

计算机行业深度：大模型研究框架（2025）

——“大模型”系列（5）

评级：推荐(维持)

刘熹(证券分析师)
S0350523040001
liux10@ghzq.com.cn

郭义俊(联系人)
S0350124070015
guoyj@ghzq.com.cn

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

最近一年走势



相关报告

《计算机“人工智能”系列专题：AutoGLM沉思：DeepResearch+Operator，开启智能体新阶段（推荐）*计算机*刘熹》——2025-04-17

《计算机行业深度报告：关税对自主可控的影响拆解——计算机“自主可控”系列报告（3）（推荐）*计算机*刘熹》——2025-04-11

《服务器电源：AI芯片功耗提升，高功率电源景气上行——AI算力“卖水人”系列（五）（推荐）*计算机*刘熹》——2025-03-10

相对沪深300表现

表现	1M	3M	12M
计算机	-17.7%	3.6%	26.0%
沪深300	-5.9%	-1.0%	5.7%

◆ 大模型发展回顾：以Transformer为基，Scaling law贯穿始终

2017年谷歌团队提出Transformer架构，创造性推动注意力层以及前馈神经网络层的发展，加速提升模型性能。2018–2020年是预训练Transformer模型时代，GPT-3以1750亿参数突破大规模预训练的可能性界限，而SFT及RLHF等技术帮助模型加速对齐人类价值观。此后随着训练侧Scaling Law描述的幂律关系出现收益递减，叠加高质量文本数据或逐步被AI耗尽，推理模型开始进入人们视野；以OpenAI发布o1-preview将AIME 2024的模型回答准确率从GPT4o的13.4%提升至56.7%，模型维持加速迭代更新。

◆ 国内大模型进展：行业充分竞争，降本提效为主旋律

资源有限的条件下，预计低成本高性能追平海外SOTA为2025年国产大模型的主题。我们以DeepSeek、豆包、阿里千问为例，1) DeepSeek-R1/V3依靠创新的降本提效手段，核心旨在资源有限的条件下，极大提升GPU在计算/通信上的利用率。2) 豆包大模型在2024年下半年发力，月活数据冲上全球第二和国内第一；同样在降本增效范式上依靠稀疏MoE架构实现小参数高性能；3) 阿里Qwen引领国产开源模型标杆的同时，依靠强化学习范式推出的QwQ-32B已登顶全球最强开源模型，以32B参数模型追平DeepSeek-R1满血模型性能，小参数高性能持续成为主旋律。

◆ 海外大模型进展：资源头部集中，押注AGI

算力充沛条件下，资源倾斜押注AGI。1) OpenAI：推理模型o1、多模态模型Sora均实现了行业引领，2025年来CEO Altman多次提及将发布OpenAI的首款Agent，且2025年也会是Agent爆发的元年；2) Google：前瞻布局原生多模态Gemini，2024年底发布多款Agent产品，同时布局轻量化模型Gemma抢占端侧生态；3) Meta：2024年12月Llama3.3以70B参数实现Llama3.1 405B的性能；基于Meta Live已实现实时语音交互、跨设备协作能力，发力通用智能体；4) 2024年10月Claude3.5 Sonnet升级新增computer use能力，让Claude像人一样使用电脑；此外，2025年抢先发布混合推理模型Claude-3.7-sonnet。

◆ 模型未来研判：投注后训练+算法大幅优化，低成本落地+实现AGI为终极目标

模型在架构以及pre-training——post training——落地层面均迎来加速变革。1) 模型架构层面，MoE与Transformer融合当前逐步成为主流架构，2024年全球MoE大模型数量呈爆发增长态势；2) pre-training层面，高质量数据或逐步耗尽的背景下，合成数据已然成为数字经济时代的“新型石油”，继续支撑模型的训练迭代；3) post-training方面，推理模型性能飞跃的关键也逐步转向该阶段RL计算量和测试推理阶段的思考时间，同时DeepSeek带动了纯强化学习的新范式；4) 模型落地层面，DeepSeek带动模型加速低成本部署趋势，通过MLA等低秩分解的方式实现显存占用的大幅降低，实现本地化部署DeepSeek-R1-32B及以下模型仅需要消费级显卡，大模型落地迎来真正意义上的元年。

◆ 大模型技术稳步提升，推动AGI时代加速到来，以大模型为底座的技术迭代或将持续驱动国产AI估值迎来重塑，维持计算机行业“推荐”评级。

◆ 相关公司

1) 算力：①云计算：中国电信、中国移动、中国联通、金山云、优刻得、青云科技、深信服；②IDC：云赛智联、光环新网、奥飞数据、数据港、润泽科技、科华数据、大位科技、ST鹏博、ST华通、ST证通；③芯片：海光信息、寒武纪；④服务器/一体机：中科曙光、浪潮信息、华勤技术、云从科技、恒为科技、中国软件国际、神州数码、烽火通信、朗科科技；⑤交换机：星网锐捷、紫光股份、中兴通讯；⑥液冷：飞荣达、英维克、申菱环境、高澜股份；⑦电源：欧陆通、麦格米特、中国长城；⑧柴油发电机：科泰电源、泰豪科技、潍柴重机、苏美达、动力新科、玉柴国际；⑨边缘计算：网宿科技、顺网科技、中科创达、云天励飞。

2) AI应用：①2G：中国软件、太极股份、深桑达、电科数字、广电运通、数字政通、中科星图、新点软件、国投智能、云从科技、税友股份、航天信息、拓尔思、能科科技、博思软件、华宇软件、通达海、金桥信息；②2B：金蝶国际、用友网络、泛微网络、致远互联、卫宁健康、创业慧康、广联达、石基信息、明源云、新致软件、汉得信息、鼎捷软件、赛意信息、莱斯信息、四川九州、东方财富、同花顺、恒生电子、新开普、佳发教育、拓维信息、远光软件、润和软件、索辰科技、中望软件、百融云、托普云农、焦点科技、盛视科技；③2C：金山办公、三六零、万兴科技、福昕软件、合合信息。

3) IT服务：宇信科技、京北方、中科软、软通动力、中科创达、新炬网络、天玑科技。

◆ 风险提示：大模型产业发展不及预期；中美博弈加剧；宏观经济影响下游需求；市场竞争加剧；相关标的公司业绩不及预期；国内外公司并不具备完全可比性；对标的相关资料和数据仅供参考；数据安全风险。

一、大模型发展回顾：以Transformer为基，Scaling law贯穿始终

- 1.1.1、大语言模型（LLMs）的兴起——自回归架构强化文本生成能力
- 1.1.2、Transformer架构克服RNN长文本局限性，标志着NLP的分水岭时刻
- 1.1.3、Transformer拆解：包括Encoder/Decoder、注意力层、前馈神经网络层
- 1.2.1、预训练Transformer模型时代 (2018–2020)：GPT VS BERT
- 1.2.2、GPT-3以1750亿参数开启了预训练侧Scaling law叙事
- 1.3.1、Transformer受限于长序列场景，计算复杂度与输入序列表现为指数增长关系
- 1.3.2、Mamba架构集成Transformer+RNN优势，成为Transformer架构的强劲挑战者

二、国内大模型进展：行业充分竞争，降本提效为主旋律

- 2.1、国内大模型：行业充分竞争，降本提效为主旋律
- 2.2、DeepSeek：早期确立AI战略，模型家族涵盖标准语言模型/推理模型/多模态模型
- 2.3、豆包大模型：实时语音、视频生成/理解领域布局，2024H2发力月活冲上全球第二
- 2.4、Qwen：AI为阿里巴巴未来战略核心，Qwen系列掀起国内模型开源革命

三、海外大模型进展：资源头部集中，压铸AGI

- 3.1、海外大模型：格局头部集中马太效应显著，集中押注面向AGI
- 3.2、OpenAI：全球AI大模型风向标，自然语言/多模态/推理模型上均作为引领角色
- 3.3、Google：Gemini面向智能体时代新作，原生多模态领域前瞻布局
- 3.4、Meta：10年布局跻身全球AI巨头，Llama成为全球开源模型标杆
- 3.5、Antropic：Claude-3.5对标OpenAI，Agent系列computer use推动人机交互变革

四、模型未来研判：投注后训练+算法的持续优化

- 4.1、模型架构的演进：从Dense到MoE，模型大幅降本提效
- 4.2、合成数据作为AI时代新石油，支撑模型继续在pre training上scaling
- 4.3、DeepSeek带动纯强化学习新范式，引领通向AGI之路
- 4.4、DeepSeek带动模型加速私有化+低成本部署趋势

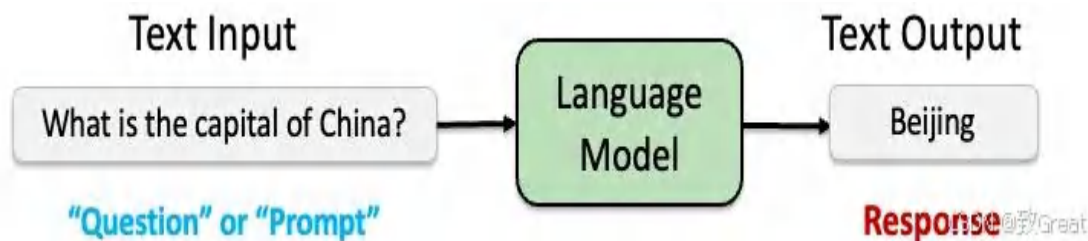
五、投资建议及风险提示

一、大模型发展回顾：以Transformer为基，Scaling law贯穿始终

● 语言模型是一种人工智能系统，旨在处理、理解和生成类似人类的语言。它们从大型数据集中学习模式和结构，使得能够产生连贯且上下文相关的文本，应用于翻译、摘要、聊天机器人和内容生成等领域。大语言模型（LLMs）是语言模型系统的子集。大语言模型规模显著更大，通常包含数十亿个参数（例如，GPT-3 拥有 1750 亿个参数），使得大语言模型在广泛的任务中表现出卓越的性能。大语言模型这一术语在 2018 至 2019 年间随着基于 Transformer 架构的模型出现开始受到关注，在 2020 年 GPT-3 发布后，LLMs 开始被广泛使用。

● 大多数 LLMs 以自回归方式操作，根据前面的文本预测下一个字（或 token / sub-word）的概率分布。这种自回归特性使模型能够学习复杂的语言模式和依赖关系，从而善于文本生成。在文本生成任时，LLM 通过解码算法确定下一个输出的字，这一过程可以采用的策略包括：1）选择概率最高的下个字；2）从预测的概率分布中随机采样一个字。

图：语言模型系统概念：旨在处理、理解和生成类似人类的语言



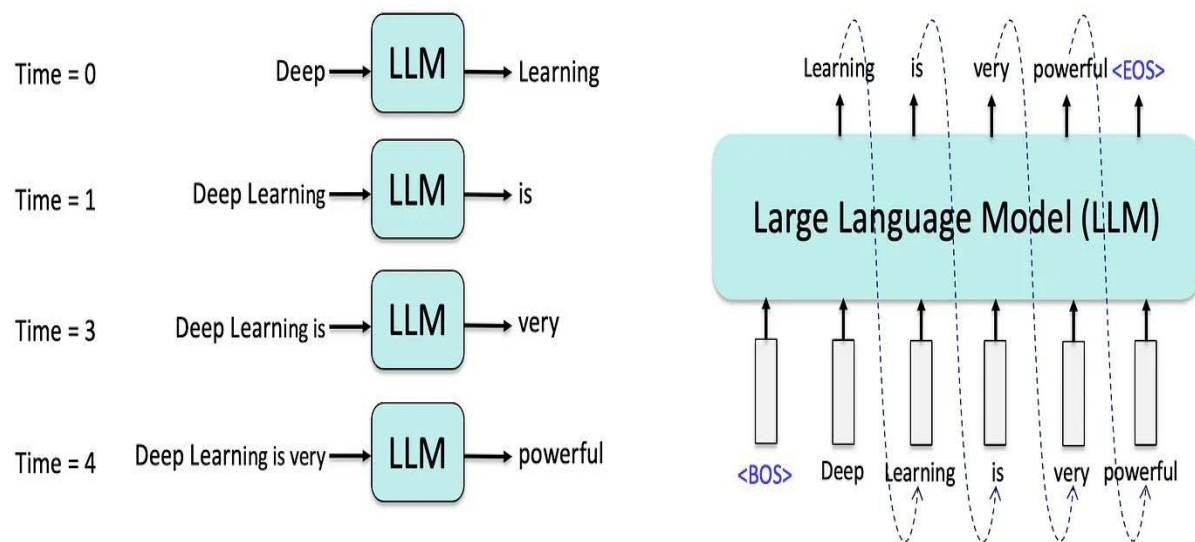
图：LLM通过解码算法来确定下一个输出的字



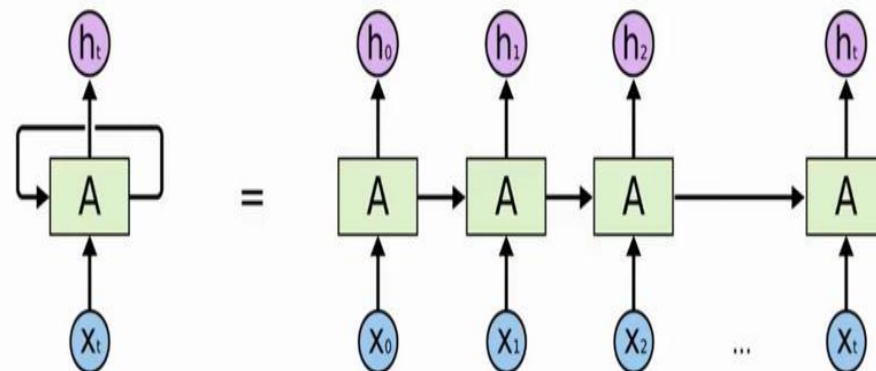
1.1.2、Transformer架构克服RNN长文本局限性，标志着NLP的分水岭时刻

2017年谷歌团队提出Transformer模型，Transformer架构也标志着NLP的分水岭时刻。Transformer突破了早期模型如循环神经网络（RNNs）与长短期记忆网络（LSTMs）在捕捉长依赖及顺序处理上的难点，同时RNN和LSTMs计算低效且易受梯度消失等问题困扰。Transformer的横空出世扫清了这些障碍，重塑了该领域格局，并为当代大型语言模型发展铺设了基石。

图：从提示词开始，模型迭代预测下一个词直到生成完整的序列或达到预定的条件

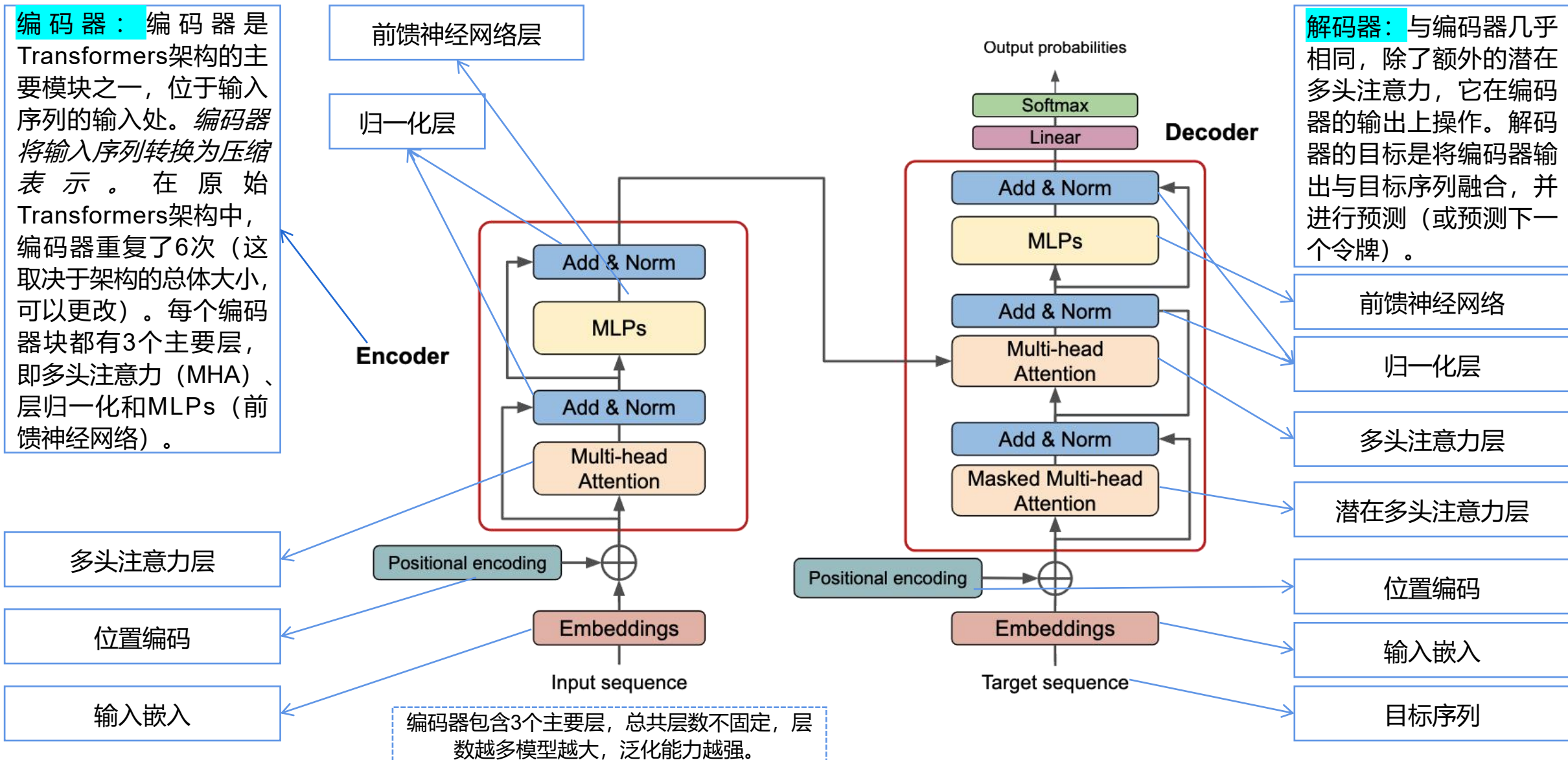


图：RNNs（循环神经网络）



1.1.3、Transformer拆解：包括Encoder/Decoder、注意力层、前馈神经网络层

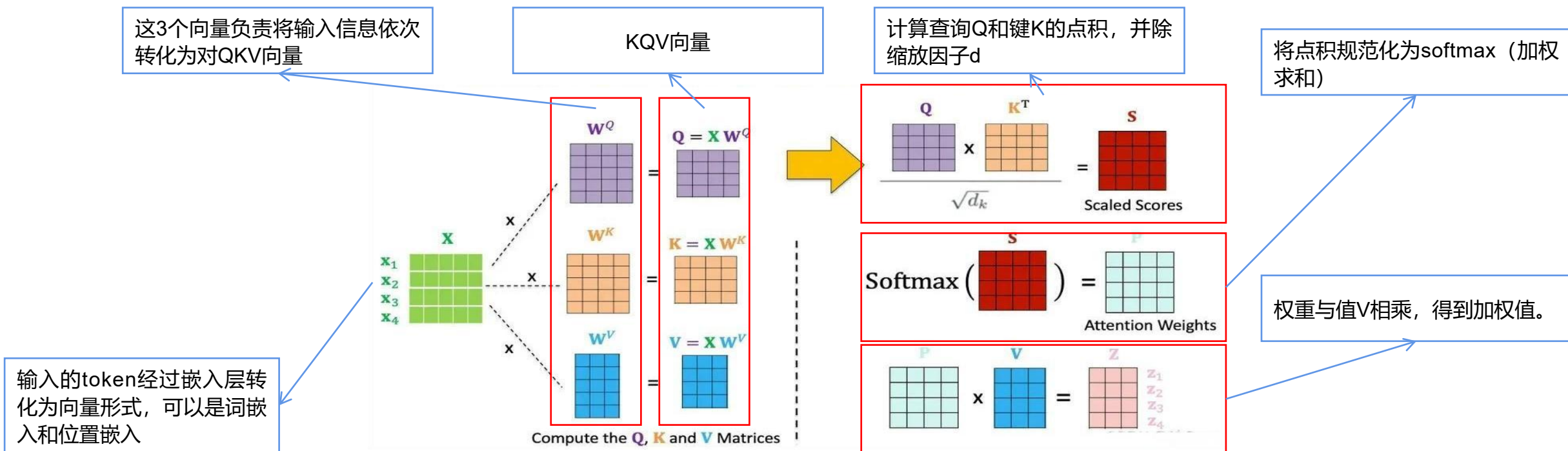
编码器：编码器是Transformers架构的主要模块之一，位于输入序列的输入处。编码器将输入序列转换为压缩表示。在原始Transformers架构中，编码器重复了6次（这取决于架构的总体大小，可以更改）。每个编码器块都有3个主要层，即多头注意力（MHA）、层归一化和MLPs（前馈神经网络）。



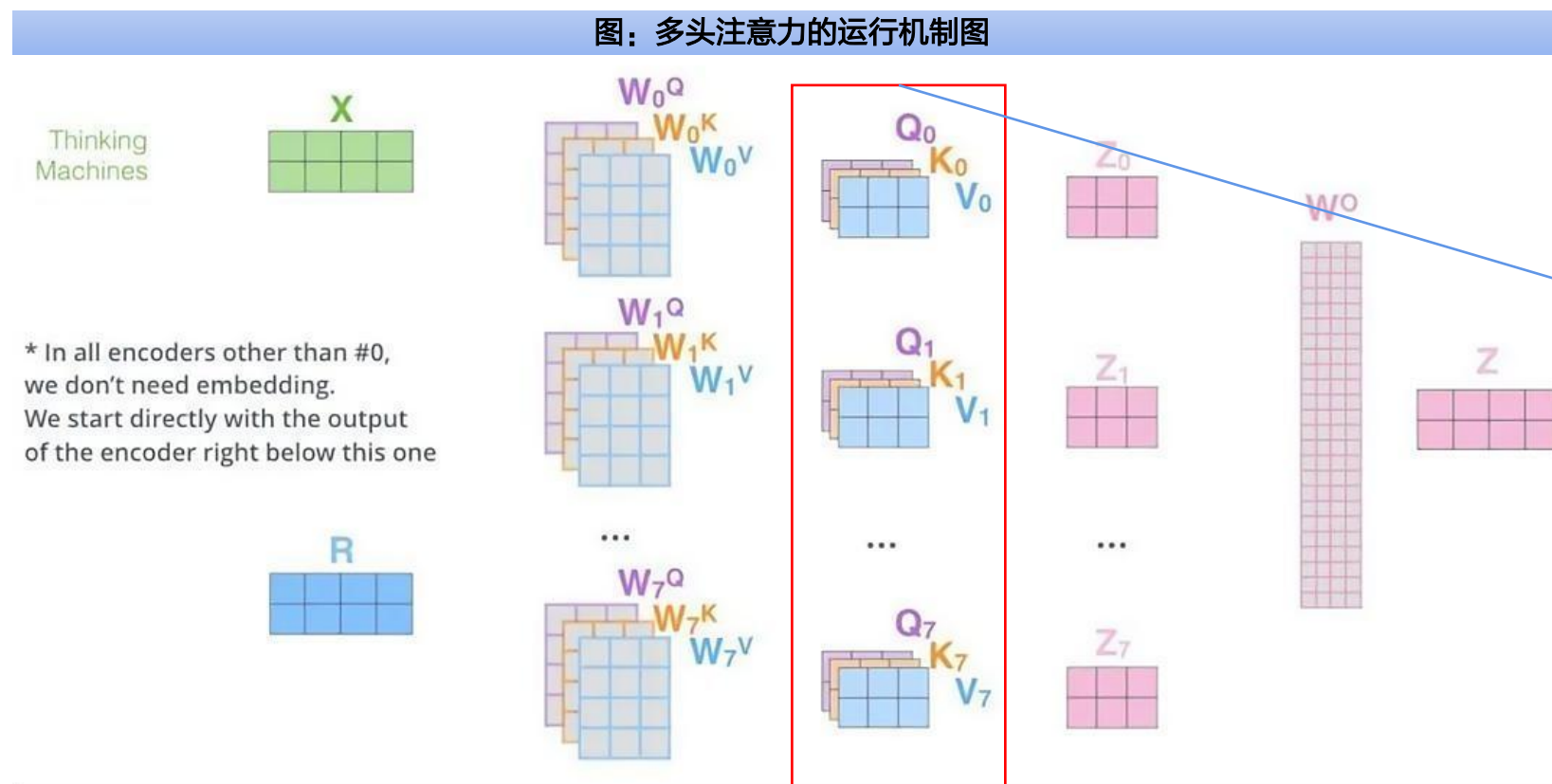
解码器：与编码器几乎相同，除了额外的潜在多头注意力，它在编码器的输出上操作。解码器的目标是将编码器输出与目标序列融合，并进行预测（或预测下一个令牌）。

