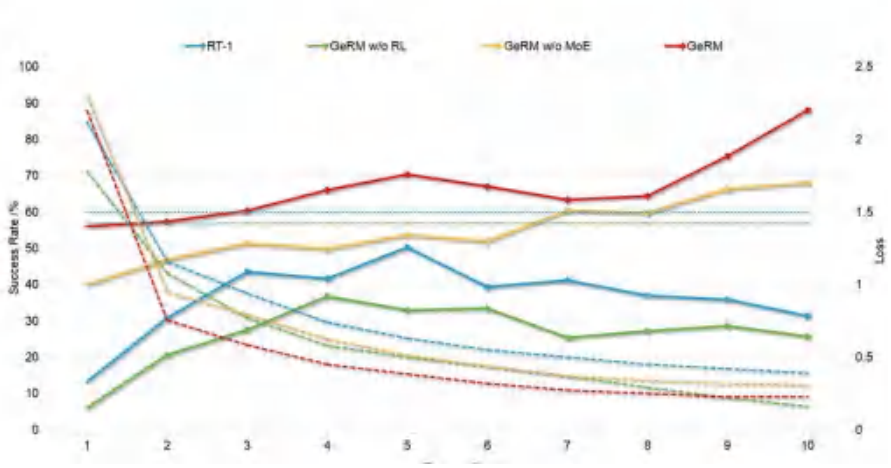
图17：小米应用商汤大模型



数据来源：小米发布会

MoE 赋予机器人更强的性能，节约更多的算力成本。机器人领域视觉-语言-动作多模态模型是大模型落地的绝佳场景，《GeRM: A Generalist Robotic Model with Mixture-of-experts for Quadruped Robot》提出用于四足强化学习的基于 MoE 架构的 GeRM 通用机器人模型，在 99 个任务中相比不使用 MoE 架构的模型，表现出更低参数阈值的涌现能力，提高性能的同时更加节省算力成本。基于 MoE 架构的更多变式的多模态通用机器人模型有望在未来展现出更高的性能，节省更多的成本，带动通用机器人行业更快速成长。

图18：基于 MoE 架构的 GeRM 有低阈值的涌现能力

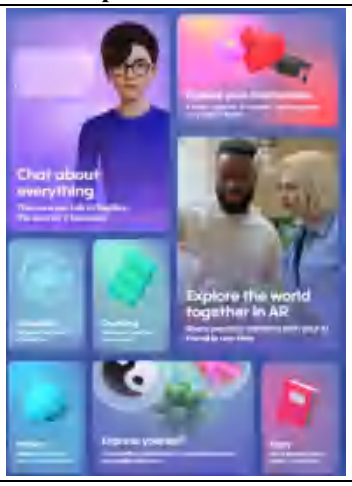


数据来源：《GeRM: A Generalist Robotic Model with Mixture-of-experts for Quadruped Robot》

MoE 架构助力 AI 更快打破游戏行业的不可能三角。游戏集合了美术、动化、文字，因为等多模态内容为一体，开发流程复杂，难以实现“成本、质量、速度”的不可能三角,大模型有望打破这一三角：（1）角色生成：Replika，Character.AI、AI Dungeon 等基于语言模型的角色生成已经非常成熟，同类技术可以应用在 NPC 对话中，玩家和 NPC 的交互有望展现出更高的自由度，有望开发出极致仿真的游戏；（2）视觉生成：GPT-4、腾讯混元文生图目前都支持视觉生成，策划使用视觉生成技术，先生成图再与美术沟通可以大大降低沟通成本，甚至“策划+文生图”模型的游戏开发将更加高速，成本更低；（3）元素生成：《微软模拟飞行 2020》应用 AI 技术生成世界各地约 15 亿座 3D 建筑物，突破了人工的限制，MoE 技术有望在元素生成中表

