



Deepseek对中国算力产业的影响

2025年2月

目录

CONTENT

1

DeepSeek的技术突破与市场定位

2

DeepSeek驱动算力需求变革

3

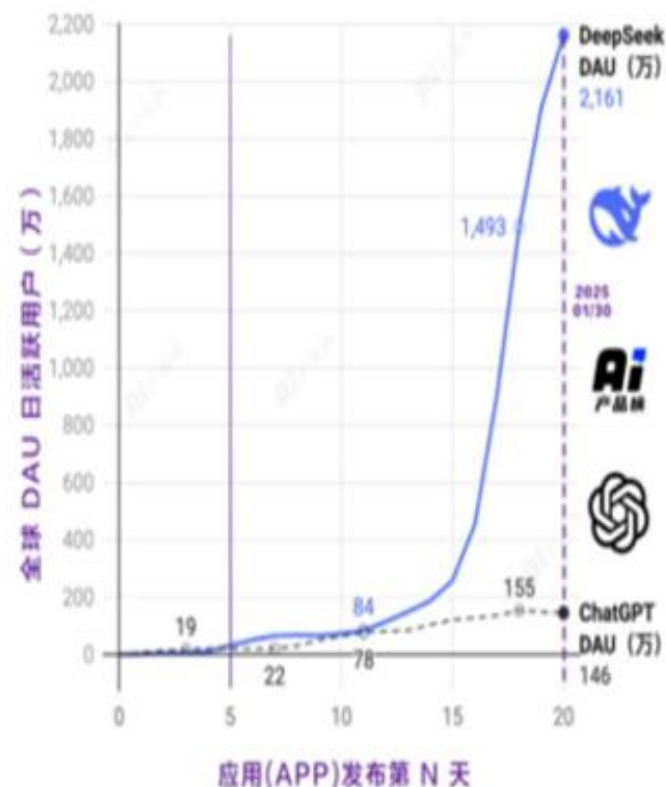
算力产业链的重构

DeepSeek爆火--C端： Deepseek全球破圈，成为用户规模增长最快的移动AI应用

超级app增长1亿用户所用时间



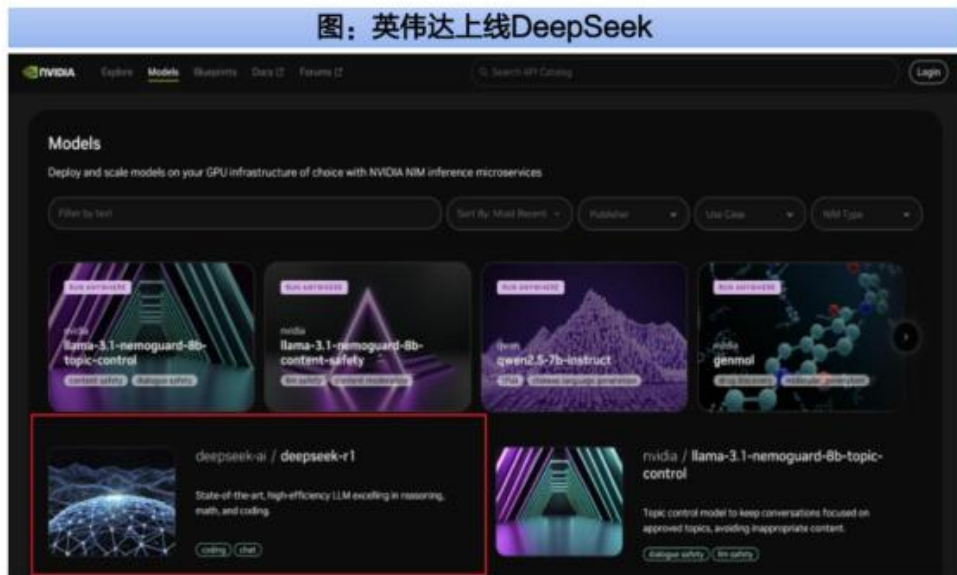
App上线后同样天数DeepSeek与ChatGPT移动端全球DAU对比情况



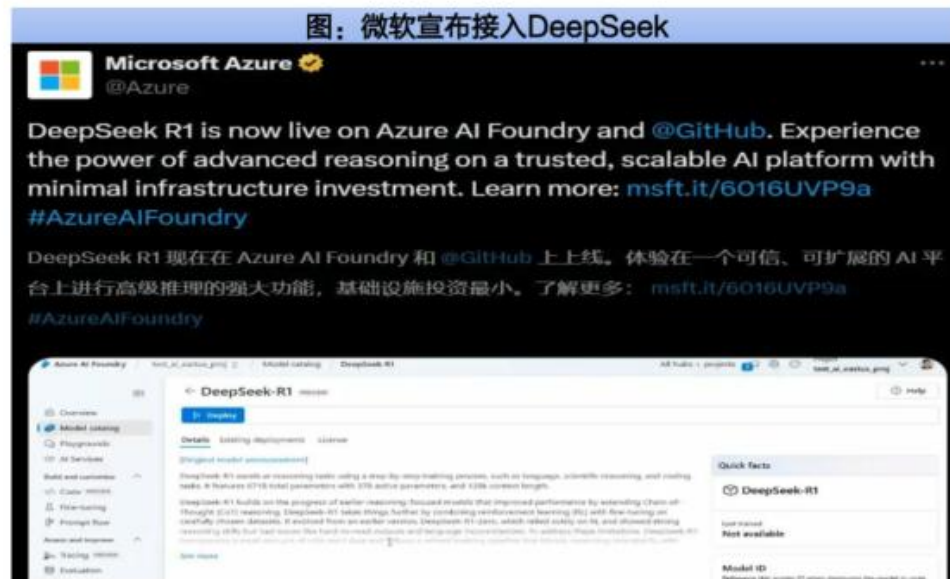
DeepSeek爆火--B端：科技巨头积极拥抱DeepSeek

- ✓ 微软、英伟达、亚马逊、英特尔、AMD等科技巨头陆续上线DeepSeek模型服务。
- ✓ 1) 1月30日，英伟达宣布DeepSeek-R1可作为 NVIDIA NIM 微服务预览版使用。
- ✓ 2) 1月，DeepSeek-R1 模型被纳入微软平台 Azure AI Foundry 和 GitHub 的模型目录，开发者将可以在Copilot +PC上本地运行DeepSeek-R1 精简模型，以及在Windows上的 GPU 生态系统中运行，此外还宣布将 DeepSeek-R1部署在云服务Azure上。
- ✓ 3) AWS（亚马逊云科技）宣布，用户可以在Amazon Bedrock 和Amazon SageMaker AI两大AI服务平台上部署DeepSeek-R1模型。
- ✓ 4) Perplexity 宣布接入了 DeepSeek 模型，将其与 OpenAI 的 GPT-o1 和 Anthropic 的 Claude-3.5 并列作为高性能选项。
- ✓ 5) 华为：已上线基于其云服务的DeepSeek-R1相关服务；
- ✓ 6) 腾讯：DeepSeek-R1大模型可一键部署至腾讯云‘HAI’上，开发者仅需3分钟就能接入调用。
- ✓ 7) 百度：DeepSeek-R1和DeepSeek-V3模型已在百度智能云千帆平台上架；
- ✓ 8) 阿里：阿里云PAI Model Gallery支持云上一键部署DeepSeek-R1和DeepSeek-V3模型。