1.1.4、Transformer核心点1——自注意力机制（Self-Attention）

- 注意力机制允许模型在解码时，根据当前生成的词动态地关注输入序列中的不同部分，有效捕捉与输出相关的输入信息，而非依赖于某个固定的上下文向量。注意力机制使得模型更容易捕捉长距离依赖关系，模型在每个时间步都可以选择关注距离较远的输入部分。数学表达上，注意力度量两个向量之间的相似性并返回加权相似性分数，标准的注意力函数接受三个主要输入，即查询、键和值向量。
- 例如，在电商平台搜索特定商品时，输入内容即为Query，搜索引擎据此匹配Key（涵盖商品种类、颜色、描述等），并通过计算Query与Key的相似度（即权值），得出匹配内容（Value）。



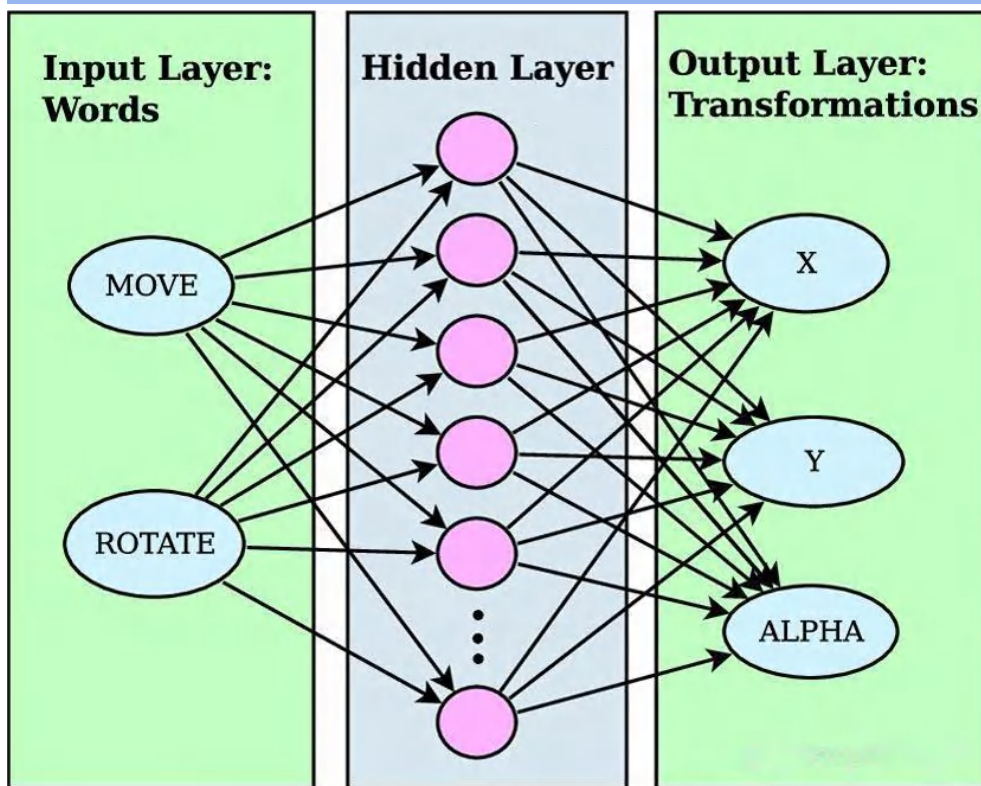
✔ **Multi-headed attention**（多头注意力机制）增强了自注意能力，扩展关注位置，同时为注意力层提供多个“表示子空间”。假设模型若用了8个注意力头，就会有8组不同的Q/K/V矩阵，每个输入的词向量都被投影到8个表示子空间中进行计算。



与自注意力机制的区别：将线性变换后的查询、键和值矩阵分割成多个头。每个头都有自己的查询、键和值矩阵。然后，在每个头中独立地计算注意力分数。

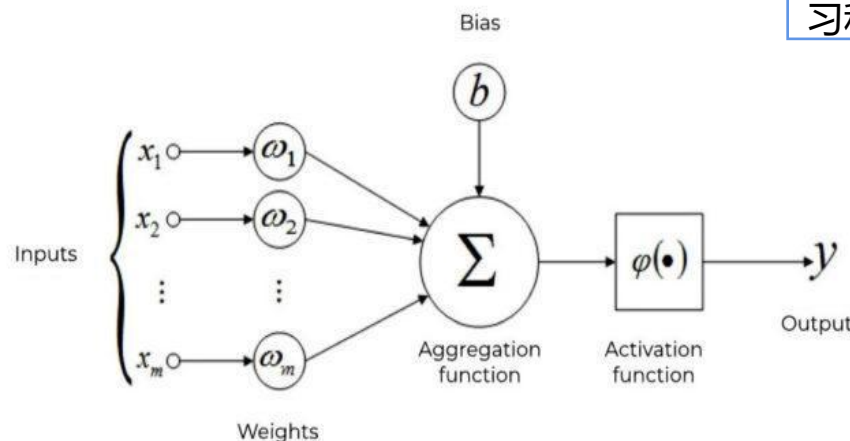
✓ 前馈神经网络是最基本的人工神经网络结构，由多层节点组成，每个节点都与下一层的所有节点相连。前馈神经网络的特点是信息只能单向流动，即从输入层到隐藏层再到输出层，不能反向流动。工作原理上表现为，输入数据首先进入输入层，然后通过权重和偏置传递到隐藏层，隐藏层中的节点对输入进行加权求和，并通过激活函数进行非线性转换，最后输出层接收到经过隐藏层处理的信号，并产生最终的输出。

图：前馈神经网络架构



偏置：加在输入上的常数，用于调整激活函数的输出。

激活函数：用于在网络中引入非线性，使得网络能够学习和模拟复杂的函数映射。



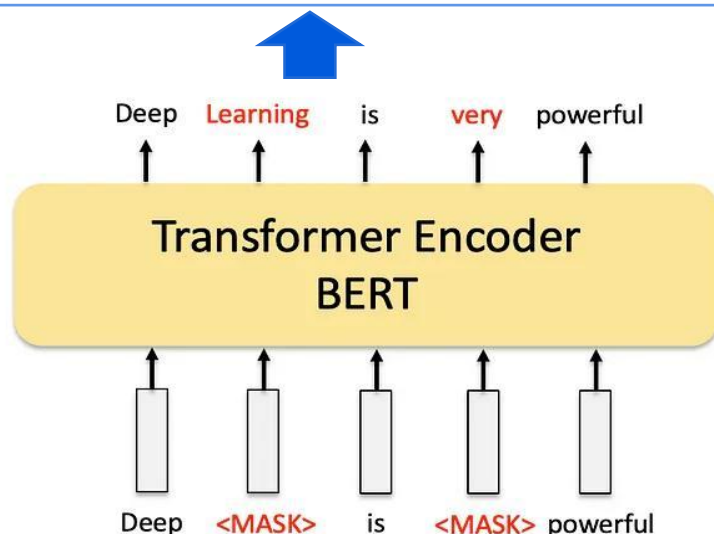
Output of a neuron

$$y = \varphi(b + \sum x_i \omega_i)$$

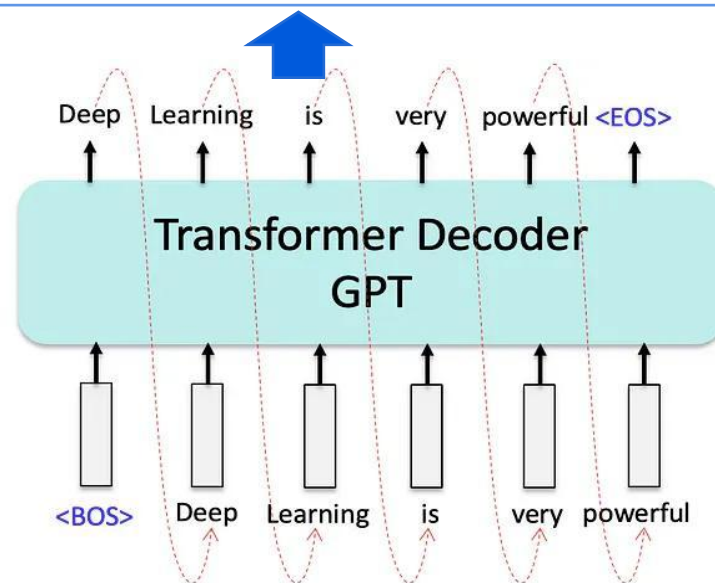
权重：连接输入层和隐藏层、隐藏层和输出层的连接强度

- ✔ **Transformer架构的出现也标志着预训练模型的崛起及对扩展性的重视。** BERT与GPT的诞生便显示了大规模预训练与微调范式的成效。
- ✔ 2018年，谷歌推出了BERT模型，模型采用Transformer编码器，在多个NLP任务中取得了突破性进展。与以往仅单向处理文本的模型不同，BERT运用了双向训练方法同时双向捕获上下文信息，以至于BERT在文本分类、命名实体识别及情感分析等语言理解任务中展现出了不俗的表现。
- ✔ **OpenAI的GPT系列采用了与BERT差异化的方法，借助自回归预训练来强化生成能力。** 2018年OpenAI发布GPT的第一个版本，凭借Transformer解码器，GPT模型在自回归语言建模及文本生成领域展现了出色的性能。

MLM (掩码语言建模): BERT不是预测序列中的下一个词，而是被训练预测句子中随机掩码的标记。这迫使模型在进行预测时考虑整个句子的上下文；
NSP (下一句预测): 模型学习预测两个句子是否在文档中连续，从而理解句子之间关系。

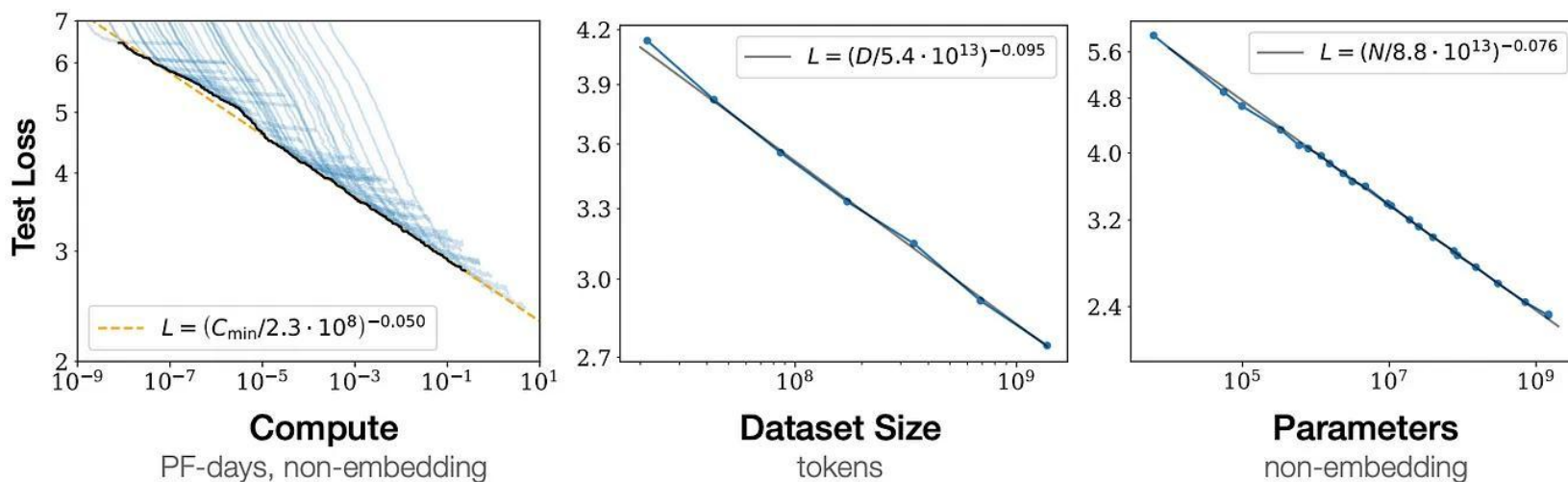


单向自回归训练: GPT使用因果语言建模目标进行训练，其中模型仅基于前面的标记预测下一个标记。因此适合于生成任务。
下游任务的微调: GPT的一个关键贡献是它能够在不需要特定任务架构的情况下针对特定下游任务进行微调。只需添加一个分类头或修改输入格式，GPT就可以适应诸如情感分析、机器翻译和问答等任务。



- 2020年OpenAI发布GPT-3（1750亿参数模型），NLP模型迎来了转折点。1750亿参数突破了大规模预训练的界限，展示了显著的少样本和零样本学习能力，在推理时只需提供最少或无需示例即可执行任务。GPT-3的生成能力扩展到了创意写作、编程和复杂推理任务，展示了超大模型的潜力。
- 预训练侧Scaling law的叙事开始出现，模型性能随着模型大小、数据集大小以及训练使用的计算量的增加而平稳提升。在2018年至2020年间，研究人员发现随着模型规模的增长，模型在捕捉复杂模式和泛化到新任务方面变得更好。这种规模效应得到了三个关键因素的支持：1）数据集大小：更大的模型需要庞大的数据集进行预训练；2）计算资源：强大硬件（如GPU和TPU）的可用性以及分布式训练技术，使得高效训练具有数十亿参数的模型成为可能；3）高效架构：混合精度训练和梯度检查点等创新降低了计算成本，使得在合理的时间和预算内进行大规模训练更加实际。

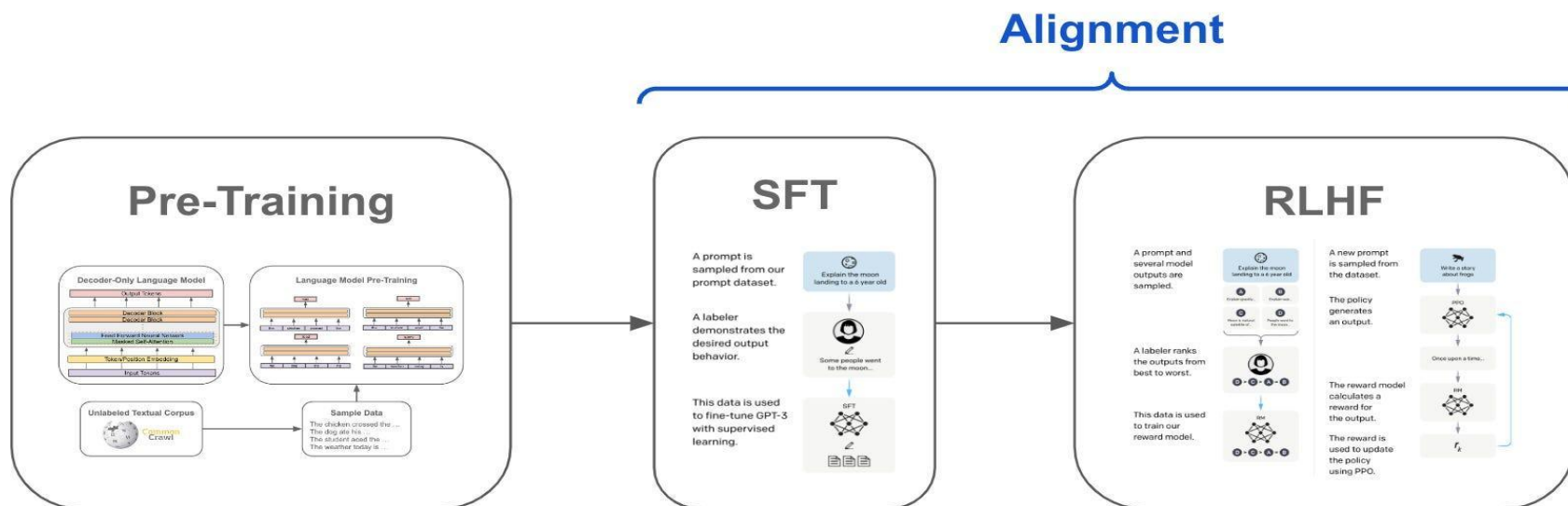
图：Scaling law的三个方向



1.2.3、Post-training重要性凸显，RLHF范式出现 (2021–2022)

- ✓ GPT-3同时也表现出大型语言模型与人类价值观、偏好及期望保持一致上的挑战。其中，“幻觉”问题尤为突出，即LLM生成的内容可能与事实不符、缺乏意义或与输入提示相悖，给人以“言之凿凿却离题万里”之感。为应对模型幻觉，2021至2022年间，研究人员推动了监督微调（SFT）及基于人类反馈的强化学习（RLHF）等技术的进展。
- ✓ SFT（有监督学习方法）通过提供明确的输入——输出对，模型学习其中的映射关系。但SFT的弊端包括有1）可扩展性问题：收集人类演示需劳动密集且耗时，尤其是对于复杂或小众任务；2）性能：简单模仿人类行为并不能保证模型会超越人类表现，或在未见过的任务上很好地泛化。
- ✓ RLHF（基于人类反馈的强化学习）解决了SFT中可扩展性和性能限制的问题。RLHF包括两个阶段，首先：1）根据人类偏好数据集训练一个奖励模型，该模型学习根据人类反馈评估输出的质量。2）使用强化学习微调LLM，奖励模型使用近端策略优化（PPO）指导LLM的微调，模型学会了生成更符合人类偏好和期望的输出。
- ✓ 2022年3月，OpenAI发布GPT-3.5，与GPT3架构相同但关键增强包括改进数据更好地遵循指令，减少了幻觉。

图：监督微调SFT以及基于人类反馈的强化学习RLHF

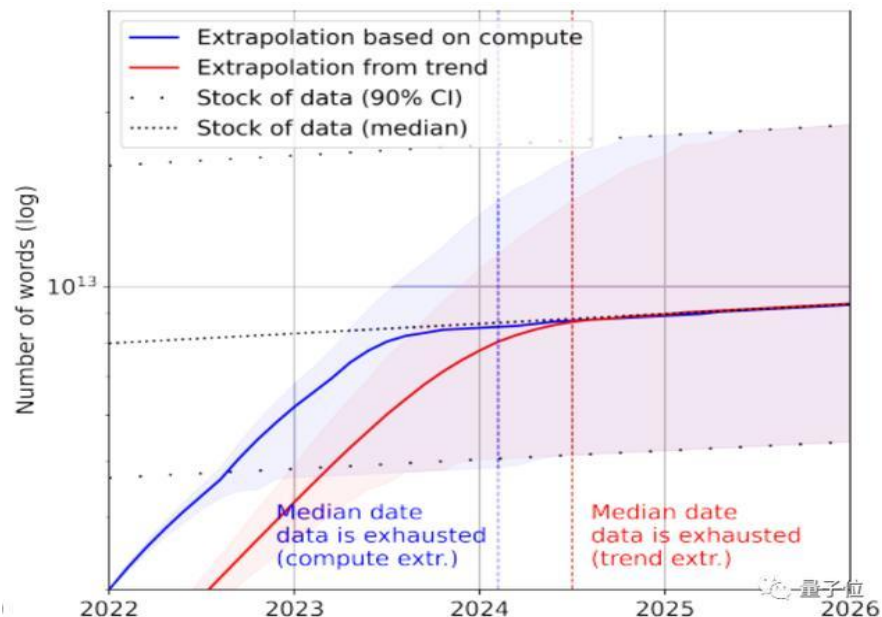


1.2.4、训练侧Scaling law瓶颈出现，推理侧接过Scaling law叙事大旗

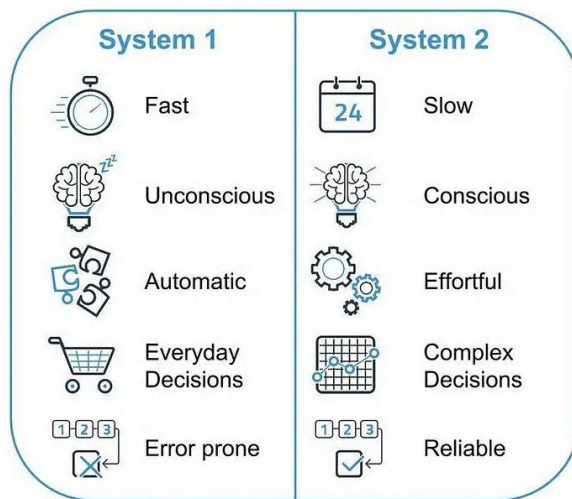
● **Scaling Law** 描述了模型性能随着模型参数、数据量和计算资源增加而提升的幂律关系，但这种提升并非线性，而是呈现出收益递减现象。在模型规模较大时，资源的增加对性能提升的影响变得有限，资源投入与性能提升之间的平衡关系并非单纯“大力出奇迹”。根据Epoch团队的论文《Will we run out of data? Limits of LLM scaling based on human-generated data》中统计，高质量语言数据存量只剩下约 $4.6 \times 10^{12} \sim 1.7 \times 10^{13}$ 个单词。结合增长率，论文预测高质量文本数据会在2023~2027年间被AI耗尽。

● **2024年**，AI开发开始强调增强推理(Reasoning)，从简单的模式识别转向更逻辑化和结构化的思维过程。2024年9月，OpenAI发布o1-preview，据Altman在x上分享的数据，AIME 2024（高水平的数学竞赛）中，o1-preview将模型回答准确率从GPT4o的13.4%提升至56.7%，o1正式版是83.3%。

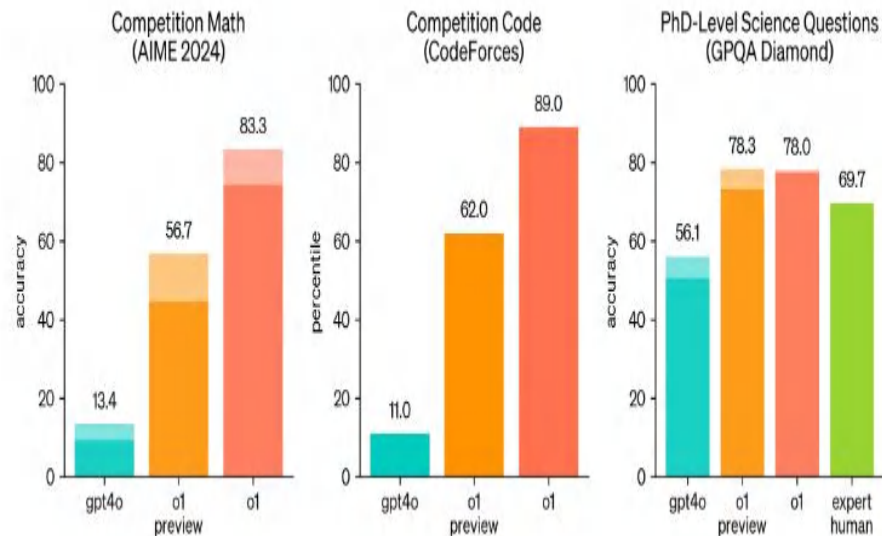
图：人类高质量文本数据或在2026年被AI耗尽



图：模型从范式1到范式2演进



图：OpenAI发布的o1实现能力的跨越式提升



- 长序列场景下Transformer计算复杂度显著提升：自注意力机制的计算复杂度为 $O(N^2, d)$ ，其中 N 代表序列长度， d 表示 token 嵌入的维度，这意味着 Transformer 模型的计算复杂度会随着输入序列长度（token 数量）的增加呈二次方增长，这种高计算复杂度会导致计算资源的大量消耗，对硬件性能提出了极高的要求。
- 随着基于 Transformer 架构的模型规模不断扩大，训练和部署成本也随之大幅增加。在计算资源方面，Transformer模型不仅需要大量的计算资源来支撑复杂的运算，还对并行处理能力有着较高的要求。训练成本不仅要涵盖高性能的 GPU，还需要大量的存储空间。并且，随着序列长度的增加，其平方级的扩展会导致内存使用量急剧上升。这使得训练和部署 Transformer 模型的成本居高不下，在一些资源受限的场景中，其应用受到了明显的限制。

图：Transformer架构下序列长度和算力需求

TABLE 2
Efficiency comparison of some novel non-Transformer models. Note that we denote n as the input length and d as the input dimension.