现出更好的性能，更快应用 AI 技术降低策划、美术和技术的高昂成本，打破游戏行业的不可能三角。

图19：角色生成：Replika AI



数据来源：replika 官网

图20：视觉生成：混元文生图

数据来源：腾讯混元

图21：元素生成：《微软模拟飞行 2020》

数据来源：游侠资讯

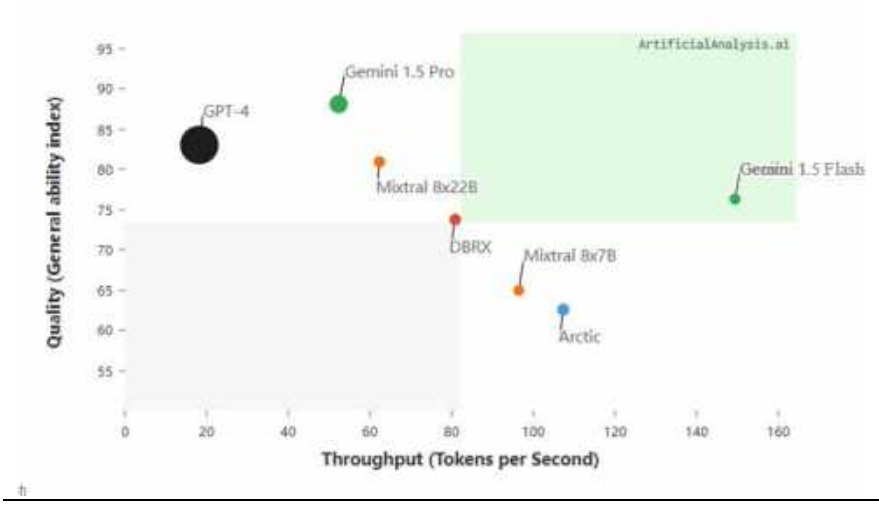
MoE 技术助力“AI+教育”行稳致远。大模型+教育领域产品已经有大量标的，例如网易有道、科大讯飞、作业帮、学而思等都在 AI 学习机器人领域积极布局。从教学端看，语言和视觉模型可以用于教案生成、素材查找、知识图谱化等，节省老师的劳动时间；从学生端看，私人的 AI 助教可以实时获取反馈，为学生提供个性化的学习方案，目前 QuillBot、Owlift、Grammarly、讯飞 AI 课程和学习机，文心大模型学习机 Z30 等大量标的已经正式提供辅助学习服务。MoE 技术在大模型端降本增效，终端辅助学习工具也将充分受益，成本更低的趋势下，AI 教育产品有望更快速渗透市场。

3.2. 国内外厂商积极应用 MoE 框架，助力降本增效

国外多个 MoE 模型已经开始商业化收费：Gemini 1.5 系列表现亮眼；多个 MoE 开源项目发布，更多 MoE 模型有望快速落地。国外以 GPT-4 为代表的 MoE 模型已经开始商业化，在综合表现、响应速度和定价的对比中，GPT-4 定价 30$/M Tokens,远远超过其它模型，Gemini 1.5 Pro 输出价格为 10.50$/M

Tokens。谷歌 Genimi 系列的综合表现位居前列，轻量版 1.5Flash 的推理速度在所有模型中最快，定价次于 Mixtral 8x7B。2024 年 4 月清华大学和微软联合发布了 MH-MoE 技术细节，开源项目已发布，Llama、Grok 等多个 MoE 开源项目也已经发布，更多项目的商业化落地有望加速。

图22：国外大模型综合表现-响应速度-定价比较



数据来源：Artificial Analysis（定价数据选取的时间节点为 2024 年 6 月 30 日之前）

国内 MoE 模型大量发布，综合表现亮眼。阿里巴巴和腾讯已经开始应用 MoE 框架，腾讯内部业务已接入 MoE 模型。2024 年 1 月 13 日 DeepSeek 发布国内开源 MoE 模型 DeepSeek MoE，5 月 6 日发布第二代模型 DeepSeek-V2，其它如天工 3.0、日日新 SenseNova 5.0、Kimi 也都选择应用 MoE 框架，推出新的或者升级后的模型。老牌厂商中，阿里巴巴和腾讯的大模型均采用了 MoE 框架。阿里巴巴 Qwen1.5-MoE-A2 参数量仅 14.3B，激活参数量仅 2.7B，对硬件资源的要求更小，推理速度更快。2024 年初腾讯应用 MoE 框架升级了混元大模型。

表 2 与 MoE 相关的国内大模型参数概览

模型	发布方	参数量（B）	上下文限制
Qwen1.5-MoE-A27B	阿里巴巴	14.3	3.2k tokens
hunyuan-standard	腾讯	1000	256k tokens
DeepSeek-V2	幻方量化	236	128k tokens
天工 2.0	昆仑万维	100	100k tokens
Step-2	阶跃星辰	1000	/
天工 3.0	昆仑万维	400	100k tokens
Kimi	月之暗面	200	200 万 汉字
日日新 SenseNova 5.0	商汤科技	600	200k token
APUS-sDAN4.0	APUS&国内新旦智能联合发布	136	32k tokens
XVERSE-MoE-A4.2B	元象科技	25.8	256k tokens
abab6	MiniMax	1000	200k tokens