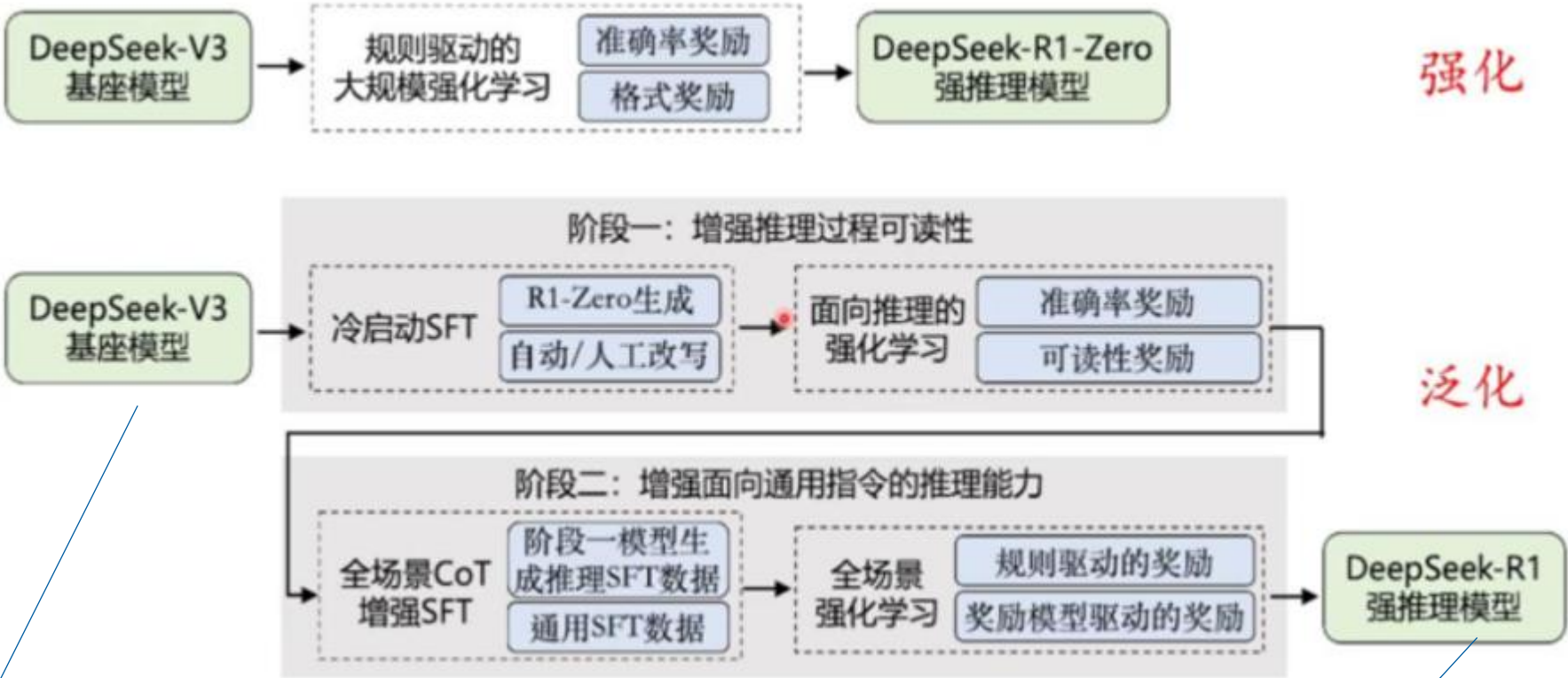
图：英伟达上线DeepSeek



图：微软宣布接入DeepSeek



DeepSeek明星产品：DeepSeek的LLM模型分为三个版本：基座模型V3、强化推理版R1-Zero、泛化推理版R1



我是 DeepSeek，很高兴见到你！

我可以帮你写代码、读文件、写作各种创意内容，请把你的任务交给我吧~

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索



我是 DeepSeek，很高兴见到你！

我可以帮你写代码、读文件、写作各种创意内容，请把你的任务交给我吧~

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索



DeepSeek爆火的原因：一流的性能表现、大幅降低的算力成本、开源模式

高性能模型架构创新

DeepSeek的模型架构创新，如MoE和FP8混合精度训练，大幅提升模型性能和训练效率。