Model	Training Form	Training Computational Complexity	Training Memory Complexity	Inference Form	Inference Computational Complexity	
					Prefilling	Decoding (per token)
Transformer [25]	Transformer-like	$O(n^2d)$	$O(n^2 + nd)$	Transformer-like	$O(n^2d)$	$O(nd)$
S4 [64]	Convolution	$O(nd^2 \log n)$	$O(nd)$	Recurrence	$O(nd^2)$	$O(d^2)$
Mamba [73]	Recurrence	$O(nd^2 \log n)$	$O(nd)$	Recurrence	$O(nd^2)$	$O(d^2)$
Hyena [59]	Convolution	$O(nd \log n)$	$O(nd)$	Convolution	$O(nd \log n)$	$O(nd \log n)$
RetNet [61]	Transformer-like	$O(n^2d)$	$O(n^2 + nd)$	Recurrence	$O(n^2d)$	$O(d^2)$
RWKV [60]	Recurrence	$O(nd^2)$	$O(nd)$	Recurrence	$O(nd^2)$	$O(d^2)$

图：国内外大模型上下文已经有大幅提升

国内外关注度较高的模型上下文可接受长度表

公司/机构/团队	模型/产品名称	上下文Tokens
OpenAI	GPT-3.5	4K-16K
	GPT-4	8K-32K
Anthropic	Claude	100K
	Claude2	100K
Meta	LLaMA	2k
	LLaMA2	4k
	Llama 2 Long	32K
IDEAS NCBR 、Google DeepMind等	Long LLaMA	256k
Moonshot（月之暗面）	Kimi Chat	400K
港中文贾佳亚团队、MIT	LongAlpaca	32K-100K

- Mamba融合Transformer、RNN架构的特点，实现在推理和训练上的加速。结构化的状态空间序列模型（SSM）能高效捕获序列数据中的复杂依赖关系，其一大关键是融合了卷积神经网络以及循环神经网络的特点，让计算开销随序列长度而线性或近线性变化，大幅降低计算成本，而Mamba则为SSM的一种变体。
- Mamba 可根据输入对 SSM 进行重新参数化，让模型在滤除不相关信息的同时无限期地保留必要和相关的信息。在CVPR 2025上，英伟达推出混合新架构MambaVision视觉骨干网络，打破精度/吞吐瓶颈。MambaVision是首个针对计算机视觉应用，结合Mamba和Transformer的混合架构的尝试。主要贡献包括1）引入了重新设计的适用于视觉任务的Mamba模块，提升了相较于原始Mamba架构的准确性和图像处理能力；2）系统性地研究了Mamba和Transformer模块的融合模式，并展示了在最终阶段加入自注意力模块，显著提高了模型捕捉全局上下文和长距离空间依赖的能力。

图：Mamba、Transformer、RNNs架构对比



图：Mamba逐步成长为新模型基础

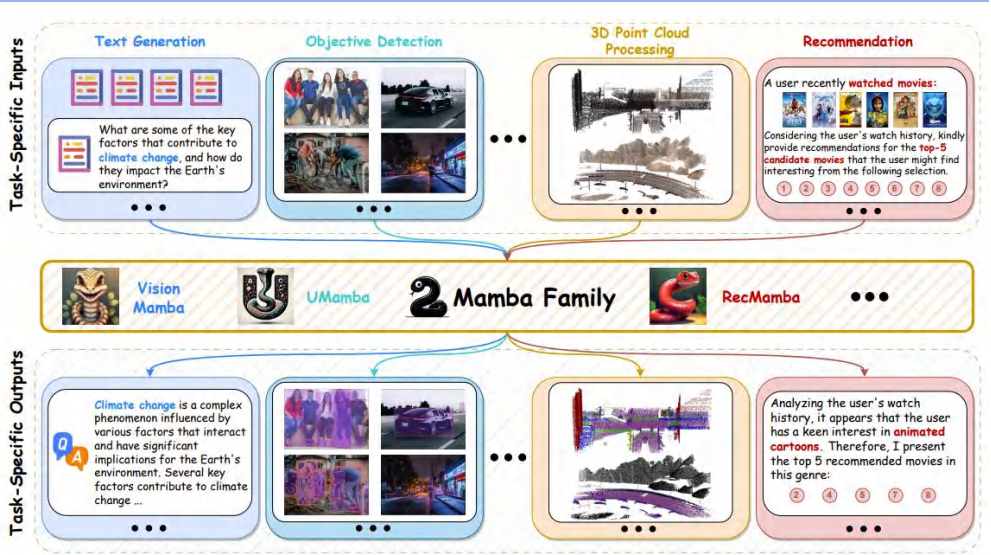


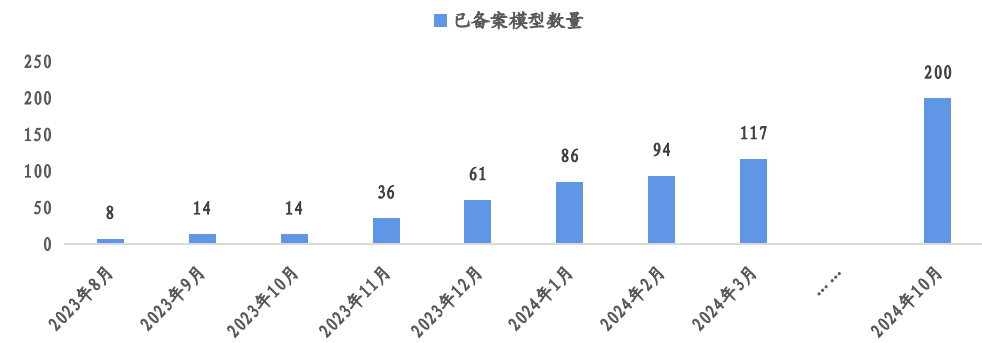
Fig. 1. Examples of the applications of Mamba-based models for different downstream tasks.

二、国内大模型进展：行业充分竞争，降本提效为主旋律

2.1、国内大模型：行业充分竞争，降本提效为主旋律

- 国产大模型生产蓬勃发展。据工信部数据，截至2024年10月，现有完成备案并上线为公众提供服务的生成式人工智能服务大模型近200个，注册用户超过了6亿，相较2024年初实现了翻倍以上的增长。
- 国产模型中，典型代表包括不限于：具备先发优势的百度文心一言、清华大学学术血脉的智谱清言、B端市场发力的讯飞星火、文字生成领域具备领先优势的Kimi、媲美Sora视频生成能力的可灵、聚焦B端发力的华为盘古、霸榜开源社区下载量的Qwen、依托腾讯生态优势的元宝、依托字节巨大流量入口的豆包以及凭借算法优化媲美GPT-o1的DeepSeek。

图：我国大模型备案数量



模型名称	核心能力	技术特点	典型应用场景
DeepSeek-R1	思维链深度推理、数学能力出色、深度学习算法	强化学习训练、GRPO算法、冷启动技术、模型蒸馏	科研分析、编程辅助
阿里Qwen2.5	多语言处理、128Ktokens上下文、代码生成修复、数学推理	集成阿里云生态、大规模预训练、逐层推理	文案策划、编程开发、数据分析
豆包1.5Pro	多模态交互、知识理解、代码生成、逻辑推理	大规模稀疏MoE架构、自主数据生产体系、7倍MoE性能杠杆	视频创作、电商导购、语音助手、文案撰写
腾讯混元	多轮交互、长上下文理解、知识增强、多模态能力	文本、图像、视频处理多种数据处理、灵活微调能力	文档生成、智能客服、媒体创作、游戏角色生成
百度文心4.0	搜索引擎增强、多模态生成	知识增强的ERNIE架构、多模态融合、中文处理优势	文本创作、智能客服、知识检索和分析
讯飞星火	多模态交互、代码生成、复杂问题逻辑分析、深度理解自然语言语义	语音识别准确率行业领先、多语言支持	智能教育、会议转录、智能客服
月之暗面Kimi	200万字长文本处理、多格式文本解析、信息检索	超长上下文理解、支持文本、语音、视觉信息处理、个性化调优	学术研究、内容创作
智谱GLM-4	多任务语言理解、128K上下文处理、文生图、多模态理解能力	Transformer架构、推理速度迅捷、智能体能力	PPT生成、文本创作、智能客服、教学辅导
天工大模型4.0	中文逻辑推理、多模态交互、情感理解与个性化服务能力	三阶段自研训练方案、自蒸馏和拒绝采样技术、端到端语音对话	智能学习助手、医疗辅助分析、智能客服、文本创作
Baichuan-M1-preview	语言、视觉和搜索推理能力、精准视觉分析	Transformer架构、万亿级token医疗数据、ELO算法	医疗辅助决策、文献信息提取、代码优化
MiniMax-01	400万token上下文处理、长文本理解与检索	线性注意力机制、优化的混合专家架构、硬件友好与计算优化	智能助手、文案撰写、创意设计、文献数据解读
零一万物Yi-Lightning	高效推理、数学与代码能力、多语言处理	MoE混合专家模型架构、混合注意力机制、动态top-p路由	智能客服、内容创作、零售电商、文本数据分析、算法验证
阶跃星辰Step-2	自然语言理解与生成、多领域推理、模糊指令处理	自主研发MFA架构、百万亿参数MoE架构、多阶段训练模式	内容创作、智能客服、教学资源生成、科研文献分析

2.2、DeepSeek：早期确立AI战略，模型家族涵盖标准语言模型/推理模型/多模态模型



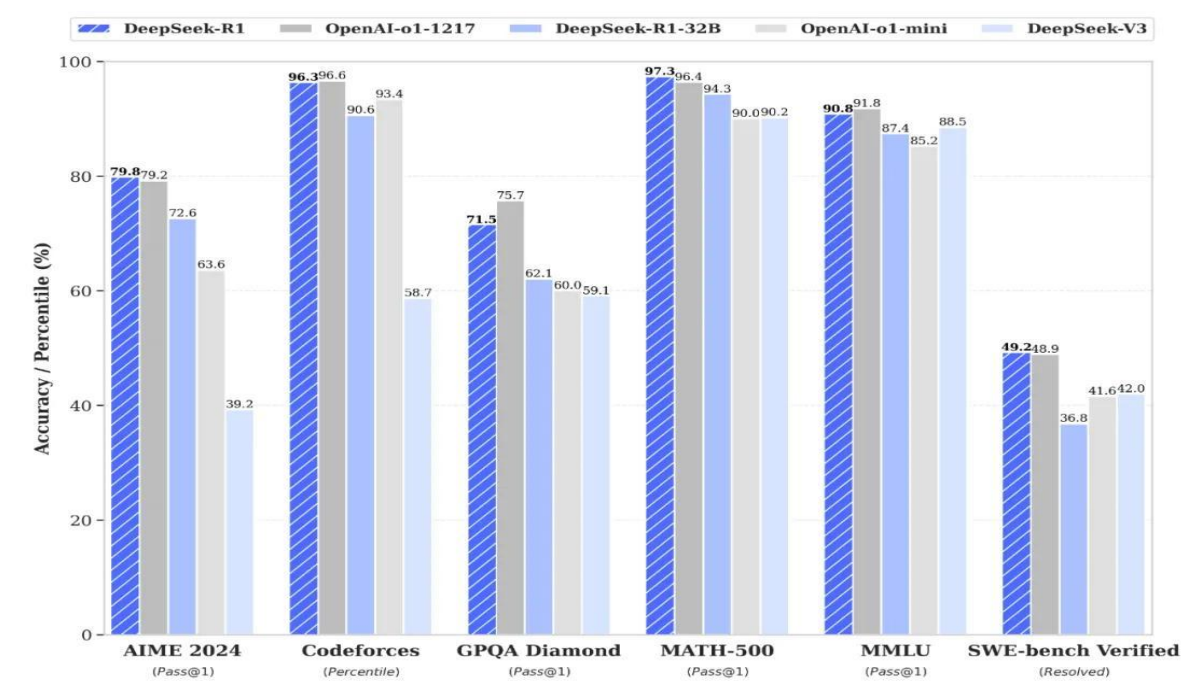
- DeepSeek是一家于2023年成立的中国初创企业，创始人是AI驱动量化对冲基金幻方量化的掌门人梁文锋。2021年，幻方量化的资产管理规模突破千亿大关，跻身国内量化私募领域的“四大天王”之列。2023年梁文锋宣布正式进军通用人工智能领域，创办DeepSeek，专注于做真正人类级别的人工智能。
- DeepSeek团队以年轻化为主，具备深厚技术底蕴。创始人梁文锋曾在36氪的采访中，给出了DeepSeek的员工画像：“都是一些Top高校的应届毕业生、没毕业的博四、博五实习生，还有一些毕业才几年的年轻人。”自2023年5月诞生以来，DeepSeek始终维持约150人的精英团队，推行无职级界限、高度扁平化的文化，以此激发研究灵感，高效调配资源。2023年5月DeepSeek正式成立时，团队已汇聚近百名卓越工程师。如今，即便不计杭州的基础设施团队，北京团队亦拥有百名工程师。

模型类别	日期	名称	内容	对标
LLM	2023年11月2日	DeepSeek Coder	模型包括 1B, 7B, 33B 多种尺寸，开源内容包含 Base 模型和指令调优模型。	Meta的CodeLlama是业内标杆，但DeepSeek Coder展示出多方位领先的架势。
	2024年6月17日	DeepSeek Coder V2	代码大模型，提供了 236B 和 16B 两种版本。DeepSeek Coder V2 的 API 服务也同步上线，价格依旧是「1元/百万输入，2元/百万输出」。	能力超越了当时最先进的闭源模型 GPT-4-Turbo。
	2023年11月29日	DeepSeek LLM 67B	首款通用大语言模型，且同步开源了 7B 和 67B 两种不同规模的模型，甚至将模型训练过程中产生的 9 个 checkpoints 也一并公开，	Meta的同级别模型 LLaMA2 70B，并在近20个中英文的公开评测榜单上表现更佳。
	2024年3月11日	DeepSeek-VL	多模态 AI 技术上的初步尝试，尺寸为 7B 与1.3B，模型和技术论文同步开源。	
	2024年5月	DeepSeek-V2	通用 MoE 大模型的开源发布，DeepSeek-V2 使用了 MLA（多头潜在注意力机制），将模型的显存占用率降低至传统 MHA 的 5%-13%	对标 GPT-4-Turbo，而 API 价格只有后者的 1/70
	2024年9月6日	DeepSeek-V2.5 融合模型	Chat模型聚焦通用对话能力，Code模型聚焦代码处理能力合二为一，更好的对齐了人类偏好，	
	2024年12月10日	DeepSeek-V2.5-1210	DeepSeek V2 系列收官之作，全面提升了包括数学、代码、写作、角色扮演等在内的多方能力。	
	2024年12月26日	DeepSeek-V3	开源发布，训练成本估算只有 550 万美金	性能上全面对标海外领军闭源模型，生成速度也大幅提升。
推理模型	2024年2月5日	DeepSeekMat	数学推理模型，仅有 7B 参数	数学推理能力上直逼 GPT-4
	2024年8月16日	DeepSeek-Prover-V1.5	数学定理证明模型	在高中和大学数学定理证明测试中，均超越多款知名的开源模型。
	2024年11月20日	DeepSeek-R1-Lite	推理模型，为之后 V3 的后训练，提供了足量的合成数据。	媲美 o1-preview
	2025年1月20日	DeepSeek-R1	发布并开源，开放了思维链输出功能，将模型开源 License 统一变更为 MIT 许可证，并明确用户协议允许“模型蒸馏”。	在性能上全面对齐 OpenAI o1 正式版
多模态模型	2023年12月18日	DreamCraft3D	文生 3D 模型，可从一句话生成高质量的三维模型，实现了 AIGC 从 2D 平面到 3D 立体空间的跨越。	
	2024年12月13日	DeepSeek-VL2	多模态大模型，采用了 MoE 架构，视觉能力得到了显著提升，有 3B、16B 和 27B 三种尺寸，在各项指标上极具优势。	
	2025年1月27日	DeepSeek Janus-Pro	开源发布的多模态模型。	

2.2、DeepSeek：DeepSeek-R1性能对标OpenAI o1正式版，实现发布即上线

- DeepSeek-R1性能比肩OpenAI-o1。DeepSeek-R1 在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版。
- 开放的许可证和用户协议。DeepSeek在发布并开源 R1 的同时，同步在协议授权层面也进行了如下调整：1）模型开源 License 统一使用 MIT，开源仓库（包括模型权重）统一采用标准化、宽松的 MIT License，完全开源，不限制商用，无需申请。2）产品协议明确可“模型蒸馏”；为了进一步促进技术的开源和共享，支持用户进行“模型蒸馏”，明确允许用户利用模型输出、通过模型蒸馏等方式训练其他模型。

图：DeepSeek-R1性能比肩 OpenAI o1 正式版



图：DeepSeek-R1发布即上线



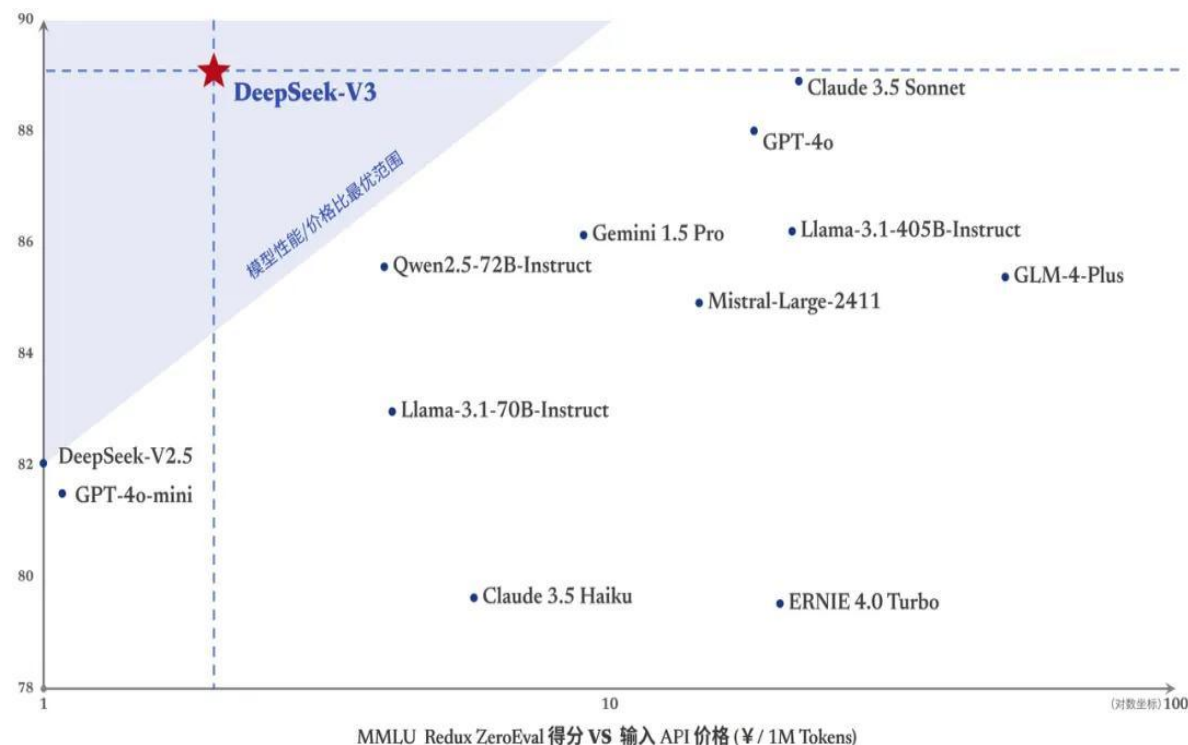
2.2、DeepSeek：DeepSeek-V3/R1均具备领先的性价比优势

✓ DeepSeek 系列模型均具备显著定价优势。

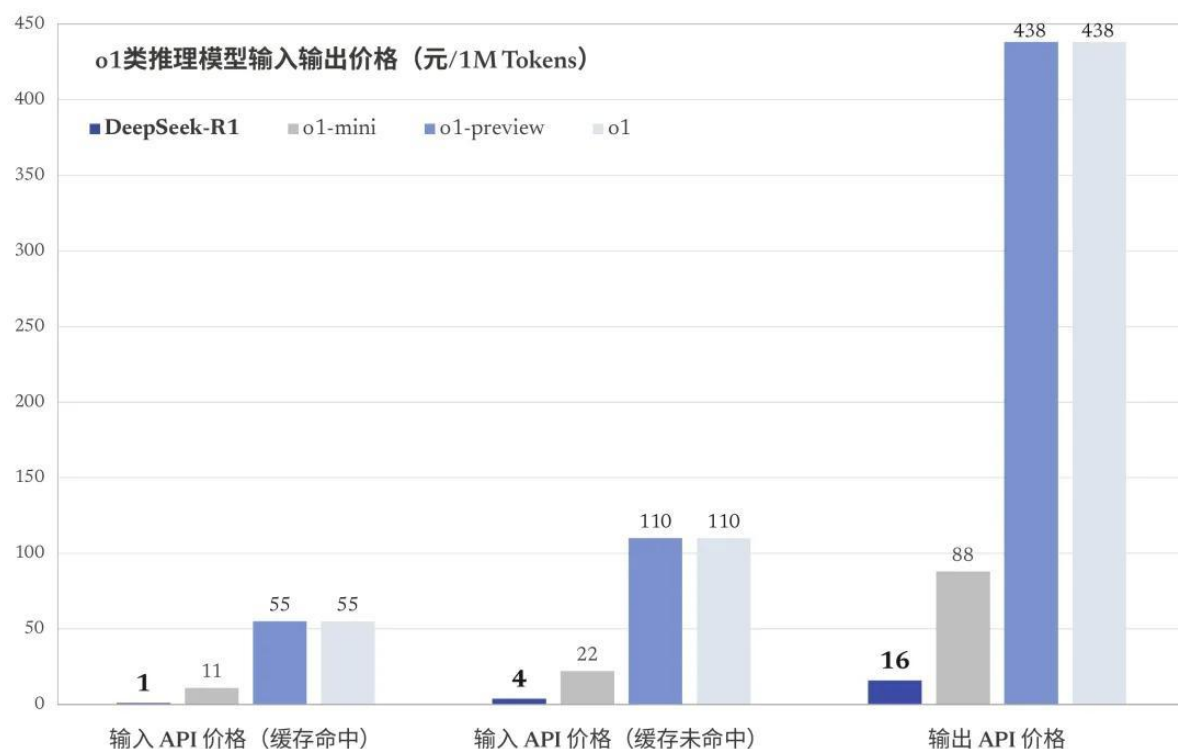
✓ DeepSeek V3模型定价：随着性能更强、速度更快的 DeepSeek-V3 更新上线，模型API服务定价也将调整为每百万输入tokens 0.5 元（缓存命中）/ 2 元（缓存未命中），每百万输出tokens 8元。