数据来源：Qwen、腾讯、幻方量化等官网及公众号，国泰君安证券研究（相关参数统计时间节点为 2024 年 6 月 30 日前）

4. OpenAI o1 模型提供大模型训练及运用推理新范式

4.1. 大模型 OpenAI o1 推理侧创新运用“思维链”

北京时间 2024 年 9 月 13 日，OpenAI 发布 o1 大模型，分为两种版本 o1-preview 和 o1-mini。前者具有高级推理功能，在数学、编程、代码、优化、推理等复杂问题上具有博士水平的能力；后者成本更低，专注于 STEM 能力的训练，在编码推理领域更胜一筹。

新模型意味着大模型具备了更加复杂思考和推理的能力。对于复杂变成推理而言，新模型有变革式的训练方法和功能。因此，为了区别于传统的“GPT-4”系列，OpenAI 为其赋予崭新的系列名，将计数重置为 1。OpenAI CEO 表示“这是我们迄今为止功能最强大、最一致的模型系列 o1，也是迄今为止我们最好的推理模型。虽然 o1 仍然存在缺陷并有限，但使用时的感觉依然更加令人印象深刻”。

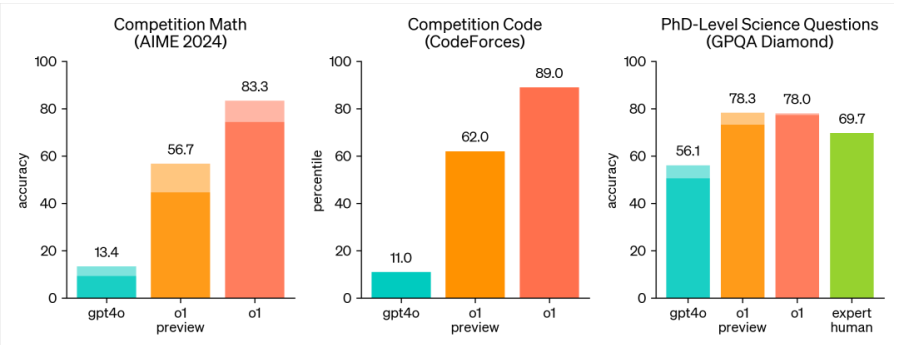
o1 模型学会人类“慢思考”，与之前的产品有着根本性的区别。先前的大模型多掌握“快思考”能力，即自动的、快速的、不用经过过多思考的推理能力。“慢思考”则是缓慢、有意识、需要努力的思考，涉及到逻辑推理、规划、专注和决策等，这种思考方式更为理性和准确，但因为它需要更多的认知资源，所以会比较慢。

o1 模型引入思维链技术，自主解决复杂问题。在响应用户提出的难题之前，o1 会产生一个缜密的内部思维链，进行长时间的思考，完善思考过程、意识逻辑错误、优化使用策略、推理正确答案。这种慢思考的能力对于处理复杂的推理任务尤为重要，因为它允许模型在给出回答之前考虑更多的信息和可能的解决方案。

强化学习标志 o1 学习方式的巨大转变。强化学习通过智能体在环境中通过执行动作来改变状态，并根据环境反馈的奖励来调整其行为。以解决难题为例，o1 面对复杂题目，分析答案并调整策略，通过一次又一次的试错，修改错误思维，找到正确答案，显著提升自身的学习能力和理解深度，这一机制标志着 AI 大模型学习推理方式的根本性转变。

o1 在科学、编程、代码等方面的能力超越以往大模型。OpenAI 声称 o1-preview 在物理、化学和生物学具有挑战性的基准任务上的表现类似于博士生。在国际数学奥林匹克资格考试中，o1 系列最新版本的准确率高达 83.3%，而 GPT-4o 仅为 13.4%。在知名的在线编程比赛 Codeforces 中，o1 拿到 89% 的百分位，GPT-4o 仅为 11%。

图23：OpenAI o1 具有超高的推理和逻辑能力



数据来源：OpenAI 官网

o1 能够形成数据飞轮效应。通过 Self-play RL，o1 可以修改思维链并完善使用策略。面对错误逻辑，o1 可以识别和纠正。面对正确的思考过程，又可以成为 o1 新的训练数据，从而不断改进推理能力，形成数据飞轮效应，