低成本实现高性能

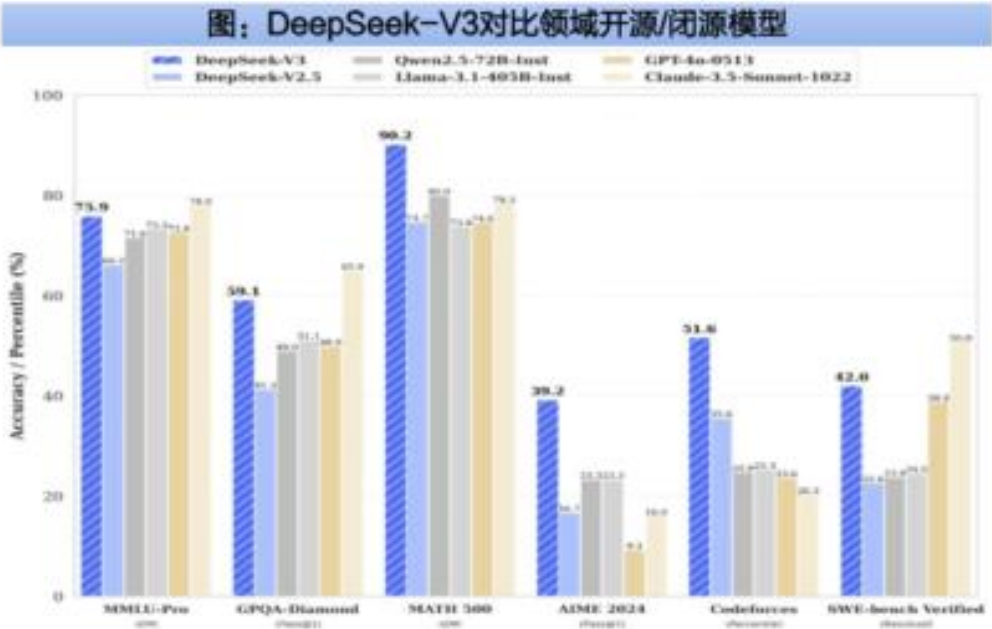
通过算法创新和硬件优化，DeepSeek以低成本实现高性能，改变AI领域的竞争规则。

开源策略推动技术普及

DeepSeek采用开源策略，降低AI技术门槛，促进全球开发者参与，推动技术快速普及和迭代。

一流的性能表现：DeepSeek-V3性能对齐海外领军闭源模型

- DeepSeek-V3 为自研 MoE 模型，671B 参数，激活 37B，在 14.8Ttoken上进行了预训练。V3多项评测成绩超越了 Qwen2.5-72B 和 Llama-3.1-405B 等其他开源模型，并在性能上和世界顶尖的闭源模型 GPT-4o 以及 Claude-3.5-Sonnet 不分伯仲。
- 在具体的测试集上，DeepSeek-V3在知识类任务上接近当前表现最好的模型 Claude-3.5-Sonnet-1022；长文本/代码/数学/中文能力上均处于世界一流模型位置。