✓ DeepSeek-R1百万tokens输出价格约为o1的1/27。DeepSeek-R1 API 服务定价为每百万输入 tokens 1 元（缓存命中）/ 4 元（缓存未命中），每百万输出 tokens 16 元。对比OpenAI-o1每百万输入tokens为55元（缓存命中），百万tokens输出为438元。

图：DeepSeek-V3 API定价对比海内外主流模型



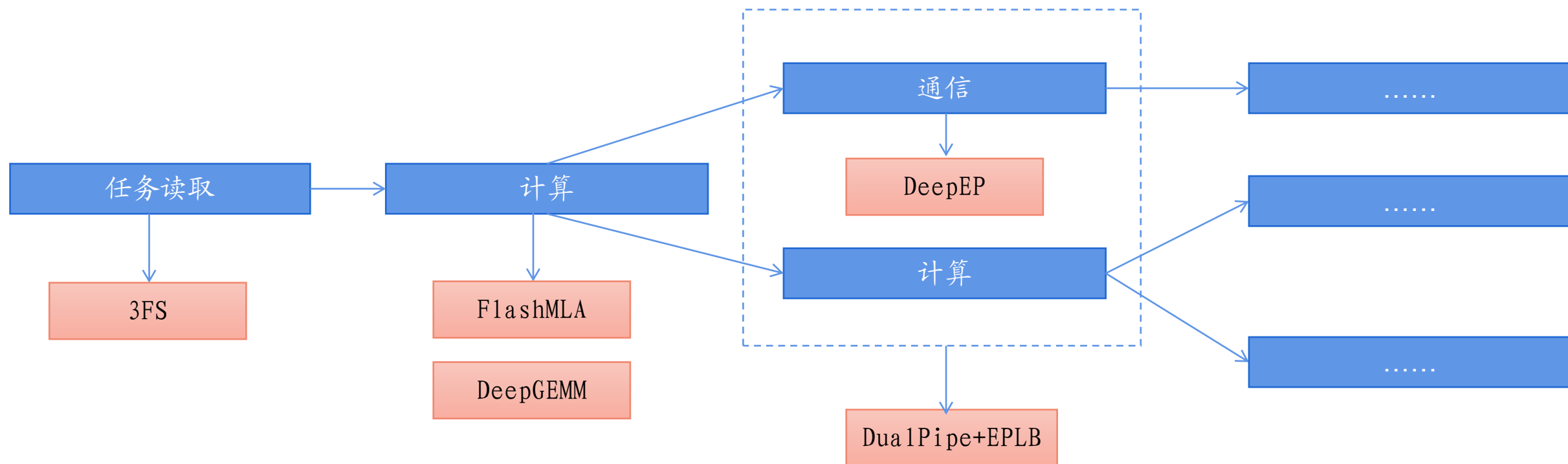
图：DeepSeek-R1定价对比同为推理模型的o1系列



2.2、以DeepSeek视角看模型降本增效范式

2025年2月24日，DeepSeek官宣开源周开放五大核心技术单元库，呈现DeepSeek高性能低成本核心架构。其中包括提升计算效率为主的FlashMLA架构及DeepGEMM库，通过降低向量计算维度实现降低显存的占用空间，进而实现GPU在显存利用率以及计算效率上的提升。而DeepEP面向通信层面，实现GPU节点间以及GPU节点内部的通信效率的大幅提升。DualPipe+EPLB则面向计算+通信任务之间的调度，通过大幅降低GPU空置率提升整体的计算效率；最后3FS则是面向数据读取环节，基于高速缓存和读取架构提升文件的随机读写效率。

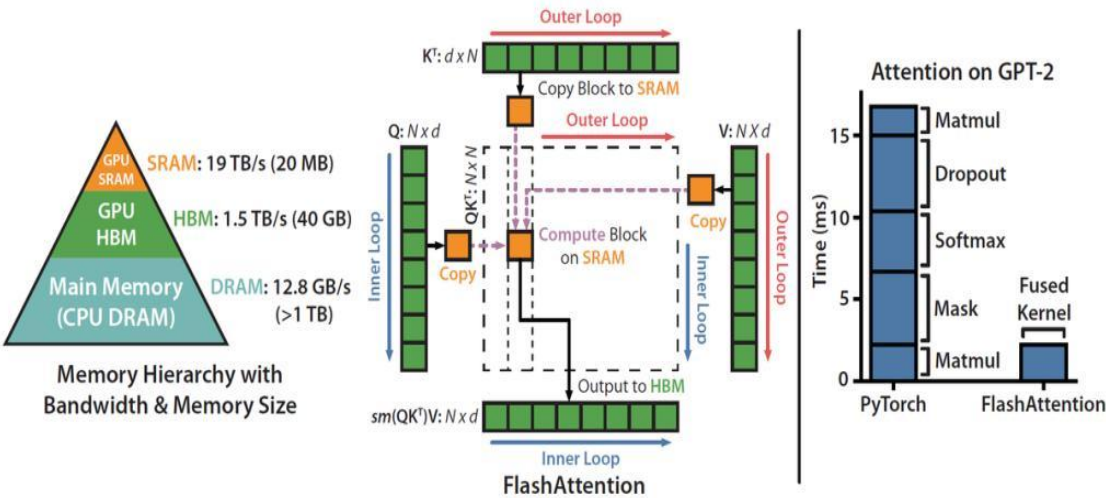
图：GPU计算维度看5大开源核心代码库的核心作用



2.2、DeepSeek开源库1：FlashMLA大幅提升显存带宽+推理效率

- 2025年2月24日，DeepSeek开启了为其5天开源周的第一个开源项目——FlashMLA。FlashMLA是一个针对 Hopper GPU 优化的高效 MLA（多头潜在注意力机制）解码内核，支持变长序列处理，现在已经投入生产使用。FlashMLA实现在H800 SXM5 GPU上具有3000GB/s的内存速度上限以及580TFLOPS 的计算上限。
- FlashMLA主要是通过优化MLA解码和分页KV缓存，提高LLM的推理效率，尤其是在H100/H800类高端GPU上发挥显著性能。当前GPU的计算速度远超内存速度，使得内存/访问成为影响 Transformer 模型执行操作的关键所在。DeepSeek 官方提到，FlashMLA 的灵感来自 FlashAttention 2&3 和 cutlass 项目。FlashAttention 是一种高效的注意力计算方法，专门针对 Transformer 模型的自注意力机制进行优化，核心目标是减少显存占用并加速计算。FlashAttention 注意力算法避免从HBM里读取或写入注意力矩阵，通过重新构建注意力计算等方式，比从HBM中读取中间注意力矩阵的传统方式速度更快。据DeepTech数据，经过测试，在减少内存/访问执行的条件下，与常见的 PyTorch Attention 相比，FlashAttention 的运行速度提高了 2-4 倍，所用内存则是前者的 5%-20%。

图：FlashMLA灵感来自FlashAttention



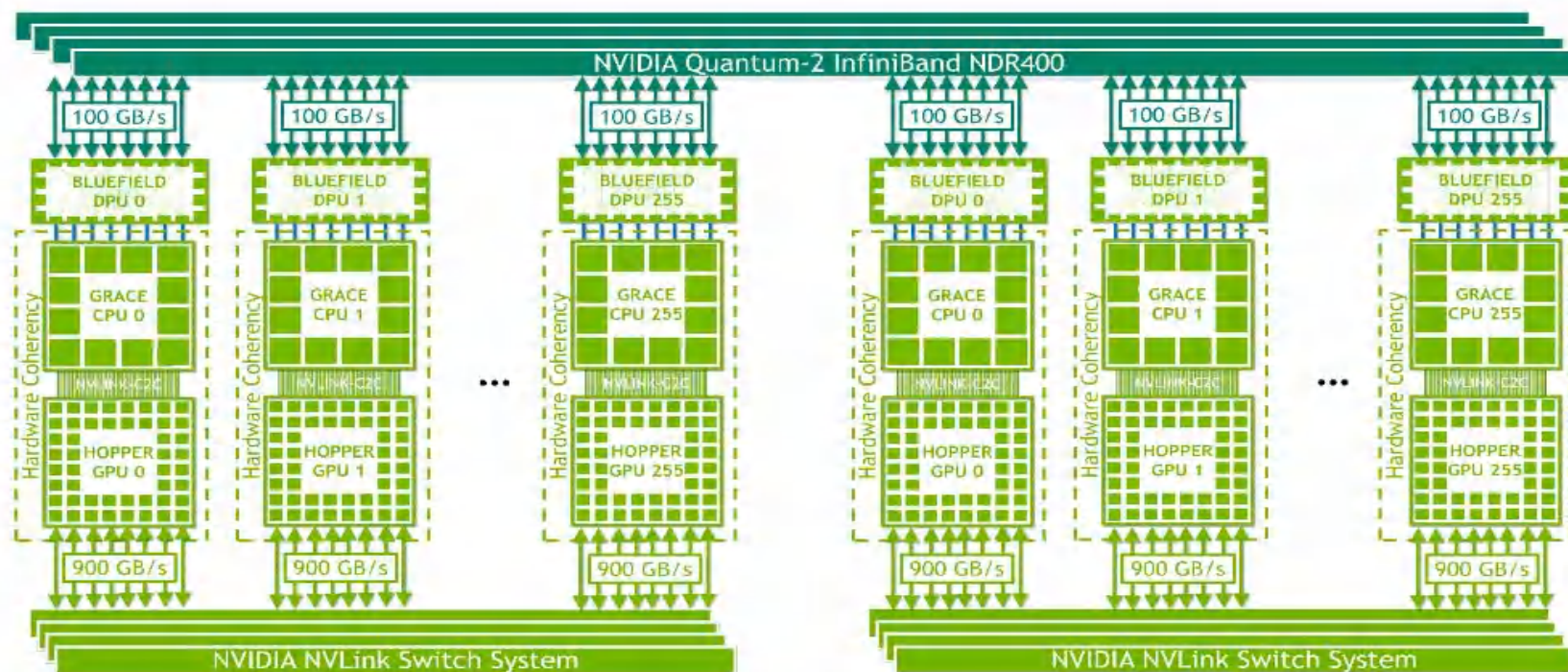
图：FlashMLA对比标准MHA的

	标准MHA	FlashMLA
时间复杂度	$O(n^2)$	$O(nk)$
内存复杂度	$O(n^2)$	$O(nk)$
适合长序列	受限	高度适合
典型用例	中短序列	长序列、实时推理

2.2、DeepSeek开源库2：DeepEP专为MoE及EP设计的通信库，高效优化通信机制

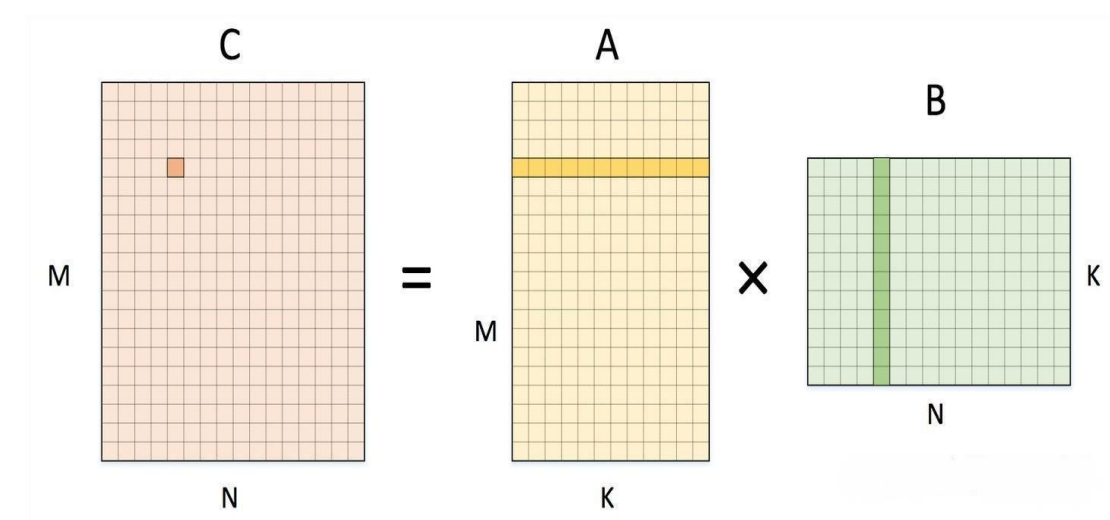
- DeepEP是DeepSeekMoE的训推EP通信库，支持FP8专为Hopper GPU设计，低延迟超高速训练推理。
- DeepEP是一个专为混合专家系统（MoE）和专家并行（EP）设计的通信库。在分布式系统中（如多GPU训练环境），所有处理单元之间需要高效地传递数据；尤其是MoE中因为不同专家需要频繁交换信息，并且MoE模型容易在专家并行中出现负载不均衡，导致每个专家分到的算力不均，不重要的专家难以发挥应有的性能。DeepEP通过高效且优化的All-to-All通信机制，支持节点内部和节点之间的通信，分别利用NVLink和RDMA实现。
- 据硅星GenAI公众号数据，在实测中，DeepEP在H800上4096个token同时处理的场景下，达到了153GB/s的传输速度，接近硬件理论极限（160GB/s）。

图：英伟达的GPU节点内部及节点间通信



✔ GEMM（通用矩阵乘法）是线性代数中的基本运算，也是科学计算、机器学习、深度学习等领域中常用的计算方式，但由于计算量较大使得GEMM的性能优化至关重要。DeepGEMM保持了DeepSeek“高性能+低成本”的技术特性，推动GEMM朝着更高效率方向发展；具体亮点包括：1）**高性能**：在Hopper架构的GPU上，DeepGEMM能够实现高达1350+FP8 TFLOPS的性能。2）**简洁性**：核心逻辑仅约 300 行代码，但性能却优于专家调优的内核。3）**即时编译（JIT）**：采用完全即时编译的方式，可以在运行时动态生成优化的代码，从而适应不同的硬件和矩阵大小。4）**无重依赖**：轻量级设计，没有复杂的依赖关系，可以让部署和使用变得简单。5）**支持多种矩阵布局**：支持密集矩阵布局和两种 MoE 布局，能够适应不同的应用场景，包括但不限于深度学习中的混合专家模型。

图：GEMM通用矩阵乘法示例



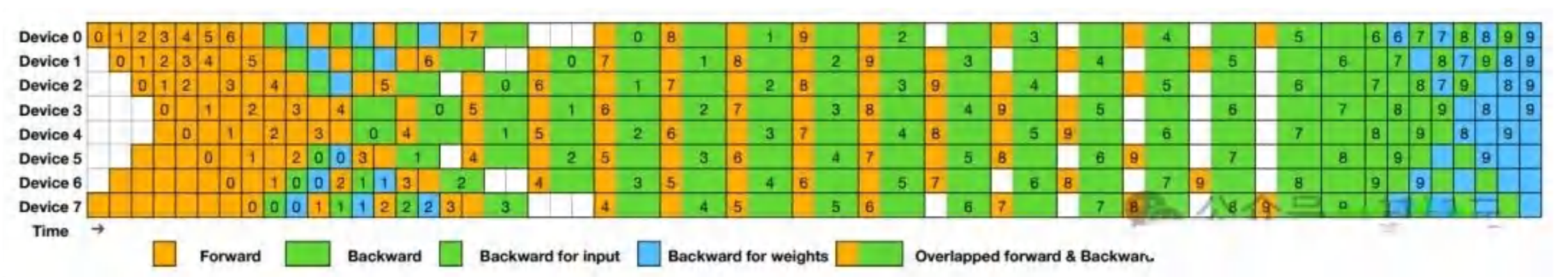
图：分组GEMM（MoE模型）中，连续性布局、掩码布局下速度可提升1.1倍~1.2倍

Grouped GEMMs for MoE models (contiguous layout)

#Groups	M per group	N	K	Computation	Memory bandwidth	Speedup
4	8192	4096	7168	1297 TFLOPS	418 GB/s	1.2x
4	8192	7168	2048	1099 TFLOPS	681 GB/s	1.2x
8	4096	4096	7168	1288 TFLOPS	494 GB/s	1.2x
8	4096	7168	2048	1093 TFLOPS	743 GB/s	1.1x

- ❖ DualPipe（双向管道并行算法）基于DeepSeek-V3技术报告提出的双向管道等值算法。现有方法无法精确控制计算任务和通信任务对硬件资源的使用，导致计算和通信无法实现无缝重叠，进而产生大量流水线气泡，增加了系统的延迟。DualPipe算法通过实现向后和向前计算通信阶段的双向重叠，大幅减少了训练过程中的空闲时间。
- ❖ EPLB（专家并行负载均衡器）具有动态负载均衡、分层与全局平衡结合以及流量优化三大特点。MoE 的专家网络分布在多个 GPU 上，每次计算需频繁执行Token分发与结果聚合，导致 GPU 计算资源大量闲置；因此如何将通信隐藏到计算的过程中、提升模型训练效率、节省计算资源，成为了 MoE 系统优化的关键。动态负载均衡功能基于混合专家（MoE）架构，通过复制高负载专家并采用启发式分配算法，优化了GPU之间的负载分布。在分层与全局平衡结合方面，EPLB不仅支持单个节点内的分层负载管理，还能实现跨节点的全局负载均衡，有效减少GPU闲置现象。此外，在流量优化方面，该技术能够在均衡负载的同时，通过调整专家分布降低节点间的数据通信量，从而提高整体训练效率。

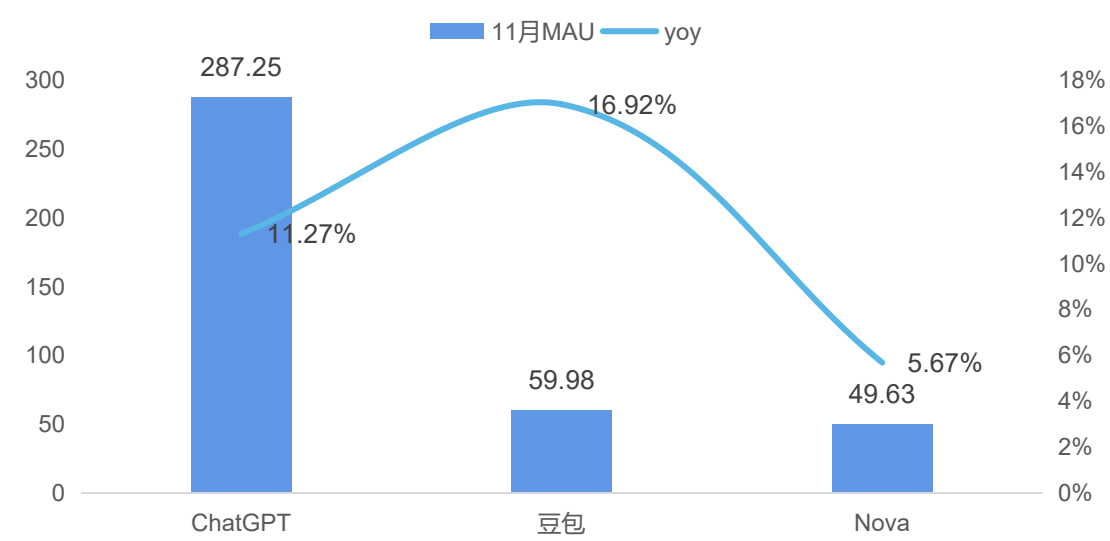
图：8个流水线并行阶段DualPipe在两个方向上的调度示例



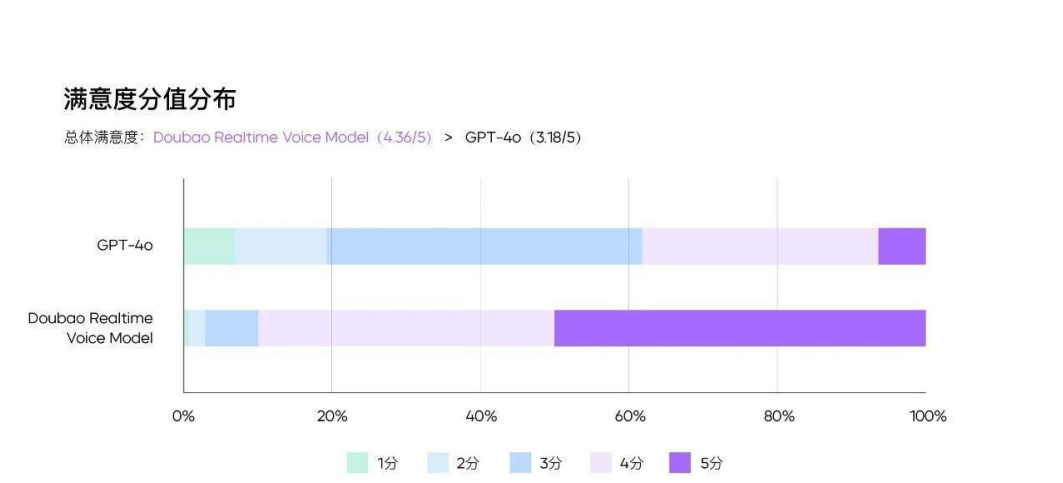
2.3、豆包大模型：实时语音、视频生成/理解领域布局，2024H2发力月活冲上全球第二

- 2023年以来，字节积极布局生成式AI。2023年2月组建“Seed”团队，专注于AI领域的语言与图像研究，6月上线“火山方舟”；11月成立专注于AI创新业务的新部门Flow，聚焦于AI大模型及AI应用层的产品研发。2024年5月，字节跳动正式发布了自研的豆包大模型，通过火山引擎正式对外提供服务。
- 2024H2豆包发力模型，月活冲刺6000万位居全球第二国内第一。2024年8月，豆包大模型正式实现用户和云端大模型的实时语音通话。2024年9月，豆包·视频生成模型正式上线；2024年12月，豆包视觉理解模型正式发布，通用模型能力全面对齐GPT-4o。2024年11月全球大模型月活跃排行榜上，豆包大模型MAU（月活）达5998万，仅次于OpenAI的ChatGPT，位列全球第二、国内第一。
- 2025年1月，豆包实时语音大模型上线开放，语音理解和生成一体化，实现端到端语音对话。相比传统级联模式，在语音表现力、控制力、情绪承接方面表现惊艳，并具备低时延、对话中可随时打断等特性；据外部用户真实反馈，该模型整体满意度较GPT-4o有明显优势。

图：2024年11月豆包大模型月活冲上全球第二（左：百万）



图：豆包实时语音模型客户使用满意度超GPT-4o



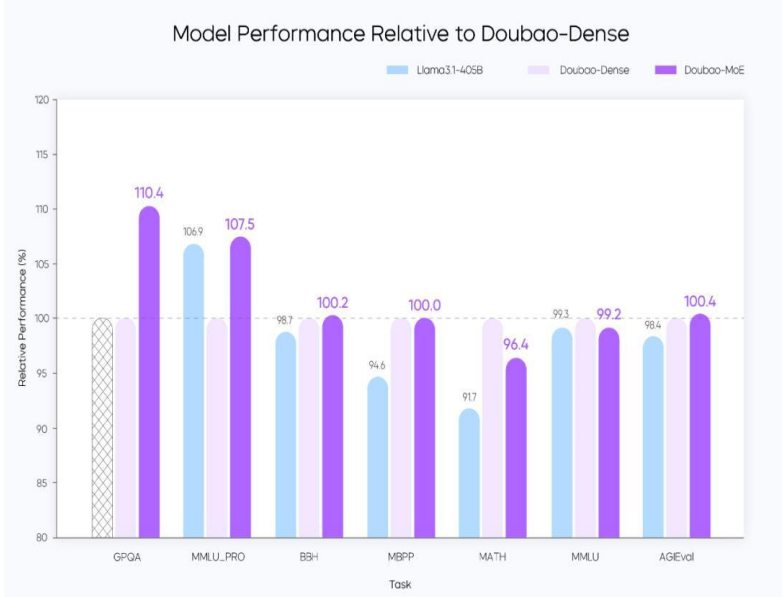
2.3、豆包大模型：新模型豆包pro 1.5依靠稀疏MoE架构实现小参数高性能，MoE杠杆提升至7x