类似 AlphaGo 的价值网络随着 MCTS 生成更多精炼数据而改进。

4.2. 大模型结合强化学习开启应用推理新范式

o1 运用的基本原理为：自我对弈强化学习+思维链+推理标记+定制数据集。据天翼智库分析，一是采用大规模自我对弈强化学习，设置奖惩机制，让模型自行学习解决问题，通过不断尝试和纠错来掌握新技能。二是内置思维链（Chain of Thought, CoT），能够在解决问题前通过内置思维链进行推导，并将其推理过程外化，使得模型的决策过程更为透明，便于理解和验证。三是引入推理标记，用于辅助模型在对话环境中进行深层思考。四是使用专门的训练数据集，包含了大量复杂问题和对应的解题步骤，有助于模型掌握推理能力。

北大团队认为 OpenAI o1 运用的技术关键还是在于强化学习的搜索与学习机制。基于 LLM 已有的推理能力，迭代式的 Bootstrap 模型产生合理推理过程（Rationales）的能力，并将 Rationales 融入到训练过程内，让模型学会进行推理，而后再运用足够强大的计算量实现 Post-Training 阶段的 Scaling，类似于 STaR 的扩展版本。帮助 o1 取得如此性能飞跃的是 Post-Training 阶段 RL 计算量的 Scaling 和测试推理阶段思考时间的 Scaling。

OpenAI o1 的发布将重塑行业对于算力分配的认知，标志着 RL 下 Post-Training Scaling Law 的时代正式到来。OpenAI 研究员 Jason Wei 也表示，o1 模型背后的核心不只是通过 Prompt 提示词完成 CoT，而是引入 RL 训练模型，从而使模型更好地执行链式思考。隐式思维链思考给 o1 带来的巨大性能提升，也将启发行业在模型规模达到一定量级后，更多的将算力投入到 Post-Training 阶段的 RL 训练和推理阶段模型的思考过程当中。

4.3. 简单的功能与高昂的成本，o1 并非完美无缺

尽管在推理和思考方面，o1 遥遥领先，但这并非说明 o1 已是万能。o1-preview 与 o1-mini 目前仅有对话裸模型，不支持浏览网页与上传文字图片，功能有待进一步完善。同时，o1 在不需要复杂推理的领域表现没有明显优势。因此在更多的常见日常应用下，“GPT-4”系列仍具有使用优势，o1 系列仍属于早期发展阶段。

o1 当前仍具有较高使用壁垒。目前，o1 仅开放给拥有 ChatGPT Plus 和 Team 的用户免费试用，并且每周仅能给 o1-preview 发送 30 条消息，给 o1-mini 发送 50 条消息。通过 API 访问，只有消费 1000 美元以上并且付费时间超过一个月以上的用户才可以在 20RPM 的限速下使用。

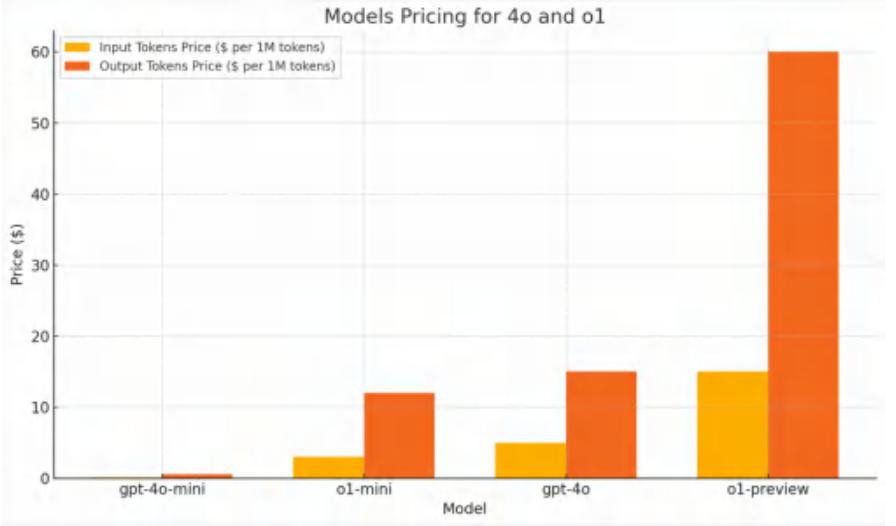
图24：OpenAI o1 使用壁垒高

数据来源：OpenAI 官网

o1 高昂的算力成本亟待解决。在 API 的价格上，o1-preview 输入每百万 token

要 15 美元，输出每百万 token 要 60 美元；o1-mini 输入每百万 token 要 3 美元，输出每百万 token 要 12 美元，输出成本都是推理成本的 4 倍，对比一下 GPT4o，分别是 5 美元和 15 美元，因此 o1 体现出高昂的推理成本。