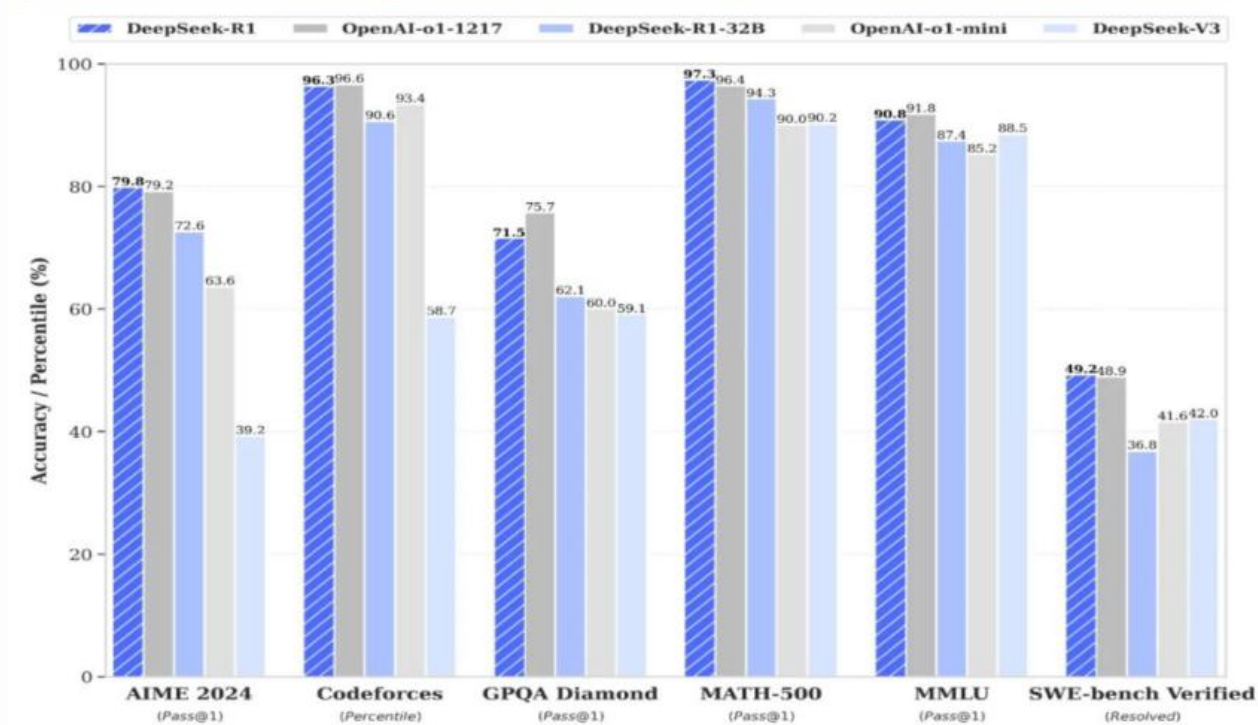
图：DeepSeek-V3在英文、代码、数学领域表现优异

测试集		DeepSeek-V3	Qwen2.5-72B-Inst.	Llama3.1-405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o-0513
	模型架构	MoE	Dense	Dense	-	-
	# 激活参数	37B	72B	405B	-	-
	# 总参数	671B	72B	405B	-	-
英文	MMLU (EM)	88.5	85.3	88.6	88.3	87.2
	MMLU-Redux (EM)	89.1	85.6	86.2	88.9	88
	MMLU-Pro (EM)	75.9	71.6	73.3	78	72.6
	DROP (3-shot F1)	91.6	76.7	88.7	88.3	83.7
	IF-Eval (Prompt Strict)	86.1	84.1	86	86.5	84.3
	GPQA-Diamond (Pass@1)	59.1	49	51.1	65	49.9
	SimpleQA (Correct)	24.9	9.1	17.1	28.4	38.2
	FRAMES (Acc.)	73.3	69.8	70	72.5	80.5
	LongBench v2 (Acc.)	48.7	39.4	36.1	41	48.1
代码	HumanEval-Mul (Pass@1)	82.6	77.3	77.2	81.7	80.5
	LiveCodeBench(Pass@1-COT)	40.5	31.1	28.4	36.3	33.4
	LiveCodeBench (Pass@1)	37.6	28.7	30.1	32.8	34.2
	Codeforces (Percentile)	51.6	24.8	25.3	20.3	23.6
	SWE Verified (Resolved)	42	23.8	24.5	50.8	38.8
	Aider-Edit (Acc.)	79.7	65.4	63.9	84.2	72.9
	Aider-Polyglot (Acc.)	49.6	7.6	5.8	45.3	16
数学	AIME 2024 (Pass@1)	39.2	23.3	23.3	16	9.3
	MATH-500 (EM)	90.2	80	73.8	78.3	74.6
	CNMO 2024 (Pass@1)	43.2	15.9	6.8	13.1	10.8
中文	CLUEWSC (EM)	90.9	91.4	84.7	85.4	87.9
	C-Eval (EM)	86.5	86.1	61.5	76.7	76
	C-SimpleQA (Correct)	64.1	48.4	50.4	51.3	59.3