- 2025年1月22日，豆包全新基础模型 Doubao-1.5-pro 正式发布，模型能力全面升级，融合并进一步提升了多模态能力。从训练和推理效率出发，Doubao-1.5-pro 使用稀疏 MoE 架构。在预训练阶段，仅用较小参数激活的 MoE 模型，性能即可超过 Llama-3.1-405B 等超大稠密预训练模型。
- 通过模型结构和训练算法优化，豆包将 MoE 模型的性能杠杆提升至 7 倍，此前业界的普遍水平为不到 3 倍。MoE 模型的性能通常可以用表现相同的稠密模型的总参数量和MoE模型的激活参数量的比值来确定，据豆包大模型团队公众号数据，如IBM的 Granite 系列模型中，800M激活的MoE模型性能可以接近2B总参数的稠密模型，性能比值大约在2.5倍（2000M/800M）。此前，业界的普遍水平在不到3倍的性能杠杆上。团队通过模型结构和训练算法优化，在完全相同的部分训练数据（9T tokens）对比验证下，用激活参数仅为稠密模型参数量 1/7 的 MoE 模型，超过了稠密模型的性能，将性能杠杆提升至 7 倍。

图：Doubao-1.5-pro多个基准测试结果

	Doubao-1.5-pro	Llama3.1-405B	GPT4o-0806	Gemini-exp-1205	Claude-3.5-Sonnet-latest	Qwen2.5	DeepseekV3
Knowledge	MMLU	88.6	88.6	88.7	86.8	88.5	88.5
	MMLU_PRO	80.1	73.3	74.9	76.4	78.0	75.9
	GPQA	65.0	51.1	53.1	62.1	65.0	49.0
MATH	Math	88.6	73.8	75.9	89.7	78.3	83.1
	OlympiadBench	59.8	34.1	40.7	64.7	43.5	50.0
Code	MBPP+	78.0	72.8	78.3	78.6	76.5	76.9
	McEval	70.2	58.7	68.2	67.0	68.2	61.7
	FullStackBench	65.1	53.6	61.8	62.6	60.3	56.9
Reasoning	BBH	91.6	89.2	91.7	92.6	88.3	92.3
	DROP	93.0	91.2	79.8	89.7	88.3	87.4
Instruction Following	IFEval	89.5	86.0	85.7	89.8	89.3	84.1
	SysBench	67.6	58.9	62.2	69.0	47.2	66.3
Chinese	CMMLU	90.9	75.4	77.3	84.3	81.2	84.3
	C-Eval	91.8	72.7	76.0	83.9	80.0	84.1

图：豆包模型与Dense模型性能对比



图：豆包视觉多模态进一步提升

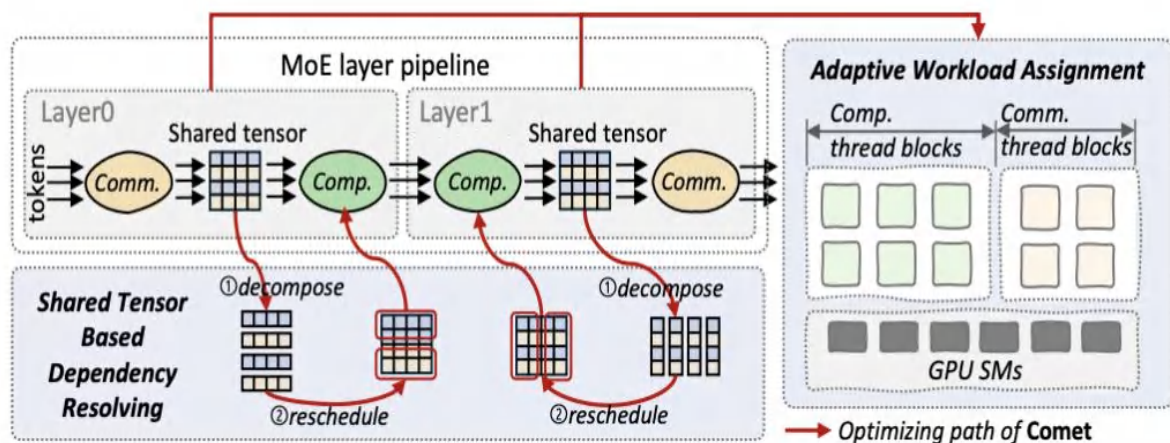
Benchmark	Doubao-1.5-pro	GPT4o-1120	Claude3.5-Sonnet	Gemini-2-flash	Qwen2-VL-72B	InternVL-2.5-78B
College-level Problems	MMLUval	73.8	70.7	70.4	70.7	64.5
	MMLU-Pro	59.3	54.5	54.7	57.0	46.2
	MathVision	48.6	30.4	38.3	41.3	25.9
Mathematical Reasoning	OlympiadBench	48.5	25.9	27.8	43.6	11.2
	MathVista	78.8	63.8	65.4	73.1	70.5
Document and Diagrams Reading	TextVQAval	84.7	81.4	76.5	75.6	85.5
	ChartQAval (avg)	88.0	86.7	90.8	85.2	88.3
	InfoQAval (test)	88.0	80.7	74.3	77.8	84.5
	DocVQAval (test)	96.7	91.1	95.2	92.1	96.5
	CharxivQAval (DQ)	54.4 / 84.3	52.0 / 86.5	60.2 / 84.3	55.2 / 81.8	43.0 / 81.3
General Visual Question Answering	RealWorldQA	78.9	75.4	66.6	74.5	77.8
	MixStar	71.9	63.9	65.1	69.4	68.6
	MixBench-en	87.5	83.5	81.7	83.0	85.9
Spatial and Counting Understanding	MixBench-cn	86.0	82.1	83.4	82.9	83.4
	Bink	68.4	68.0	59.6	62.6	61.1
	CountBench	89.6	85.1	86.8	88.2	88.6
Video Understanding	Video-MME	74.1	73.4	61.7	78.2	71.2
	EgoSchema-subset	75.4	74.8	64.4	71.8	80.6

2.3、豆包大模型：开源MoE通信优化技术COMET、万卡集群部署已节省数百万 GPU 小时

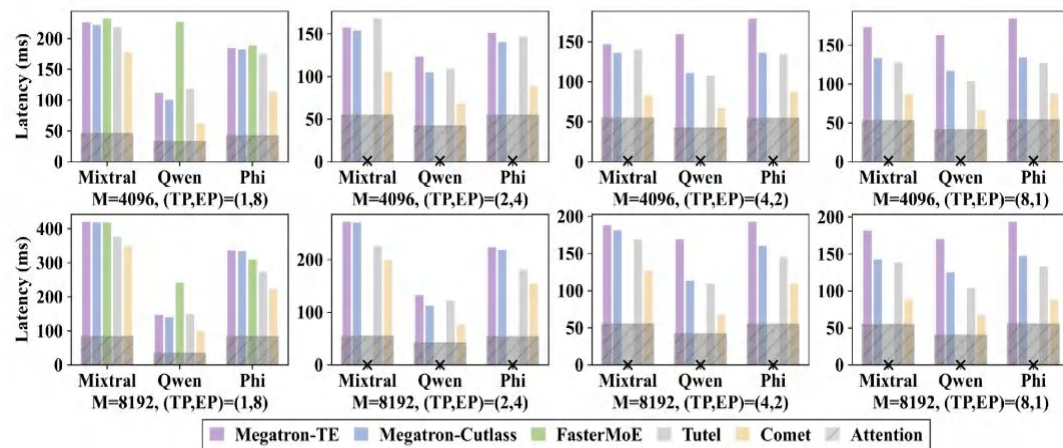
COMET面向模型通信优化系统，再实现通信效率的显著提升。面对MoE架构跨设备通信开销巨大的问题，严重制约了训练效率和成本，COMET针对 MoE 模型的通信优化系统，通过细粒度计算-通信重叠技术，助力大模型训练优化；其中引入两项关键机制，1）共享张量依赖解析：通过分解和重调度共享张量，解决通信与计算之间的粒度错配问题，实现细至单 Token 级的重叠。2）自适应负载分配：动态分配 GPU 线程块资源，精准平衡通信与计算负载，消除流水线气泡。

COMET已落地万卡集群，累计节省数百万 GPU 小时。豆包团队在多个大规模 MoE 模型中评估了 COMET 的端到端性能，COMET在 8 卡 H800 的实验集群中，端到端 MoE 模型（Mixtral-8x7B、Qwen2-MoE 等）的前向时延较其他基线系统可降低 31.8%-44.4%，且在不同并行策略、输入规模及硬件环境下均表现稳定。目前COMET 已实际应用于万卡级生产集群，助力 MoE 模型高效训练，并已累计节省数百万 GPU 小时。

图：COMET设计结构



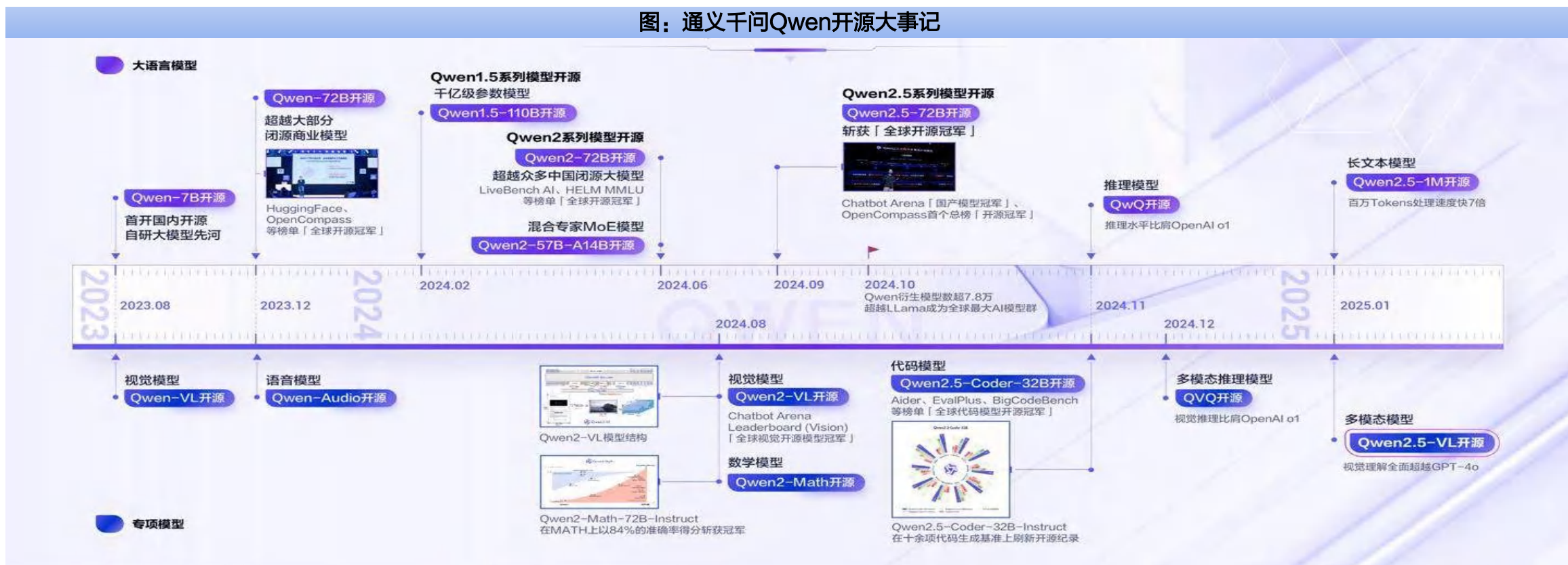
图：COMET在多个MoE模型中的测试结果



2.4、Qwen：AI为阿里巴巴未来战略核心，Qwen系列掀起国内模型开源革命

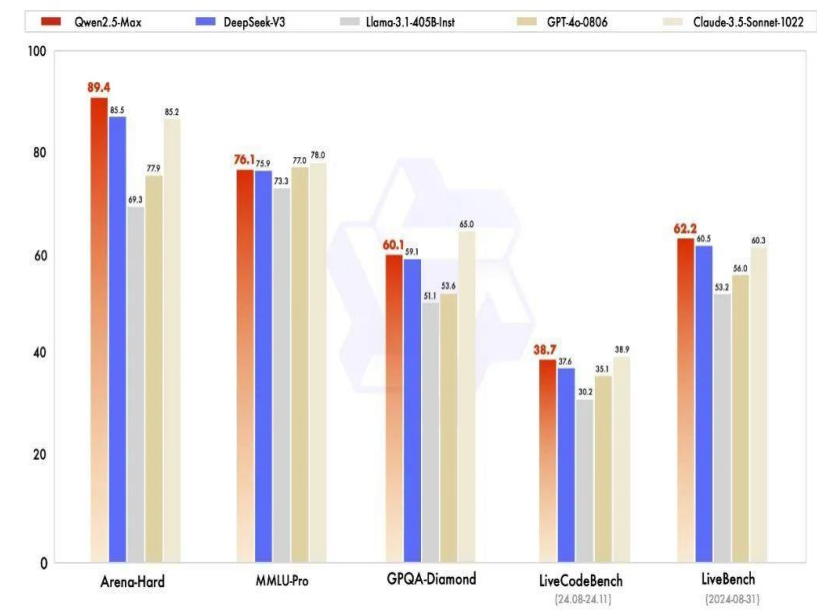
- 以AI为战略核心，阿里布局Qwen系列大模型。阿里将AI视为未来战略核心，依托中国市场份额第一的阿里云为通义千问提供算力支持。阿里大模型技术骨干成员来自达摩院、清华大学等机构，参与过全球首个10万亿参数多模态模型M6的研发。2023年8月，Qwen首次亮相并发布了7B版本；2024年6月，阿里开源发布Qwen2，2023年8月以来，Qwen/Qwen1.5/Qwen2/Qwen2.5相继开源，覆盖多模态/数学/代码模型等数十种，掀起了国内模型的开源革命。
- Qwen系列模型位居全球开源模型榜首。据全球最大AI开源社区Hugging Face数据显示，截至2025年2月，阿里Qwen开源大模型的衍生模型数量已突破10万，持续领先美国Llama等开源模型，稳居全球最大开源模型榜首。

图：通义千问Qwen开源大事记



- 2025 年 1 月 29 日：阿里云通义千问超大规模的 MoE 模型 Qwen2.5-Max 正式上线。该模型采用 MoE 架构，预训练数据量达 20 万亿 tokens，基座模型在 11 项基准测试中全面领先开源模型，指令模型则在多项任务中与 Claude-3.5-Sonnet 持平。据三方基准测试平台，Qwen2.5-Max 在 Chatbot Arena 盲测榜单中以 1332 分位列全球第七，超越 DeepSeek V3、Claude-3.5-Sonnet 等国际主流模型。
- 2025年2月25日，Qwen上线新推理功能——深度思考 (QwQ)，发布预览版推理模型 。QwQ是在QWQ-MAX-PREVIEW（推理模型）支持下，同时是基于Qwen2.5-Max的推理模型。类似DeepSeek R1和kimi的推理模型，QwQ可同时支持深度思考和联网搜索，并会展示完整的思维链。Qwen团队称，QWQ-MAX官方版本即将发布，同步会发布Android和iOS应用程序，还会发布更小的可在本地设备部署的模型，如QWQ-32B等。

图：Qwen2.5性能测试



图：Qwen2.5-Max性能测试

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	+6/-9	10022	Google	Proprietary
1	1	Gemini-2.0-Pro-Exp-02-05	1379	+7/-6	7853	Google	Proprietary
3	1	ChatGPT-4o-latest_(2024-11-20)	1365	+4/-3	38148	OpenAI	Proprietary
3	1	DeepSeek-R1	1361	+8/-7	4193	DeepSeek	MIT
3	7	Gemini-2.0-Flash-001	1357	+8/-7	5576	Google	Proprietary
4	1	o1-2024-12-17	1352	+6/-6	12021	OpenAI	Proprietary
7	5	o1-preview	1335	+4/-3	33174	OpenAI	Proprietary
7	7	Qwen2.5-Max	1332	+12/-11	3394	Alibaba	Proprietary
8	9	DeepSeek-V3	1316	+5/-5	16583	DeepSeek	DeepSeek
9	9	Gemini-2.0-Flash-Lite-Preview	1306	+9/-8	5083	Google	Proprietary
9	12	GLM-4-Plus-0111	1305	+11/-10	2791	Zhipu	Proprietary
9	13	Step-2-16K-Exp	1304	+8/-9	5128	StepFun	Proprietary

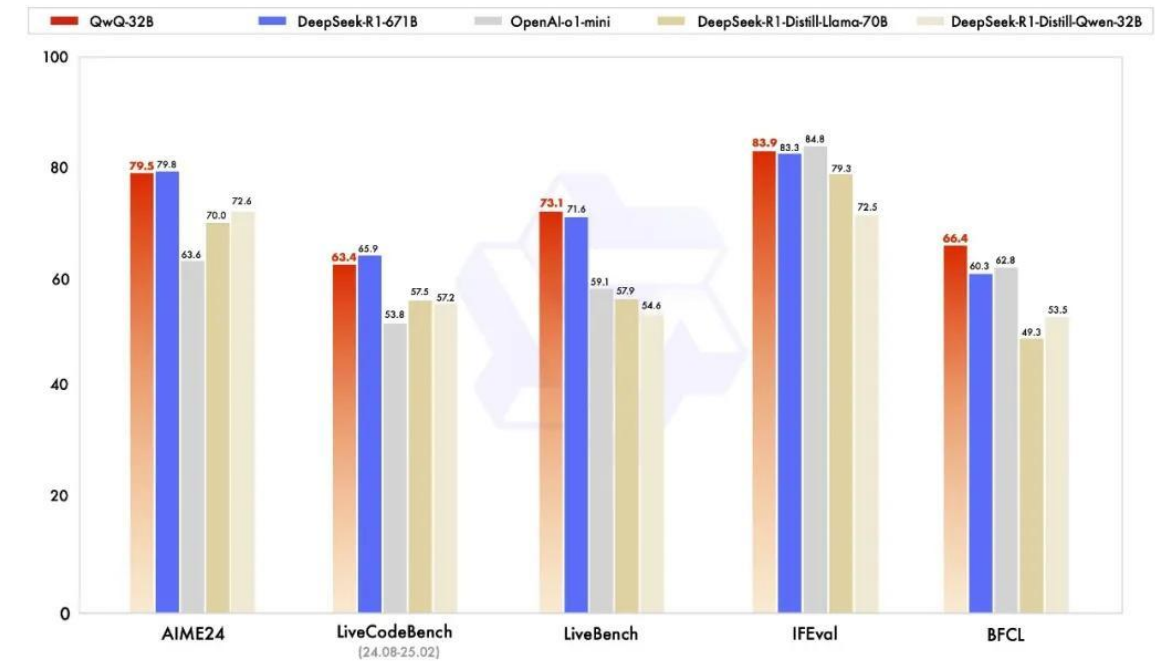
图：Qwen推理模型

阿里通义千问
上线推理模型
DeepSeek R1 平替

2.4、Qwen：QwQ-32B登顶全球领先开源模型，强化学习能力再验证

- 2025年3月6日，阿里云发布并开源了全新的推理模型通义千问QwQ-32B，性能比肩DeepSeek-R1满血版。通过大规模强化学习，千问QwQ-32B在数学、代码及通用能力上实现质的飞跃，整体性能比肩Deepseek-R1。在保持强劲性能的同时，千问QwQ-32B还大幅降低了部署使用成本，其参数量约为 DeepSeek-R1 满血版的 1/21 且推理成本是后者的1/10。
- QwQ-32B已成为全球领先开源模型，核心为强化学习。在LiveBench榜单中，QwQ-32B以综合评分92.3分位列全球第五，超越OpenAI GPT-4.5（91.8分）、Google Gemini 2.0（90.1分）等头部闭源模型。而QwQ-32B的核心为强化学习，Qwen 团队在冷启动的基础上进行了大规模强化学习，专攻数学 & 编程领域，同时也强化通用能力确保整体性能均衡提升。

图：QwQ-32B测试集对比



图：QwQ-32B已经登顶全球最强开源模型

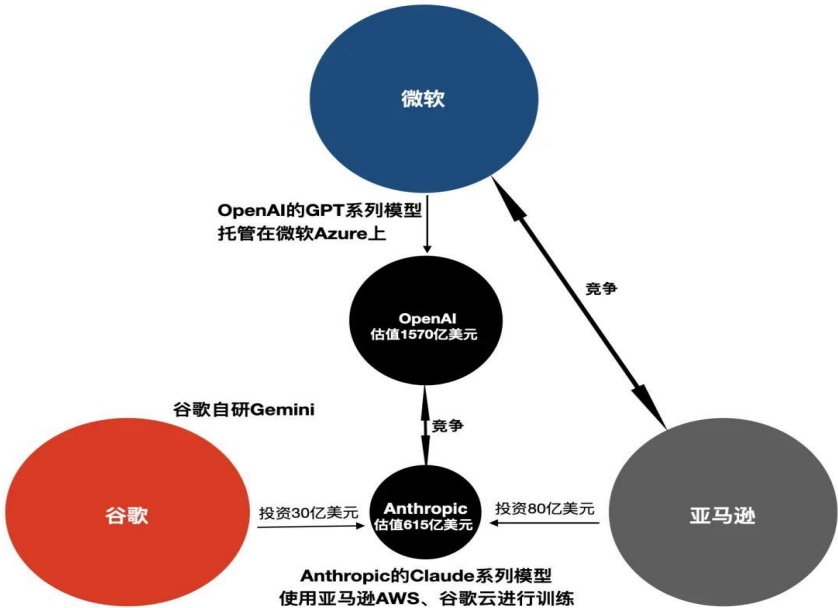
Model	Organization	Global Average	Reasoning Average	Coding Average	Mathematics Average	Data Analysis Average	Language Average	IF Average
claude-3-7-sonnet-thinking	Anthropic	76.10	87.83	74.54	79.00	74.05	59.93	81.25
o3-mini-2025-01-31-high	OpenAI	75.88	89.58	82.74	77.29	70.64	50.68	84.36
o1-2024-12-17-high	OpenAI	75.67	91.58	69.69	80.32	65.47	65.39	81.55
grok-3-thinking	xAI	-	-	67.38	-	-	-	-
qwq-32b	Alibaba	71.96	83.50	72.23	77.82	65.03	51.35	81.83
deepseek-r1	DeepSeek	71.57	83.17	66.74	80.71	69.78	48.53	80.51
o3-mini-2025-01-31-medium	OpenAI	70.01	86.33	65.38	72.37	66.56	46.26	83.16
gpt-4.5-preview	OpenAI	68.95	71.08	75.18	69.33	64.33	61.45	72.33
gemini-2.0-flash-thinking-exp-01-21	Google	66.92	78.17	53.49	75.85	69.37	42.18	82.47
claude-3-7-sonnet	Anthropic	65.56	66.00	67.49	63.26	63.37	56.76	76.49
gemini-2.0-pro-exp-02-05	Google	65.13	60.08	63.49	70.97	68.02	44.85	83.38

三、海外大模型进展：资源头部集中，压铸AGI

3.1、海外大模型：CSP大厂重点参与，格局头部集中

● 海外头部大模型依托资源壁垒形成强马太效应。大模型随着2022年ChatGPT的发布进入大众视野，同时与OpenAI资源匹敌的Google、Meta同样成为了底层模型的主要竞争者，Google、Meta基于自身超过30亿的用户体量，不断基于用户数据反哺模型训练；而亚马逊则通过投资Anthropic来布局AI领域。当前海外主流的AI模型竞争玩家包括技术能力以及用户数全球领先的OpenAI系GPT模型、依托亚马逊/谷歌投资的Anthropic模型Claude、谷歌自研模型Gemini、Meta自研模型Llama、马斯克旗下自研模型xAI等。