图25：OpenAI o1 算力成本高昂



数据来源：OpenAI 官网

为了最有效使用 o1，OpenAI 给出了提问示范。（1）保持提示简单直接：模型擅长理解和响应简短、清晰的指令，而不需要大量的指导；（2）避免思路链提示：由于这些模型在内部进行推理，因此不需要提示它们“逐步思考”或“解释你的推理”；（3）使用分隔符来提高清晰度：使用三重引号、XML 标签或章节标题等分隔符来清楚地指示输入的不同部分，帮助模型适当地解释不同的部分；（4）限制检索增强生成 (RAG) 中的附加上下文。提供附加上下文或文档时，仅包含最相关信息，以防止模型过度复杂化其响应。

4.4. o1 加速 AGI 实现，孕育应用蓝海

o1 应用于专业化场景，替代脑力劳动成为可能。o1 大模型推理与思考能力，将赋能各行各业的发展，未来的大模型不仅可能取代传统的体力劳动，部分脑力劳动也可能被取而代之。例如在生物领域，大模型或被用来进行基因序列的测量、癌症药物的研发；在数理领域，大模型会被用来证明数学定理、进行天体探查。此外，还可以在教育、医疗、工程、金融、软件等领域增添更多可能性。

拉动算力需求，指导算力投资。此前市场普遍担心算力资本开支 26 年持续性问题，o1 的推出证明算力的重要性与需求紧迫性，未来算力需求可能超过训练预期。近期 OpenAI、xAI 和 Meta 均加大算力投入，抢先将万卡集群提升为十万卡集群，算力市场仍是广阔蓝海。

加速大模型向真正的 AGI 发展。目前，o1 普遍进行一分钟的“慢思考”后，会给出准确答案。通过强化学习与思维链的解决问题范式，未来的真正的 AGI 思考时间或更长，也许会给出更多颠覆性的回答。在 o1 的开创性示范下，AGI 相关研究进程会大大加快，离实现真正 AGI 更进一步。

颠覆人机交互模式，o1 成为决策智能伙伴。随着 o1 系列的不断迭代升级，必将在代码编程、计算研究、工程设计等领域有着举足轻重的地位并且彻底革新人与模型之间的关系。届时，AI 不再只会对简单命令做出机械反应的工具，而是能够通过程序运算、挖掘大数据网络、拥有自主“思考”能力的决策伙伴。

5. 风险提示

1）技术迭代不及预期

AI 大模型研发及应用尚处早期，前沿相关学术研究层出不穷，当前技术路径有可能在很短时间内被更优路径和方法论取代，而新的方法论从理论到落地尚需一定时间，因此相关公司在市场拓客过程中，可能存在技术迭代升级不及预期的风险。

2）AI 应用市场拓展节奏不及预期

考虑到 AI 应用跟市场需求匹配具有一定的不确定性，并且大模型即产品，AI 大模型从研发出来的那一刻便具备产品属性，因此在拓展市场过程中存在用户需求不能适当匹配的风险，导致获客节奏不及预期。

3）市场竞争加剧的风险

当前开源大模型的性能提升十分明显，参与厂商有望持续攀升，因此相关应用领域的竞争程度有可能更为激烈。因此存在市场竞争加剧的风险。

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

本报告仅供国泰君安证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

评级说明

投资建议的比较标准

投资评级分为股票评级和行业评级。

以报告发布后的 12 个月内的市场表现为比较标准，报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的沪深 300 指数涨跌幅为基准。

	评级	说明
股票投资评级	增持	相对沪深 300 指数涨幅 15%以上
	谨慎增持	相对沪深 300 指数涨幅介于 5%～15%之间
	中性	相对沪深 300 指数涨幅介于-5%～5%
	减持	相对沪深 300 指数下跌 5%以上
行业投资评级	增持	明显强于沪深 300 指数
	中性	基本与沪深 300 指数持平
	减持	明显弱于沪深 300 指数

国泰君安证券研究所

	上海	深圳	北京
地址	上海市静安区新闸路 669 号博华广场 20 层	深圳市福田区益田路 6003 号荣超商务中心 B 栋 27 层	北京市西城区金融大街甲 9 号 金融街中心南楼 18 层
邮编	200041	518026	100032
电话	（021）38676666	（0755）23976888	（010）83939888
E-mail:	gtjaresearch@gtjas.com		