一流的性能表现：DeepSeek-R1性能对标OpenAI-o1正式版

- 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。
- R1 在 2024 年 AIME 测试中取得了 79.8% 的成绩，与 OpenAI o1 的 79.2% 水平相当。在 MATH-500 基准测试中，DeepSeek-R1 以 97.3% 的成绩略微超越了 o1 的 96.4%。在编程领域，该模型在 Codeforces 平台上表现优异。

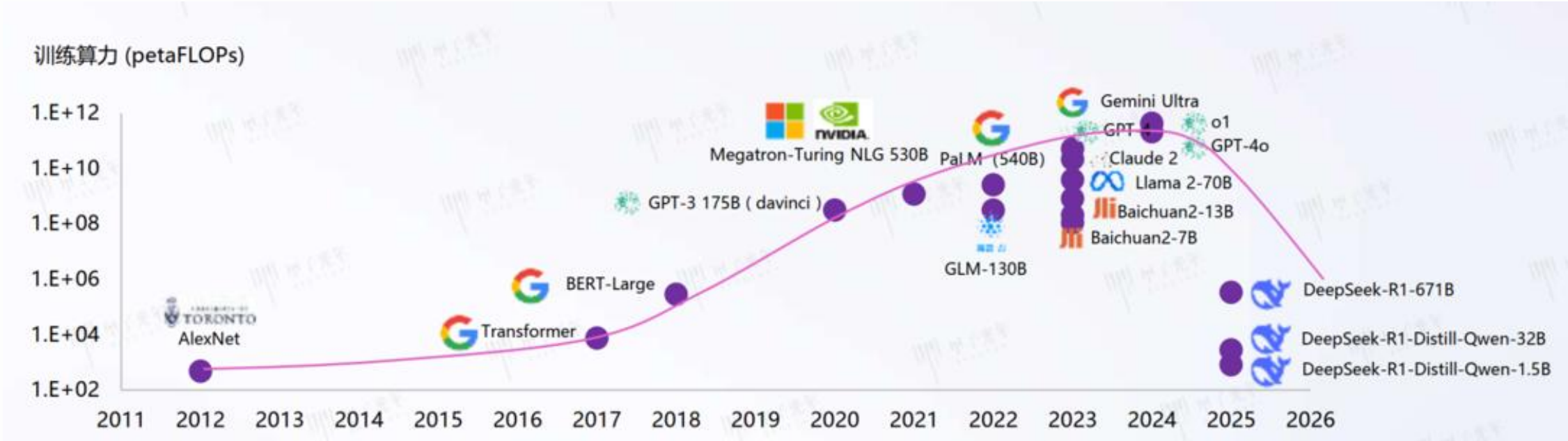
图：DeepSeek-R1性能比肩 OpenAI o1 正式版



图：DeepSeek R1与其他模型的性能对比

Benchmark (Metric)	Claude-3.5- Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
Architecture	-	-	MoE	-	-	MoE
# Activated Params	-	-	37B	-	-	37B
# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2
	IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	75.7
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	47.0
	FRAMES (Acc)	72.5	80.5	73.3	76.9	-
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6
	Codeforces (Rating)	717	759	1134	1820	2061
	SWE Verified (Resolved)	50.8	38.8	42.0	41.6	48.9
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	61.7
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-
	C-Eval (EM)	76.7	76.0	86.5	68.9	-
	C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-

大幅降低的算力成本：训练算力下降90%



DeepSeek-V3和R1模型不仅性能出色，训练成本也极低。V3模型仅用2048块H800 GPU训练2个月，消耗278.8万GPU小时。相比之下，Llama3-405B消耗了3080万GPU小时，是V3的11倍。按H800 GPU每小时2美金计算，V3的训练成本仅为**557.6万美金**，而同等性能的模式通常需要**0.6-1亿美金**。R1模型在V3基础上，通过引入大规模强化学习和多阶段训练，进一步提升了推理能力，成本可能更低。