图：亚马逊、微软、谷歌之间的AI代理人竞争



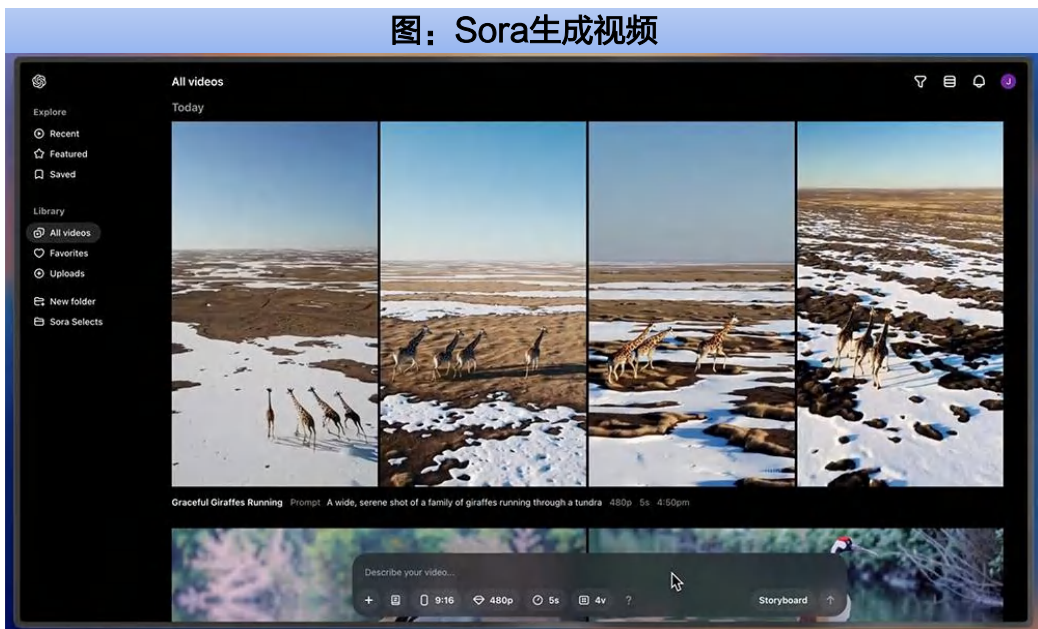
	大模型	核心能力	技术特点	典型应用场景
海外	OpenAI GPT-4	自然语言语义理解、多模态交互、知识集成与复杂推理	强化Transformer架构、多模态融合、对抗训练机制、优化训练策略	编程辅助、学术研究、商业分析、内容创作
	OpenAI o3-mini	强大推理链条、多种编程语言、安全可控输出、多语言处理	推理链条技术，安全评估机制，强化学习，可调节智能推理	编程开发、教育领域、医疗行业、多语言交流、智能客服
	OpenAI Sora	视频内容生成、场景理解与想象	先进的视频生成算法，能够基于文本描述生成高质量视频	影视制作、广告宣传、虚拟现实内容创作
	谷歌Gemini2.0	多模态无缝融合、自然语言处理、文生图、数学复杂任务推理	优化Transformer架构、联合训练技术、层次化多任务学习、生成对抗网络技术	AI搜索、智能助手、医疗诊断
	Anthropic Claude3.5	复杂问题推理、强大视觉理解、编程能力出色	多任务学习架构、UnstructuredGeneralization算法	数据分析、游戏开发、内容创作
	xAI Grok-2	快速信息检索、幽默交互、实时动态感知	与X平台数据深度结合，响应速度快	社交媒体运营、实时问答、娱乐互动
	Meta Llama3.3	开源模型定制、128K上下文窗口、推理表现出色	Transformer架构、15万亿token训练数据、在线偏好优化	文案策划、智能客服、代码优化
	MistralAI LeChat	多领域对话、高质量图像生成、带引文的网页搜索	低延迟响应迅速、Mistral预训练模型知识、先进视觉与OCR技术、生成对抗算法	创意设计、代码补全、数据统计分析、智慧助手

3.2、OpenAI：全球AI大模型风向标，自然语言/多模态/推理模型上均作为引领角色

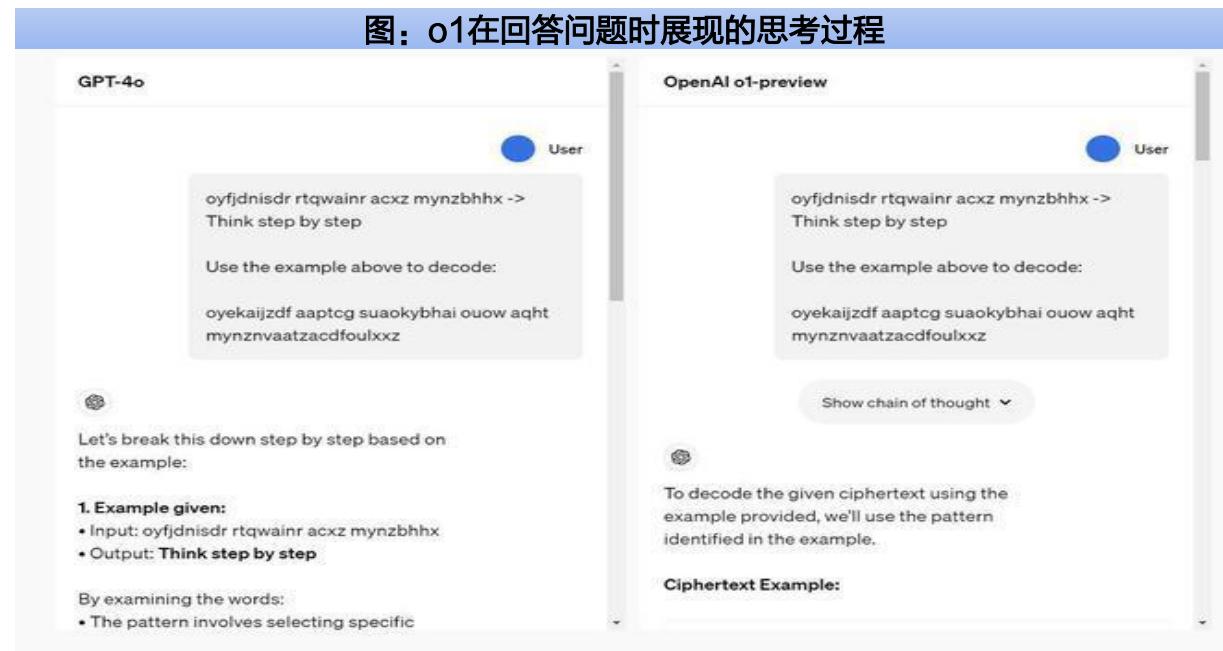
❖ **OpenAI以非营利性组织形式创立，GPT1-GPT4迭代实现能力跃升。**2015年，OpenAI由马斯克、Sam Altman等人共同创立，以非营利性组织的形式成立。2018年，OpenAI基于Transformer推出第一代GPT模型GPT-1，核心在于采用了Transformer的解码器架构，通过生成式预训练以及微调的方式开启了自然语言处理领域的新篇章。2019年GPT-2发布，参数规模扩大到15亿，并引入多样化数据集进行训练；2020年发布GPT-3，参数规模达到1750亿，在自然语言处理任务上取得了突破性的进展。2023年4月，GPT-4的发布在多个维度上实现了跨越式提升，无论是专业领域知识问答以及创意写作等方面能力远超以往模型水平，多模态能力上能同时处理文本和图像信息。

❖ **多模态+推理模型上均实现引领作用。**2024年2月，OpenAI发布视频生成模型Sora，首次由AI生成了长达1分钟的多镜头长视频，输入寥寥数语便能生成效果炸裂视频，镜头感堪比电影。2024年9月，OpenAI发布o1推理模型，在回答问题之前像人类一样将问题拆解成多个思维步骤，经此，美国数学邀请赛（AIME）中，GPT-4o的平均解题成功率为13.4%，而o1的正确率则高达83.3%。

图：Sora生成视频



图：o1在回答问题时展现的思考过程



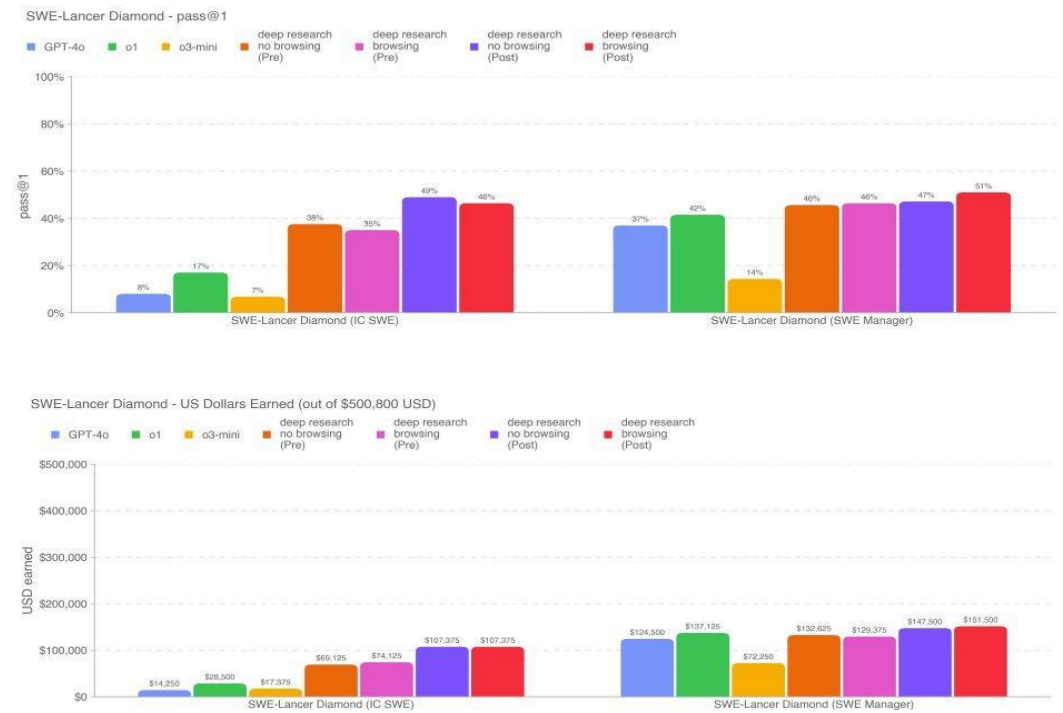
3.2、OpenAI：GPT-5即将发布，Agent领域加速布局

- **GPT-5即将迎来发布，实现OpenAI技术的极大集成。**2025年2月，OpenAI的CEO Sam Altman表示未来几个月OpenAI将发布GPT-5，GPT-5将融合OpenAI包括包括o3在内的大量技术，同时o3将作为GPT-5的一部分推出。
- **Agent领域大步迈进，推出深度研究领域智能体Deepresearch。**2025年2月，OpenAI推出面向深度研究领域的智能体产品Deep Research，使用推理来综合大量在线信息并为用户完成多步骤研究任务的智能体，旨在帮助用户进行深入、复杂的信息查询与分析，用户将可以在几十分钟内完成人类需要数小时才能完成的工作。据第三方消息，OpenAI计划为专业人士推出量身定制版Agent，用于执行销售线索分类、软件工程和博士级研究等高级任务，最高定价可达20000美金/月。

图：GPT-5将迎来发布



图：Deepresearch测评结果



3.3、Google：Gemini面向智能体时代新作，原生多模态领域前瞻布局



- 早期谷歌BERT是经典的NLP模型，主要面向单向/双向文本理解，Gemini是谷歌面向智能体以及多模态时代的下一代模型。2023年12月，谷歌正式发布原生多模态模型Gemini 1.0，原生支持文本、图像、音频、视频跨模态处理，在多方面能力测评中超越GPT-4。2024年2月发布Gemini 1.5 Pro，上下文窗口扩展至100万token，支持更复杂的推理与长文本处理。
- 谷歌全线产品进入Gemini2.0时代，原生多模态能力再提升。2024年12月，谷歌正式发布人工智能大模型 Gemini 2.0，首个版本为 Gemini 2.0 Flash；2025 年 2 月谷歌宣布产品线全面升级，所有用户进入“Gemini 2.0” 时代，推出了正式版 Gemini 2.0 Flash、Gemini 2.0 Flash - Lite 以及新一代旗舰大模型 Gemini 2.0 Pro 实验版，同时在 Gemini App 中推出其推理模型 Gemini 2.0 Flash Thinking 实验版。Gemini2.0集成谷歌搜索、代码执行以及第三方用户定义函数等工具，进一步拓展原生工具集成，同时拓展上下文窗口，Gemini 2.0 Flash 和 Flash - Lite 支持100 万tokens，Gemini 2.0 Pro 实验版支持 200 万 tokens；性能提升上，Gemini 2.0 Flash 速度是 Gemini 1.5 Pro 的两倍，

图：Gemini2.0发布时在Chatbot Arena大模型排行榜上翻炒OpenAI和DeepSeek

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	+6/-9	10022	Google	Proprietary
1	1	Gemini-2.0-Pro-Exp-02-05	1379	+7/-6	7853	Google	Proprietary
3	1	ChatGPT-4o-latest (2024-11-20)	1365	+4/-3	38148	OpenAI	Proprietary
3	1	DeepSeek-R1	1361	+8/-7	4193	DeepSeek	MIT
3	7	Gemini-2.0-Flash-001	1357	+8/-7	5576	Google	Proprietary
4	1	o1-2024-12-17	1352	+6/-6	12021	OpenAI	Proprietary

图：Gemini2.0多版本定价

Gemini Developer API (Per Million Tokens)				
MODEL	TEXT/IMAGE/VIDEO INPUTS	AUDIO INPUTS	TEXT OUTPUTS	CONTEXT CACHING*
Gemini 2.0 Flash	\$0.10	\$0.70**	\$0.40	Text/Image/Video \$0.025 Audio \$0.175
Gemini 2.0 Flash-Lite	\$0.075	\$0.075	\$0.30	\$0.01875
Gemini 1.5 Flash (Provided for reference)	\$0.075 (Prompts <= 128k)	\$0.075 (Prompts <= 128k)	\$0.30 (Prompts <= 128k)	\$0.01875 (Prompts <= 128k)
	\$0.15 (Prompts > 128k)	\$0.15 (Prompts > 128k)	\$0.60 (Prompts > 128k)	\$0.0375 (Prompts > 128k)

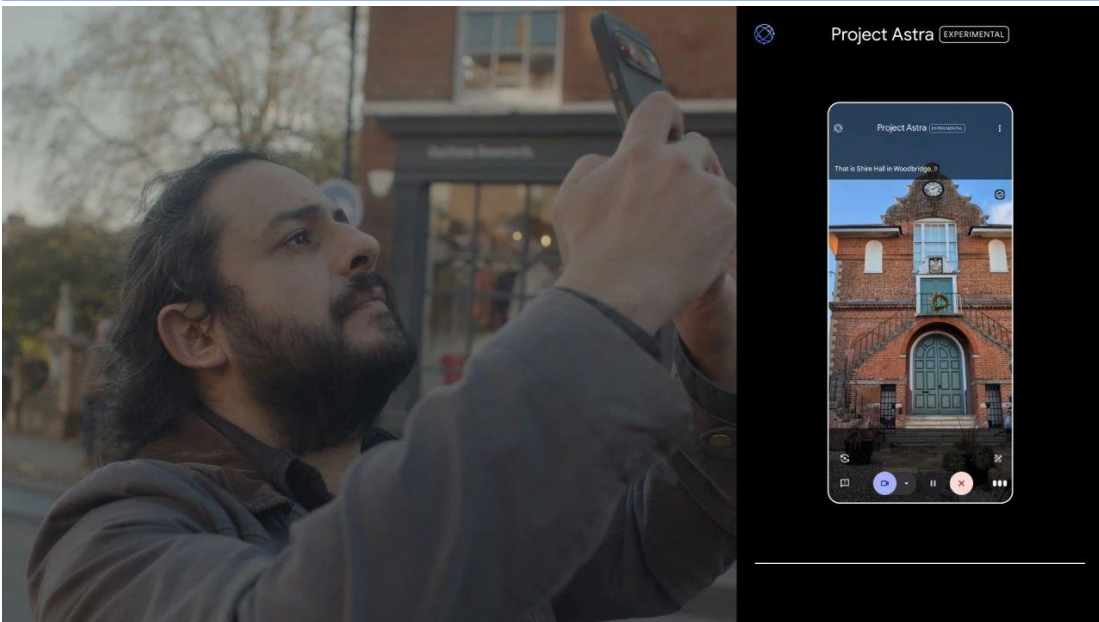
* Caching coming soon.
** The audio input pricing shown above will become effective February 20, 2025. Audio tokens will be charged the same as other modalities until then.

3.3、Google：核心优化多模态能力+智能体构建，加速推动用户（端侧+云端）增长

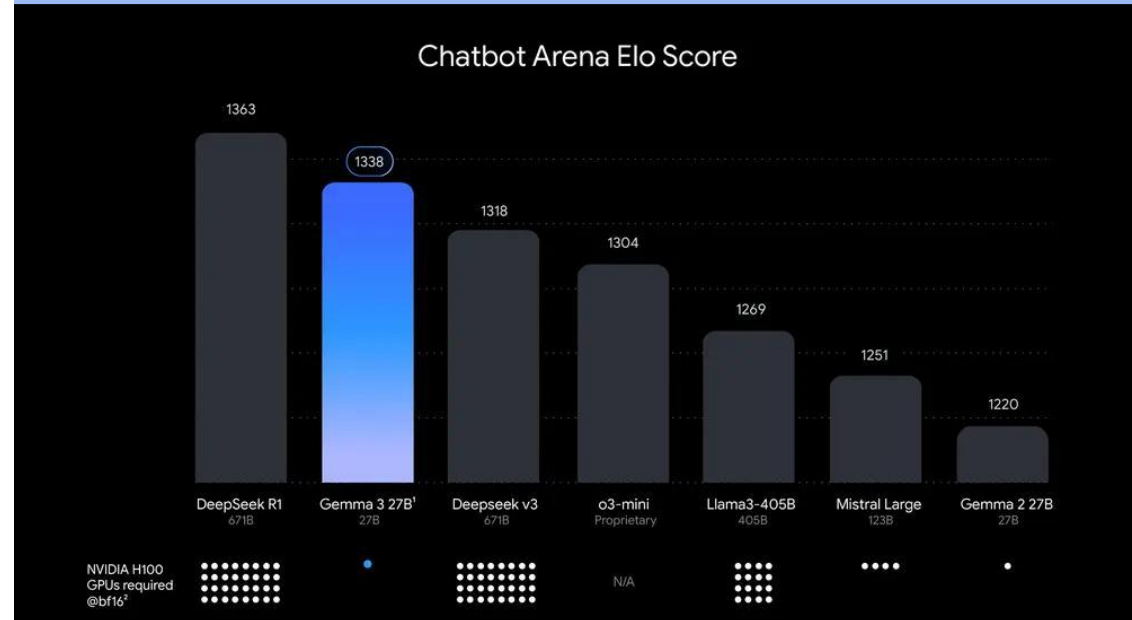
● **谷歌加速智能体构建**：2024年12月发布Gemini2.0时，首次详细披露了智能体生态的布局，包括1）Project Astra：多模态语音助手（支持 10 分钟会话记忆），集成搜索、地图等工具。2）Project Mariner：浏览器智能体，解析网页元素完成表单填写、代码调试等任务。3）Jules：GitHub 代码智能体，自动化处理开发任务（计划 / 编码 / 合并），提升异步协作效率。

● **谷歌在2024年12月的内部战略会议上明确提出，Gemini AI 计划成为第 16 个实现5 亿月活跃用户的谷歌产品。**CEO 桑达尔强调 2025 年将加速 Gemini 的消费者端扩展，目标是通过多模态功能（如语音、图像、视频交互）和全场景适配（端侧设备与云端协同）实现用户增长。Gemma基于 Gemini 的研究成果和技术架构开发，共享相同的数据集、底层 Transformer 架构优化以及安全技术，是独立训练的轻量级版本，针对特定场景（如单 GPU 运行）优化。2025年3月12日，谷歌发布Gemma3模型，并称之为“世界上最好的单加速器模型”，在配备单个GPU的主机上的性能表现超越了Facebook的Llama、DeepSeek和OpenAI等竞争对手。

图：Project Astra多模态智能体应用

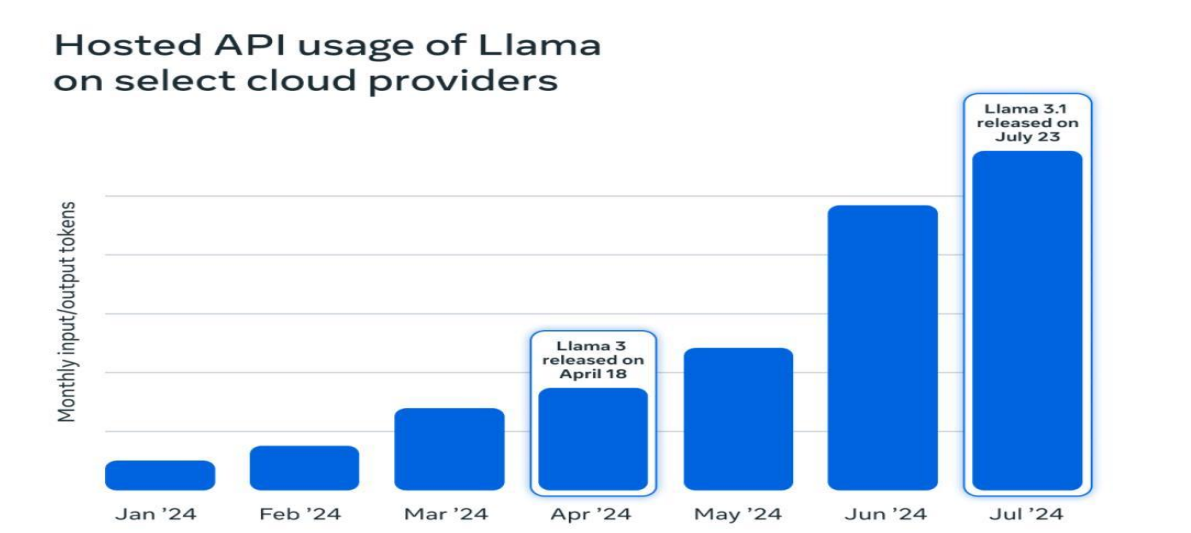


图：Gemma3在测试靠前的同时需要的GPU数量极少



- 十年布局，Meta跻身全球AI巨头。2013年Meta的AI布局起始，经历了实验室成立、技术路线图发布、产品应用探索、组织架构调整和生成式AI崛起等阶段。2022年AI推荐引擎成熟，成为驱动Meta产品参与度的总要推动力。2022年2月Meta公开发布了Llama 1，基于Transformer架构包括7B-65B四种参数规模。2023年7月18日，Llama 2发布（与微软合作），有70-700亿参数规模，用于训练基础模型的数据增加了40%，数据集包含高达2万亿个Token。
- Llama开源与OpenAI闭源战略形成直接对比，形成开源模型标杆。Llama不仅赋能了Meta自身各大社交平台，包括为Facebook、Instagram和WhatsApp用户提供服务的Meta AI助手，同时包括埃森哲(ACN)、DoorDash(DASH)和高盛(GS)在内的众多知名企业，都在使用Llama开发自己的人工智能软件。

图：Llama在HuggingFace上下载量接近3.5亿，较2023年同期增长10x（2024/8）



图：Meta的AI布局蓝图



3.4、Meta：Llama3.3提升后训练实现低成本高性能、智能体生态加速构建

- **Llama3.1**：2024年7月24日，Llama推出Llama3.1 405B，同时发布了全新升级的Llama 3.1 70B和8B模型。Llama 3.1 405B支持上下文长度为128K Tokens，在基于15万亿个Tokens、超1.6万个H100 GPU上进行训练；同时Llama 3.1 将模型从BF16量化为FP8。
- **Llama3.2**：2024年9月26日，Llama3.2发布，包括110亿和900亿参数的多模态版本10亿参数和30亿参数的轻量级纯文本模型；而专为端侧打造的3B版本，在性能测试中表现也优于谷歌的Gemma 2 2B和微软的Phi 3.5-mini。
- **Llama3.3**：2024年12月，Llama3.3发布，70B版本就能实现Llama3.1 405B的性能。同时Llama3.3推理部署成本出现大幅下降，输入成本降低了 10 倍，输出成本降低了近 5 倍。Llama 3.3 采用优化的 transformer 架构，融合了SFT和RLHF等技术，支持 128K tokens的上下文长度。在多个行业基准测试中，Llama-3.3-70B 的表现超过了谷歌的 Gemini 1.5 Pro、OpenAI 的 GPT-4o 和亚马逊的 Nova Pro。
- **智能体生态构建**：基于 Llama 的智能体项目（如 Meta Live）已实现实时语音交互、跨设备协作（如雷朋眼镜集成），定位为“个人数码助手”。未来计划通过多模态输入（摄像头、传感器）和端云协同，构建具备规划、记忆和环境交互能力的通用智能体。

图：Llama3.1测评

		Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 ^(0.125)	GPT-4o	Claude 3.5 Sonnet
Category	Benchmark											
General	MMLU ^(5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU ^(0-shot, CoT)	73.0	72.3 [△]	60.5	86.0	79.9	69.8	88.6	78.7 ^d	85.4	88.7	88.3
	MMLU-Pro ^(5-shot, CoT)	48.3	—	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval ^(0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus ^(0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K ^(8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 [◇]	94.2	96.1	96.4 [◇]
	MATH ^(0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge ^(0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA ^(0-shot, CoT)	32.8	—	28.8	46.7	33.3	30.8	51.1	—	41.4	53.6	59.4
Tool use	BFCL	76.1	—	60.4	84.8	—	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	—	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	—	—	90.5	—	—	95.2	—	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	—	—	78.2	—	—	83.4	—	72.1	82.5	—
	NIH/Multi-needle	98.8	—	—	97.5	—	—	98.1	—	100.0	100.0	90.8
Multilingual	MGSM ^(0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	—	85.9	90.5	91.6

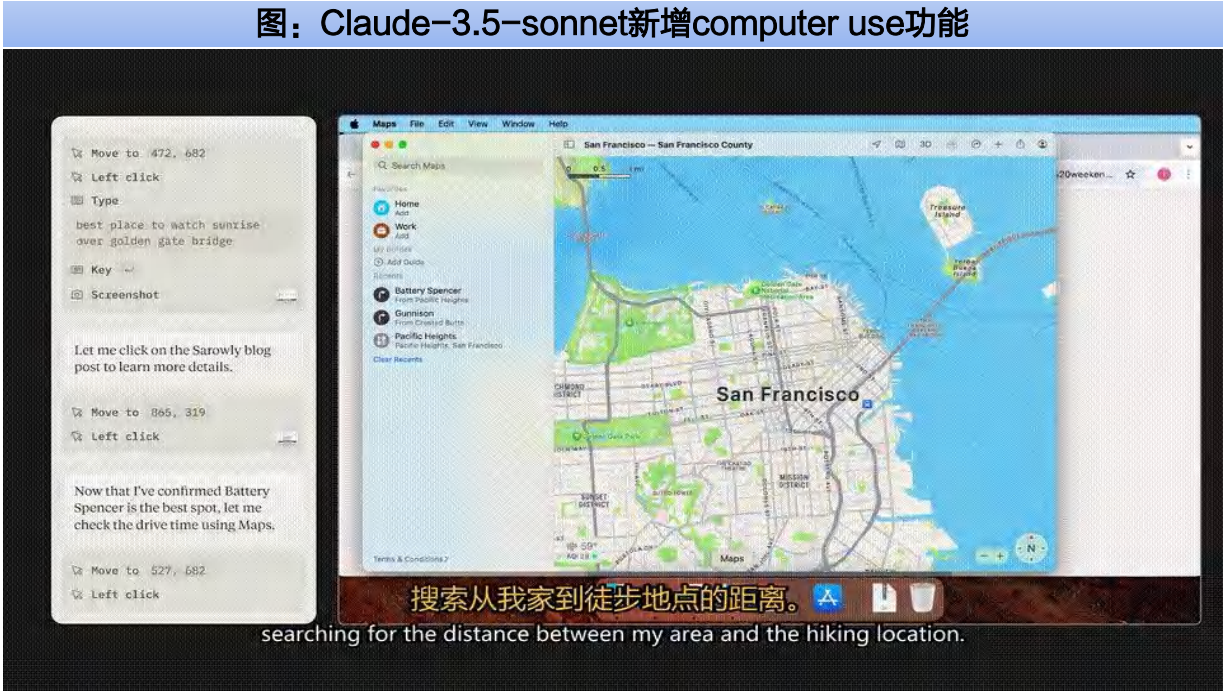
图：Llama3.3测评

Category Benchmark	Llama 3.1 70B	Llama 3.3 70B	Amazon Nova Pro	Llama 3.1 405B (4096)	Gemini Pro 1.5	GPT-4o
General						
MMLU (0-shot, CoT)	86.0	86.0	85.9	88.6	87.1	87.5
MMLU PRO (5-shot, CoT)	66.4	68.9	-	73.3	76.1	73.8
Instruction Following						
IFEval	87.5	92.1	92.1	88.6	81.9	84.6
Code						
Human Eval (0-shot)	80.5	88.4	89.0	89.0	89.0	86.0
MBPP EvalPlus (base) (0-shot)	86.0	87.6	-	88.6	87.8	83.9
Math						
MATH (0-shot, CoT)	68.0	77.0	76.6	73.8	82.9	76.9
Reasoning						
GPQA Diamond (0-shot, CoT)	48.0	50.5	-	49.0	53.5	47.5
Tool use						
BFCL v2 (0-shot)	77.5	77.3	-	81.1	80.3	74.0
Long Context						
NIH/Multi-needle	97.5	97.5	-	98.1	94.7	-
Multilingual						
Multilingual MGSM (0-shot)	86.9	91.1	-	91.6	89.6	80.6

- Anthropic是成立于2021年的人工智能初创企业，由前OpenAI资深成员Dario Amodei带领的七人精英团队共同创办。Anthropic显著区别于同行的一大特点是其对AI安全性的重视，致力于研发可靠、可解释及可操控的AI系统，尤其强调在可解释性上的视角，与OpenAI的发展路径形成鲜明对比。
- 2023年3月，Anthropic推出了Claude大模型；随后产品进一步升级迭代，于2024年3月发布了Claude 3系列。Claude 3系列根据不同的定位，按照功能升序分为：Claude 3 Haiku、Claude 3 Sonnet和 Claude 3 Opus。其中，Haiku是成本最优，市场上速度最快、成本效益最高的模型；Sonnet平衡性能和速度，性价比最高；Opus是最先进的高性能模型，号称当时已经超越GPT-4。
- 2024年10月，Claude3.5 Sonnet迎来升级发布，新增computer use功能。computer use可以让Claude像人一样使用电脑，例如自动完成表格填写、自动完成网页信息搜寻等。

图：新版Claude-3.5-sonnet测评							
	Claude 3.5 Sonnet (new)	Claude 3.5 Haiku	Claude 3.5 Sonnet	GPT-4o*	GPT-4o mini*	Gemini 1.5 Pro	Gemini 1.5 Flash
Graduate level reasoning GPQA (Diamond)	65.0% 0-shot CoT	41.6% 0-shot CoT	59.4% 0-shot CoT	53.6% 0-shot CoT	40.2% 0-shot CoT	59.1% 0-shot CoT	51.0% 0-shot CoT
Undergraduate level knowledge MMLU Pro	78.0% 0-shot CoT	65.0% 0-shot CoT	75.1% 0-shot CoT	—	—	75.8% 0-shot CoT	67.3% 0-shot CoT
Code HumanEval	93.7% 0-shot	88.1% 0-shot	92.0% 0-shot	90.2% 0-shot	87.2% 0-shot	—	—
Math problem-solving MATH	78.3% 0-shot CoT	69.2% 0-shot CoT	71.1% 0-shot CoT	76.6% 0-shot CoT	70.2% 0-shot CoT	86.5% 4-shot CoT	77.9% 4-shot CoT
High school math competition AIME 2024	16.0% 0-shot CoT	5.3% 0-shot CoT	9.6% 0-shot CoT	9.3% 0-shot CoT	—	—	—
Visual Q/A MMMU	70.4% 0-shot CoT	—	68.3% 0-shot CoT	69.1% 0-shot CoT	59.4% 0-shot CoT	65.9% 0-shot CoT	62.3% 0-shot CoT
Agentic coding SWE-bench Verified	49.0%	40.6%	33.4%	—	—	—	—
Agentic tool use TAU-bench	Retail	Retail	Retail	—	—	—	—
	69.2%	51.0%	62.6%	—	—	—	—
	Airline	Airline	Airline	—	—	—	—
	46.0%	22.8%	36.0%				

* Our evaluation tables exclude OpenAI's o1 model family as they depend on extensive pre-response computation time, unlike typical models. This fundamental difference makes performance comparisons difficult.



3.5、Antropic：先于OpenAI推出通用+推理混合模型Claude-3.7-sonnet

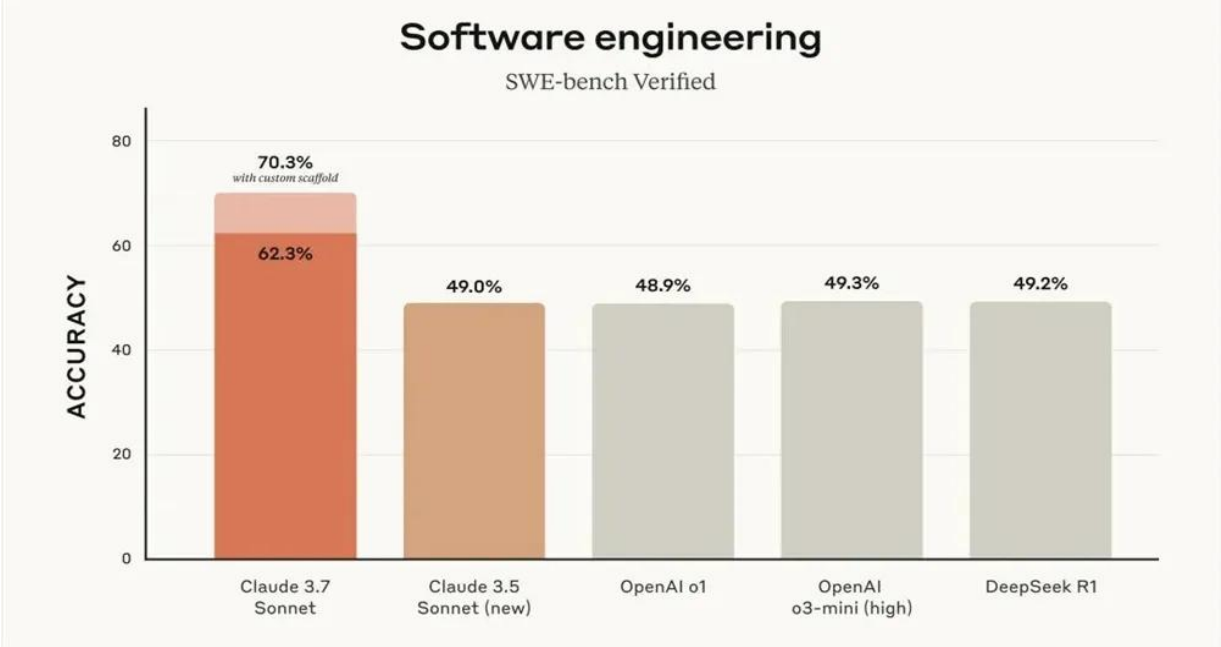
- Anthropic发布市场上首个混合推理模型Claude-3.7-sonnet。2025年2月25日，Anthropic发布了Claude 3.7 Sonnet，Claude 3.7 Sonnet是市场上首个混合推理模型，结合了快速响应和深度思考能力，用户可以通过API精细控制模型的思考时间；Claude 3.7 Sonnet性能评估上，在编码、前端开发、数学、物理等领域表现出色，尤其是真实世界的编码任务中表现卓越，超过DeepSeek R1、OpenAI o1、OpenAI o3-mini（high）模型。
- Claude-3.7-sonnet核心能力在于用户可以自己自行选择标准模式or深度思考模式，在选择深度思考模式下用户可以控制思考的预算时间。在标准模式下，Claude3.7Sonnet是Claude3.5Sonnet的升级版本；在扩展思考模式下，它在回答之前进行自我反思，这提高了在数学、物理、指令遵循、编码和其他许多任务上的性能；通过API使用Claude3.7Sonnet时，用户还可以控制思考的预算，具体表现为可以告诉Claude在回答时最多思考N个tokens，N的最大值为128K tokens的输出限制，使得用户可以在速度（和成本）与回答质量之间进行权衡。此前OpenAI表示，GPT5也将是个混合推理模型，融合GPT4以及GPTo1的能力。

图：Claude-3.7-sonnet多领域评分超OpenAI/DeepSeek

	Claude 3.7 Sonnet 64K extended thinking	Claude 3.7 Sonnet No extended thinking	Claude 3.5 Sonnet (new)	OpenAI o1 ¹	OpenAI o3-mini ¹ High	DeepSeek R1 32K extended thinking	Grok 3 Beta Extended thinking
Graduate-level reasoning GPQA Diamond ³	78.2% / 84.8%	68.0%	65.0%	75.7% / 78.0%	79.7%	71.5%	80.2% / 84.6%
Agentic coding SWE-bench Verified ²	—	62.3% / 70.3%	49.0%	48.9%	49.3%	49.2%	—
Agentic tool use TAU-bench	—	Retail 81.2%	Retail 71.5%	Retail 73.5%	—	—	—
	—	Airline 58.4%	Airline 48.8%	Airline 54.2%	—	—	—
Multilingual Q&A MMMLU	86.1%	83.2%	82.1%	87.7%	79.5%	—	—
Visual reasoning MMMU (validation)	75%	71.8%	70.4%	78.2 %	—	—	76.0% / 78.0%
Instruction-following IFEval	93.2%	90.8%	90.2%	—	—	83.3%	—
Math problem-solving MATH 500	96.2%	82.2%	78.0%	96.4%	97.9%	97.3%	—
High school math competition AIME 2024 ⁴	61.3% / 80.0%	23.3%	16.0%	79.2% / 83.3%	87.3%	79.8%	83.9% / 93.3%

Methodology: We report pass@1 averaged over several trials for most evals to reduce variance, up to an average over 16 trials for AIME and SWE-bench Verified. For our models and several others we additionally report results that benefit from "parallel test time compute" via sampling of multiple chain-of-thought sequences.
1. OpenAI o1 results on TAU-bench Retail and TAU-bench Airline originally reported by OpenAI but later deleted with no alternative sources available. Note: TAU-bench results may not be comparable.
2. SWE-bench Verified: Claude 3.7 Sonnet scores 62.3% out of 500 problems using pass@1 with bash/editor tools plus a "thinking tool" for single-attempt patches—no additional test-time compute used. The 70.3% score uses internal scoring and custom scaffold on a reduced subset of problems. OpenAI results from o3-mini system card cover a different subset of problems with a custom scaffold. DeepSeek R1 results use the "Agentless" framework.
3. GPQA/AIME: Claude 3.7 Sonnet's GPQA and AIME 2024 high scores use internal scoring with parallel test time compute, while o1 and Grok 3's high results use majority voting with N=8k samples.

图：Claude-3.7-sonnet在软件工程领域测评创下高分



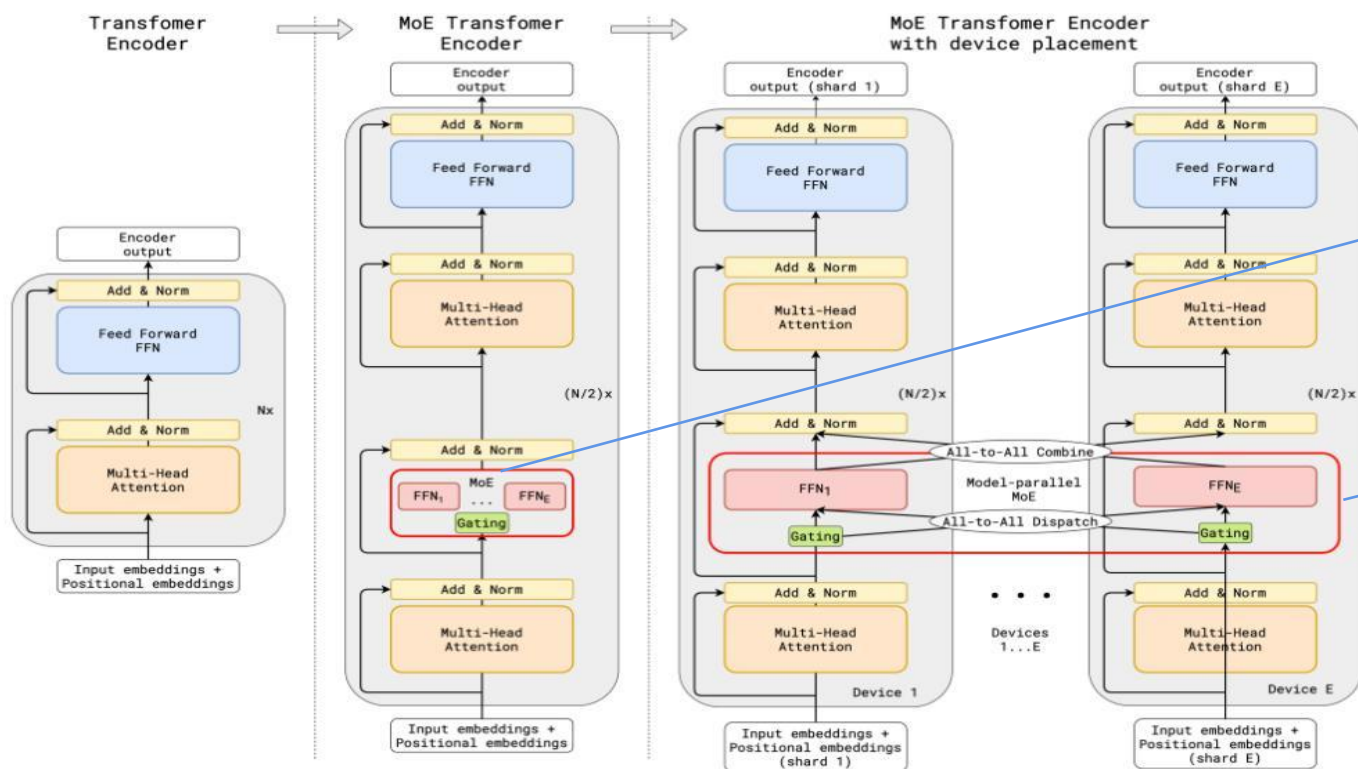
四、模型未来研判：投注后训练+算法的大幅优化

4.1、模型架构的演进：从Dense到MoE，模型大幅降本提效

● **MoE（Mixture of Experts，混合专家模型）**是一种用于提升深度学习模型性能和效率的技术架构。其主要由一组专家模型和一个门控模型组成，核心思想是在处理任务时只激活部分专家模型，并通过门控模型控制专家模型的选择和加权混合。简言之，MoE在训练过程通过门控模型实现“因材施教”，进而在推理过程实现专家模型之间的“博采众长”。

● 从Transformer架构上看，MoE使用稀疏的MoE层代替稠密的前馈网络（FFN）层，专家可以是FFN，也可以是更复杂的网络，甚至是MoE本身，这样就会形成有多层MoE的MoE；而门控网络或者路由来决定将哪个 token 发送给哪个专家。

图：MoE架构下，MoE层取代FFN层

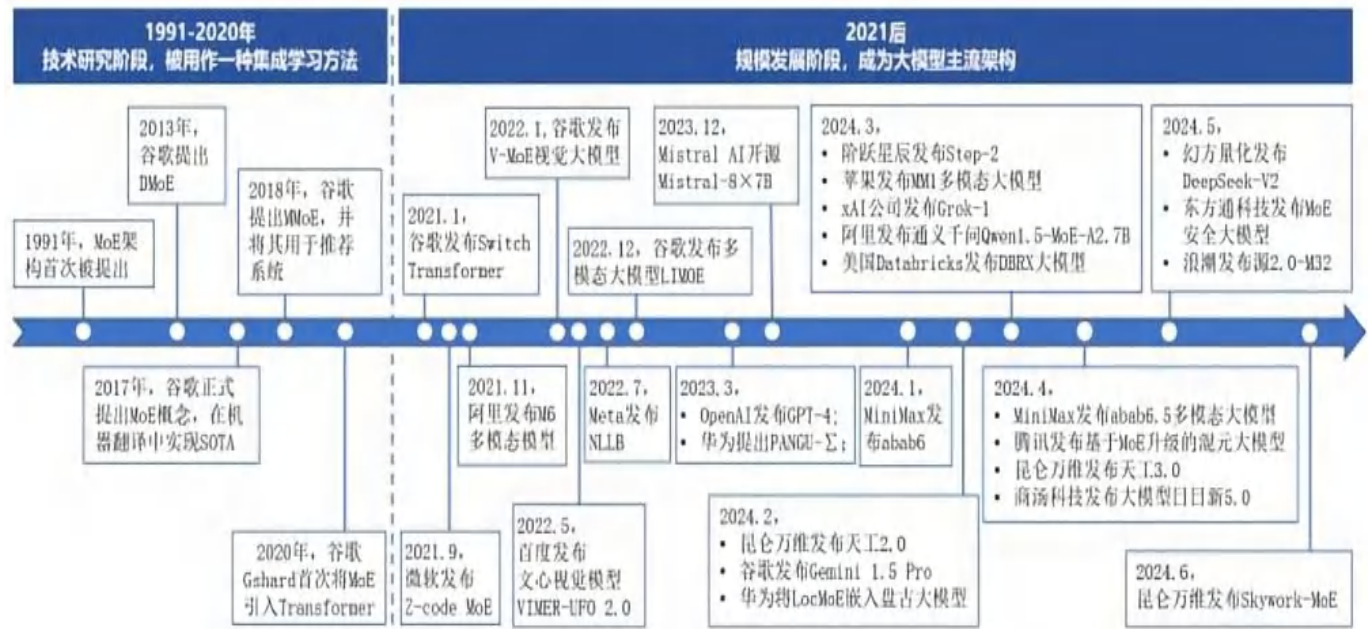


Transformer架构中，MoE层取代了传统的FFN层。

MoE架构还可以将不同专家扩展到多个设备上，MoE层在不同设备之间共享，而其他层（例如注意力层）被复制。

- MoE与Transformer融合，逐步成为主流架构。MoE 起源于 1991 年的论文《Adaptive Mixture of Local Experts》。2020年，谷歌Gshard首次将MoE引入Transformer构建分布式并行计算架构，打开MoE发展新思路。之后，MoE逐渐进入规模发展阶段，作为一种底层架构优化方法，与Transformer结合，陆续被用于推荐系统、自然语言处理、计算机视觉、多模态大模型等领域。
- 国内外主流企业差异化推进MoE大模型布局和落地，2024年全球MoE大模型数量呈爆发增长态势。据53AI网统计，2024年1-5月全球发布MoE大模型数量约20个，超2021-2023三年总量（约10个），且以多模态大模型为主（占比约90%）。谷歌、OpenAI、阿里、华为、腾讯等大型企业侧重利用MoE提升大模型性能和实用性。而Mistral AI、昆仑万维、MiniMax、幻方量化等初创企业侧重利用MoE低成本优势抢占AI市场。