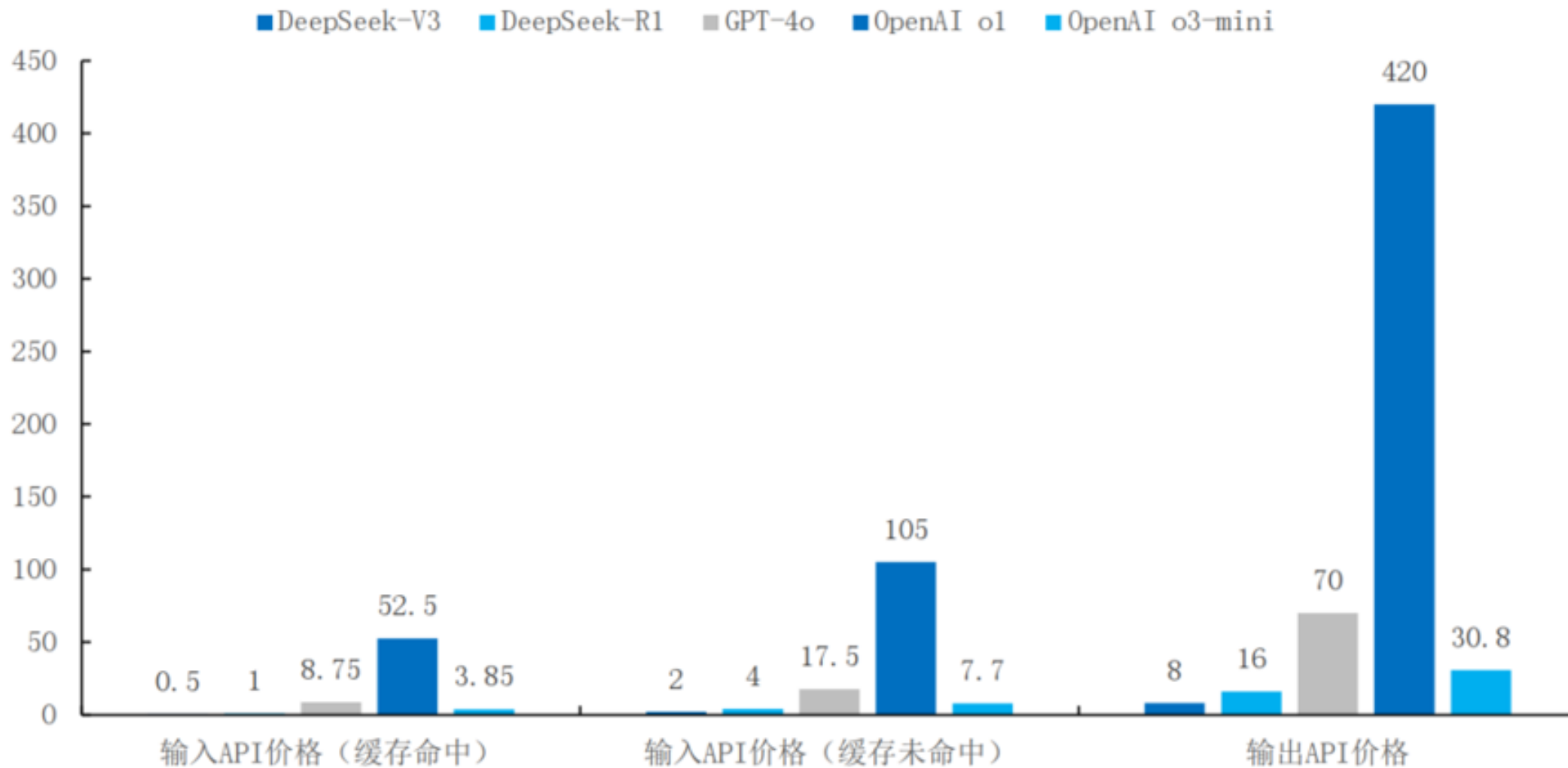
图 7：DeepSeek-V3 模型训练仅需要 278.8 万 GPU 小时训练资源

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	\$5.328M	\$0.238M	\$0.01M	\$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is \$2 per GPU hour.