图：MoE架构发展历程



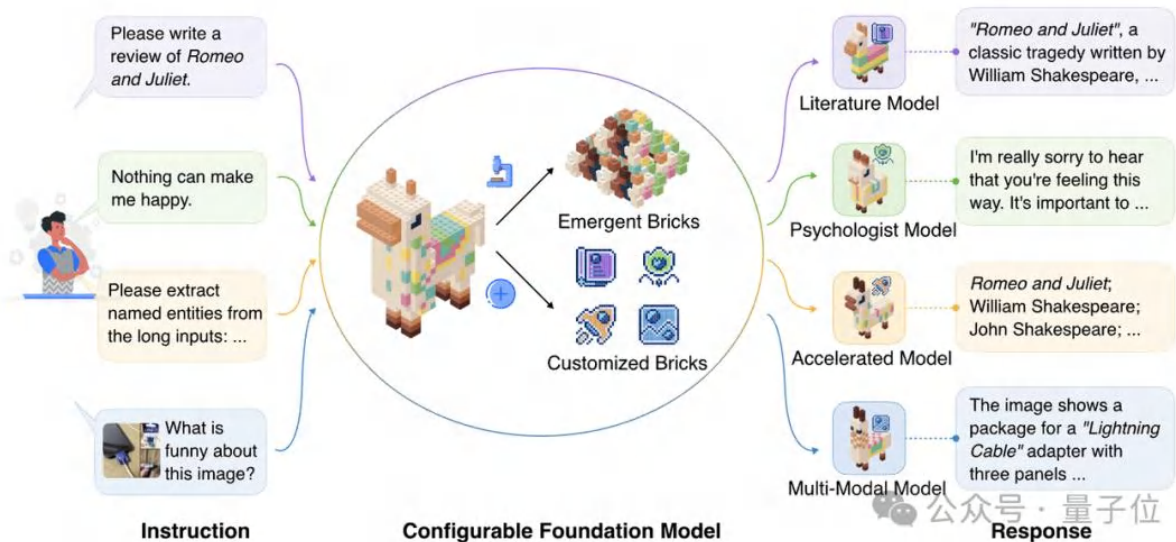
公司	布局特点	已发布模型	发布时间	参数规模
谷歌	从算法、工具、模型、硬件、生态全面布局，力争MoE发展领航者	Switch Transformer	2021.1	1.6万亿
		GLaM	2021.12	1.2万亿
		V-MoE	2022.1	150亿
		LIMO	2022.12	56亿
		Gemini 1.5 Pro	2024.2	
OpenAI	持续推出国际领先大模型，并利用MoE控制成本，GPT-4实现“性能/成本”提升70倍	GPT4	2023.3	1.76万亿
阿里	利用MoE升级万亿及以上规模参数模型，并开源“小模型”，追赶国际先进模型性能	M6	2021.11	10万亿
		Qwen1.5-MoE-A2.7B	2024.3	27亿
华为	侧重MoE架构创新，提升大模型通用性和可移植性，如提出全新LocMoE架构	盘古大模型	2024.2	
腾讯	通过MoE架构优化，结合模型量化方法，提升混元大模型性能，并赋能现有业务	混元大模型	2024.3	
Mistral AI	把握全球首个MoE大模型开源先发优势，迅速提升品牌影响力和市场地位	Mistral-8 × 7B	2023.12	560亿
		Mistral-8x22B	2024.4	1760亿
昆仑万维	把握国内先发优势，重视MoE大模型训练/推理相关技术研发及海外业务扩展	天工2.0	2024.2	千亿
		天工3.0	2024.4	4000亿
幻方量化	采用MLA(多头注意力)机制并自研MoE架构，实现成本大幅降低，应对大模型“价格战”	DeepSeek-MoE	2024.1	20/160/1450亿
		DeepSeek-V2	2024.5	2360亿

4.1、模型架构的演进：从Dense到MoE，模型大幅降本提效

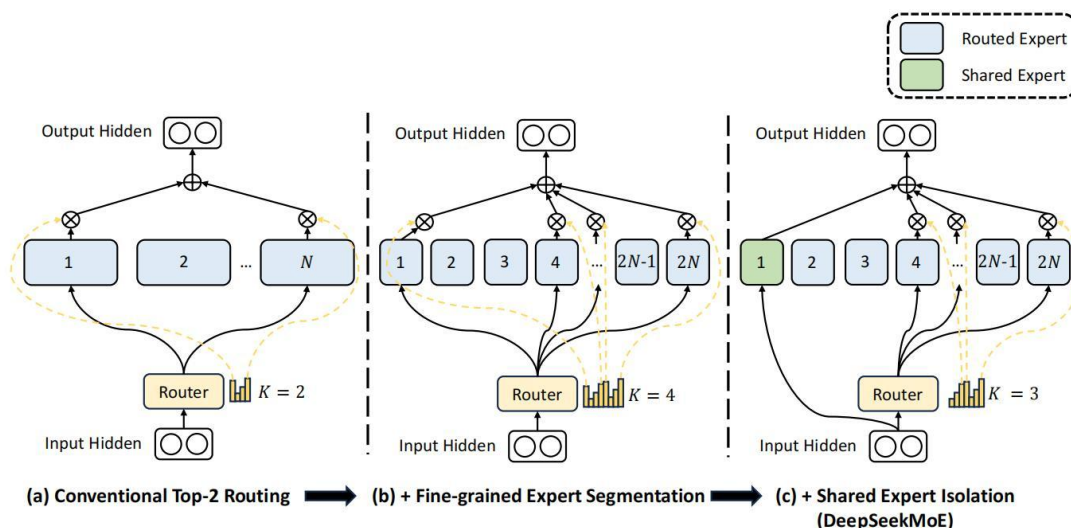
✓ **DeepSeek提出DeepSeekMoE，在传统MoE架构之上继续降本提效。包括：**1) **细粒度专家分割**：在保持模型参数和计算成本一致的情况下，用更精细的颗粒度对专家进行划分，更精细的专家分割使得激活的专家能够以更灵活和适应性更强的方式进行组合；2) **共享专家隔离**：采用传统路由策略时，分配给不同专家的token可能需要一些共同的知识或信息，因此多个专家可能会有参数冗余。专门的共享专家致力于捕获和整合不同上下文中的共同知识，有助于构建一个具有更多专业专家且参数更高效的模型。

✓ **负载均衡**：MoE架构下容易产生每次都由少数几个专家处理所有tokens的情况，而其余大量专家处于闲置状态，此外，若不同专家分布在不同计算设备上，同样会造成计算资源浪费以及模型能力局限；负载均衡则类似一个公平的“裁判”，鼓励专家的选择趋于均衡，避免出现上述专家激活不均衡的现象。DeepSeek在专家级的负载均衡外，提出了设备级的负载均衡，确保了跨设备的负载均衡，大幅提升计算效率，缓解计算瓶颈。

图：MoE架构理解框架



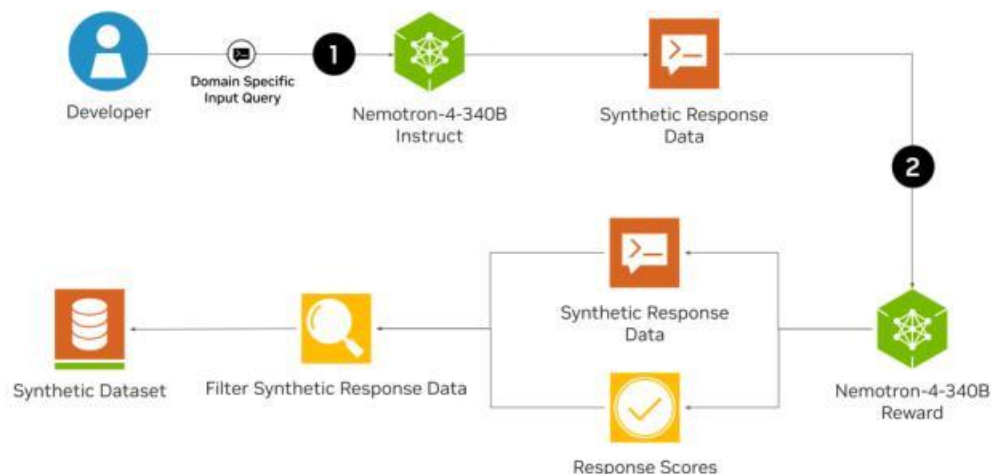
图：DeepSeekMoE对比传统MoE架构



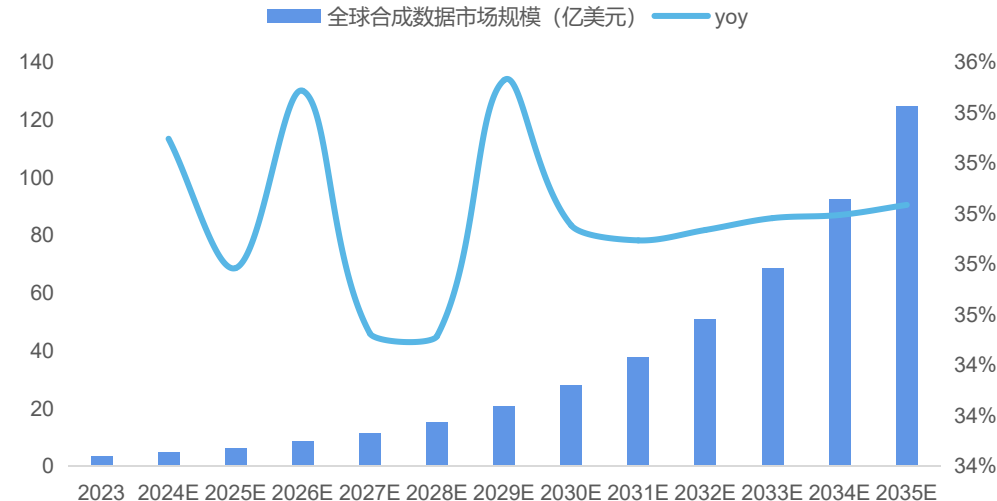
4.2、合成数据作为AI时代新石油，支撑模型继续在pre training上scaling

- 合成数据(Synthetic Data)通常由算法和数学模型创建。首先通过建模真实数据的分布，然后在该分布上进行采样，创建出新数据集，模拟真实数据中的统计模式和关系。
- 目前市面上已有许多工具可生成合成数据，如英伟达发布3D仿真数据生成引擎Omniverse Replicator、微软开源合成数据工具Synthetic Data Showcase等。6月14日，英伟达发布开源大模型Nemotron-4 340B，包含基础模型Base、指令模型Instruct和奖励模型Reward，也可用于生成高质量合成数据，其中Instruct模型用于生成基于文本的合成输出，Reward模型对生成的文本进行评估并提供反馈，指导迭代改进并确保合成数据的准确性。
- 合成数据作为数字经济时代的“新型石油”，在高质量数据或耗尽的背景下继续支撑模型scaling。据Gartner预测，2024年AI训练中用到的数据有60%是合成数据，到2030年绝大部分训练数据将是合成数据。据著名市场调研机构Nester预测，全球合成数据的市场呈现蓬勃发展趋势，年复合增长率达35%，预计到2035年底，合成数据市场规模将达124.5亿美元。

图：Nemotron-4模型生成合成数据的流程



图：全球合成数据市场规模稳步提升

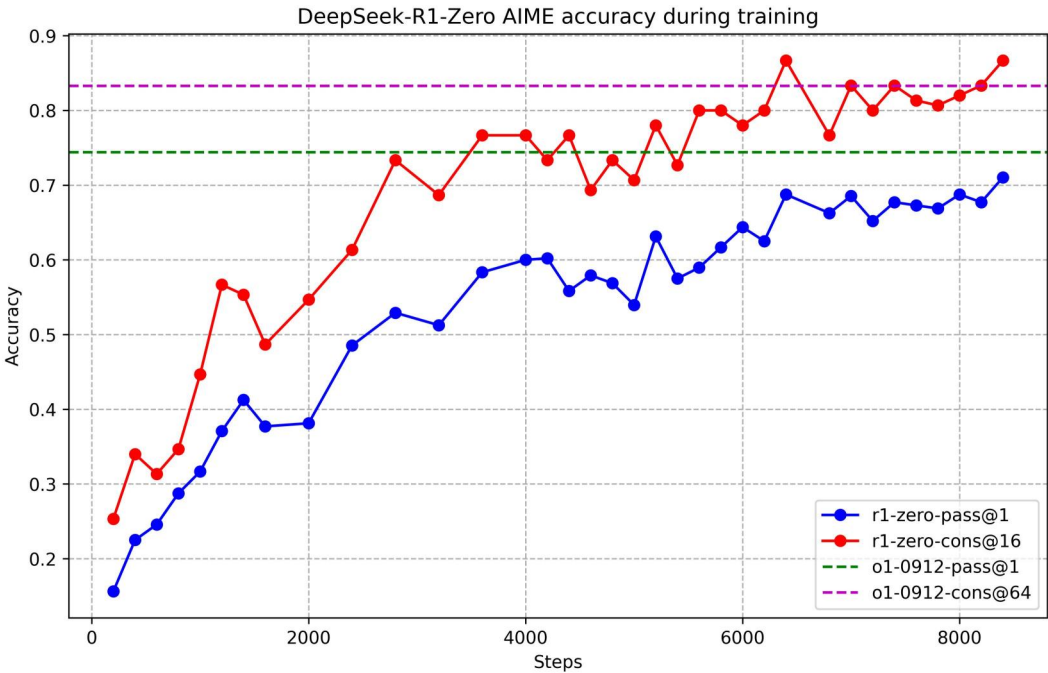


- DeepSeek探索LLM在没有任何监督数据的情况下发力推理能力的潜力，通过纯RL（强化学习）的过程实现自我进化。具体来说，DS使用DeepSeek-V3-Base 作为基础模型，并使用GRPO（群体相对策略优化）作为RL框架来提高模型在推理中的性能。在训练过程中，DeepSeek-R1-Zero自然而然地出现了许多强大而有趣的推理行为。
- 经过数千次 RL 步骤后，DeepSeek-R1-Zero 在推理基准测试中表现出卓越的性能。例如，AIME 2024 的 pass@1 分数从15.6%增加到 71.0%，在多数投票的情况下，分数进一步提高到86.7%，与OpenAI-o1-0912的性能相当。

图：R1-Zero在不同测试基准下超过o1mini甚至比肩o1的水平

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

图：随时间推移DS模型性能显著提升



- GRPO相对PPO节省了与策略模型规模相当的价值模型，大幅缩减模型训练成本。
- 传统强化学习更多使用PPO（近端策略优化），PPO中有3个模型，分别是参考模型（reference model）、奖励模型（reward model）、价值模型（value model），参考模型作为稳定参照，与策略模型的输出作对比；奖励模型根据策略模型的输出效果给出量化的奖励值，价值模型则根据对策略模型的每个输出预测未来能获得的累计奖励期望。ppo中的价值模型规模与策略模型相当，由此带来巨大的内存和计算负担。
- GRPO（群里相对策略优化）中省略了价值模型，采用基于组的奖励归一化策略，简言之就是策略模型根据输入q得到输出o（1，2，3），再计算各自的奖励值r（1，2，3），而后不经过价值模型，而是制定一组规则，评判组间价值奖励值的相对关系，进而让策略模型以更好的方式输出。

图：GRPO相对传统PPO强化学习方式对比

图：GRPO核心方法详解

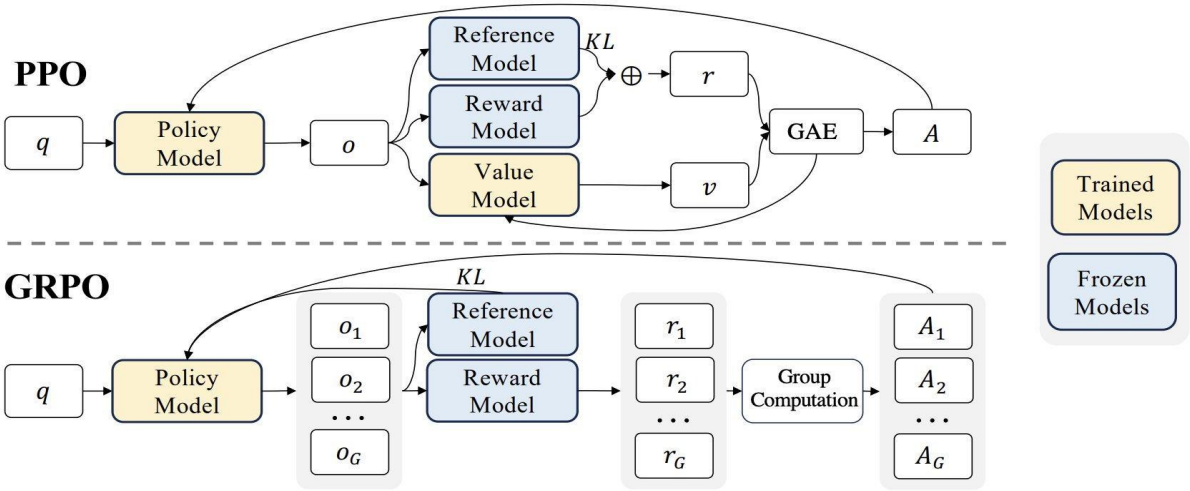


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

Group Relative Policy Optimization In order to save the training costs of RL, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q , GRPO samples a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_{θ} by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

where ϵ and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \dots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$

- DeepSeek引领大模型开源且低成本部署潮流，模型普世化趋势逐步明确。以DeepSeek为例，基于DeepSeekMoE架构，每次推理的时候仅激活37B参数；同时通过MLA等低秩分解的方式实现显存占用的大幅降低，推动计算资源以及内存消耗。DeepSeek R1/V3模型保持开源的同时，并在2月最后一周开源五大核心代码库，加速行业模型降本增效趋势。
- 本地化部署DeepSeek-R1-32B及以下模型仅需要消费级显卡。从本地化部署DeepSeek所需硬件需求上看，本地化部署满血版的DeepSeek R1需要2台A100服务器（单台8卡）；而部署32/70B的蒸馏版模型，只需要4090显卡，而对于7/8/14B等小参数模型，只需要3070/3080等基础消费级显卡即可。

表：DeepSeek本地化部署所需的硬件要求

模型版本	参数(B)	模型大小	VRAM要求(GB)	参考GPU配置
DeepSeek-R1	671B	>500G	~1,342	多GPU配置(如：NVIDIA A100 80GB × 16)
DeepSeek-R1-Distill-Llama-70B	70B	43G	~32.7	NVIDIA RTX 4090 24GB × 2
DeepSeek-R1-Distill-Qwen-32B	32B	20G	~14.9	NVIDIARTX4090 24GB
DeepSeek-R1-Distill-Qwen-14B	14B	9G	~6.5	NVIDIA RTX 3080 10GB
DeepSeek-R1-Distill-Llama-8B	8B	4.9G	~3.7	NVIDIA RTX 3070 8GB
DeepSeek-R1-Distill-Qwen-7B	7B	4.7G	~3.3	NVIDIA RTX 3070 8GB
DeepSeek-R1-DistillQwen-1.5B	1.5B	1.1G	~0.7	NVIDIA RTX 3060 12GB

五、投资建议与风险提示

◆ 大模型技术稳步提升，推动AGI时代加速到来，以大模型为底座的技术迭代或将持续驱动国产AI估值迎来重塑，维持计算机行业“推荐”评级。

◆ 相关公司

1) 算力：①云计算：中国电信、中国移动、中国联通、金山云、优刻得、青云科技、深信服；②IDC：云赛智联、光环新网、奥飞数据、数据港、润泽科技、科华数据、大位科技、ST鹏博、ST华通、ST证通；③芯片：海光信息、寒武纪；④服务器/一体机：中科曙光、浪潮信息、华勤技术、云从科技、恒为科技、中国软件国际、神州数码、烽火通信、朗科科技；⑤交换机：星网锐捷、紫光股份、中兴通讯；⑥液冷：飞荣达、英维克、申菱环境、高澜股份；⑦电源：欧陆通、麦格米特、中国长城；⑧柴油发电机：科泰电源、泰豪科技、潍柴重机、苏美达、动力新科、玉柴国际；⑨边缘计算：网宿科技、顺网科技、中科创达、云天励飞。

2) AI应用：①2G：中国软件、太极股份、深桑达、电科数字、广电运通、数字政通、中科星图、新点软件、国投智能、云从科技、税友股份、航天信息、拓尔思、能科科技、博思软件、华宇软件、通达海、金桥信息；②2B：金蝶国际、用友网络、泛微网络、致远互联、卫宁健康、创业慧康、广联达、石基信息、明源云、新致软件、汉得信息、鼎捷软件、赛意信息、莱斯信息、四川九州、东方财富、同花顺、恒生电子、新开普、佳发教育、拓维信息、远光软件、润和软件、索辰科技、中望软件、百融云、托普云农、焦点科技、盛视科技；③2C：金山办公、三六零、万兴科技、福昕软件、合合信息。

3) IT服务：宇信科技、京北方、中科软、软通动力、中科创达、新炬网络、天玑科技。

- 1) 大模型产业发展不及预期：AI行业核心是技术驱动，如果人工智能大语言模型技术进步不及预期，或应用效果不及预期，将导致AI产业发展逻辑受到影响。
- 2) 中美博弈加剧：国际形势持续不明朗，美国不断通过“实体清单”等方式对中国企业实施打压，若中美紧张形势进一步升级，将可能导致中国半导体供应链或AI技术创新受到影响。
- 3) 宏观经济影响下游需求：宏观经济环境下行，将影响客户对信息化基础设施的采购需求。
- 4) 市场竞争加剧：软件、技术和硬件是成熟且完全竞争的行业，新进入者可能加剧整个行业的竞争态势。
- 5) 相关公司业绩不及预期：市场环境变化、公司治理情况变化、其他非主营业务经营不及预期等原因或将导致相关公司的整体业绩不及预期。
- 6) 国内外公司并不具备完全可比性，对标的相关资料和数据仅供参考。
- 7) 对标的相关资料和数据仅供参考。
- 8) 数据安全风险:数据安全监管超预期，可能影响产业正常推进。

计算机小组介绍

刘熹，计算机行业首席分析师，上海交通大学硕士，多年计算机行业研究经验，善于把握由技术、政策驱动的科技产业新趋势，致力于进行前瞻重磅推荐。2024年金牌分析师，2024年同花顺最受欢迎分析师。

分析师承诺

刘熹，本报告中的分析师均具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立，客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收取到任何形式的补偿。

国海证券投资评级标准

行业投资评级

推荐：行业基本面向好，行业指数领先沪深300指数；
中性：行业基本面稳定，行业指数跟随沪深300指数；
回避：行业基本面向淡，行业指数落后沪深300指数。

股票投资评级

买入：相对沪深300 指数涨幅20%以上；
增持：相对沪深300 指数涨幅介于10%~20%之间；
中性：相对沪深300 指数涨幅介于-10%~10%之间；
卖出：相对沪深300 指数跌幅10%以上。

免责声明

本报告的风险等级定级为R3，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，也不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。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、价值及投资收入可能会波动。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。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

风险提示

市场有风险，投资需谨慎。投资者不应将本报告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，投资者务必向本公司或其他专业人士咨询并谨慎决策。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。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。

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。本报告不构成本公司向该机构之客户提供的投资建议。

任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。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。

郑重声明

本报告版权归国海证券所有。未经本公司的明确书面特别授权或协议约定，除法律规定的情况外，任何人不得对本报告的任何内容进行发布、复制、编辑、改编、转载、播放、展示或以其他方式非法使用本报告的部分或者全部内容，否则均构成对本公司版权的侵害，本公司有权依法追究其法律责任。

国海证券 · 研究所 · 计算机研究团队

心怀家国，洞悉四海



国海研究上海

上海市黄浦区绿地外滩中心C1栋
国海证券大厦

邮编：200023

电话：021-61981300

国海研究深圳

深圳市福田区竹子林四路光大银行大厦28F

邮编：518041

电话：0755-83706353

国海研究北京

北京市海淀区西直门外大街168号腾达大厦25F

邮编：100044

电话：010-88576597