API定价下降89% (V3) 、 96% (R1)

大模型API定价 (元/M Tokens)



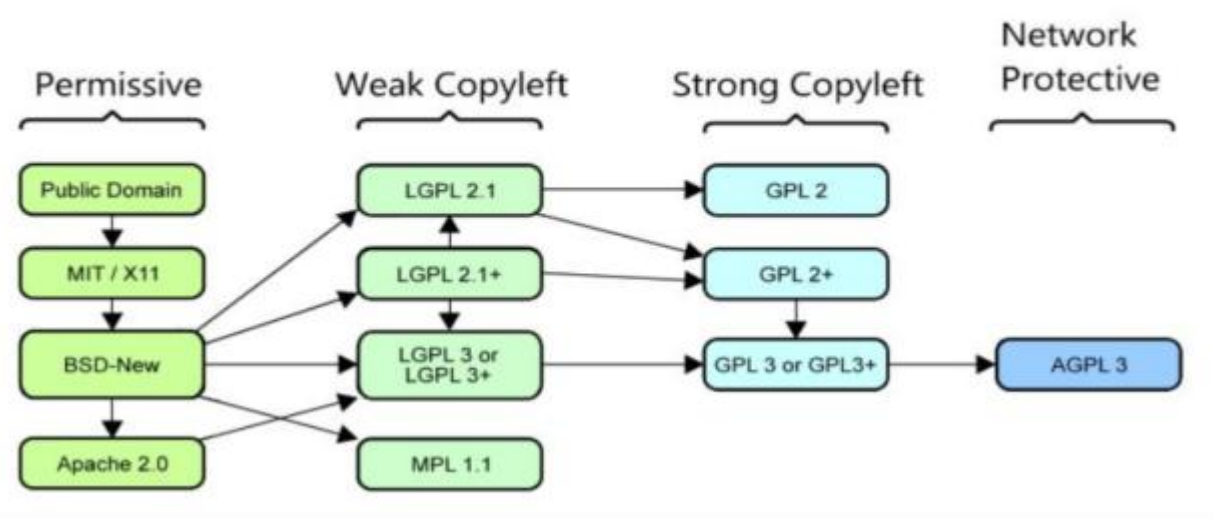
开源大模型：打破OpenAI等闭源模型生态

DeepSeek V3与R1模型实现了开源，采用MIT协议。这产生多方面影响：

- 对大模型发展**：这提升了世界对中国AI大模型能力的认知，一定程度打破了OpenAI与Anthropic等高级闭源模型的封闭生态。DeepSeek R1在多个测试指标中对标OpenAI o1，通过模型开源，也将大模型平均水平提升至类OpenAI o1等级。
- 对下游生态**：优质的开源模型可更好用于垂类场景，即使用户针对自身需求蒸馏，或使用自有数据训练，从而适合具体下游场景；此外，模型训推成本降低，将带来使用场景的普及，带动AIGC、端侧等供给和需求。

DeepSeek 不仅开源了 R1-Zero 和 R1 两个 671B 模型，还通过 DeepSeek-R1 的输出，蒸馏了 6 个小模型开源给社区，其中 32B 和 70B 模型在多项能力上实现了对标 OpenAI o1-mini 的效果。同时，DeepSeek 还修改了产品协议，支持用户进行“模型蒸馏”，即允许用户无限制商用，鼓励蒸馏（用 R1 输出结果训练其他模型），尽可能降低用户使用壁垒，全球范围出圈和更容易建立起广泛繁荣的用户生态。

图：开源许可证协议标准



基于 R1 蒸馏的小模型性能超越 OpenAI o1-mini

	AIME 2024 pass@1	AIME 2024 cons@64	MATH- 500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	1820.0
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633.0

目录

CONTENT

1

DeepSeek的技术突破与市场定位

2

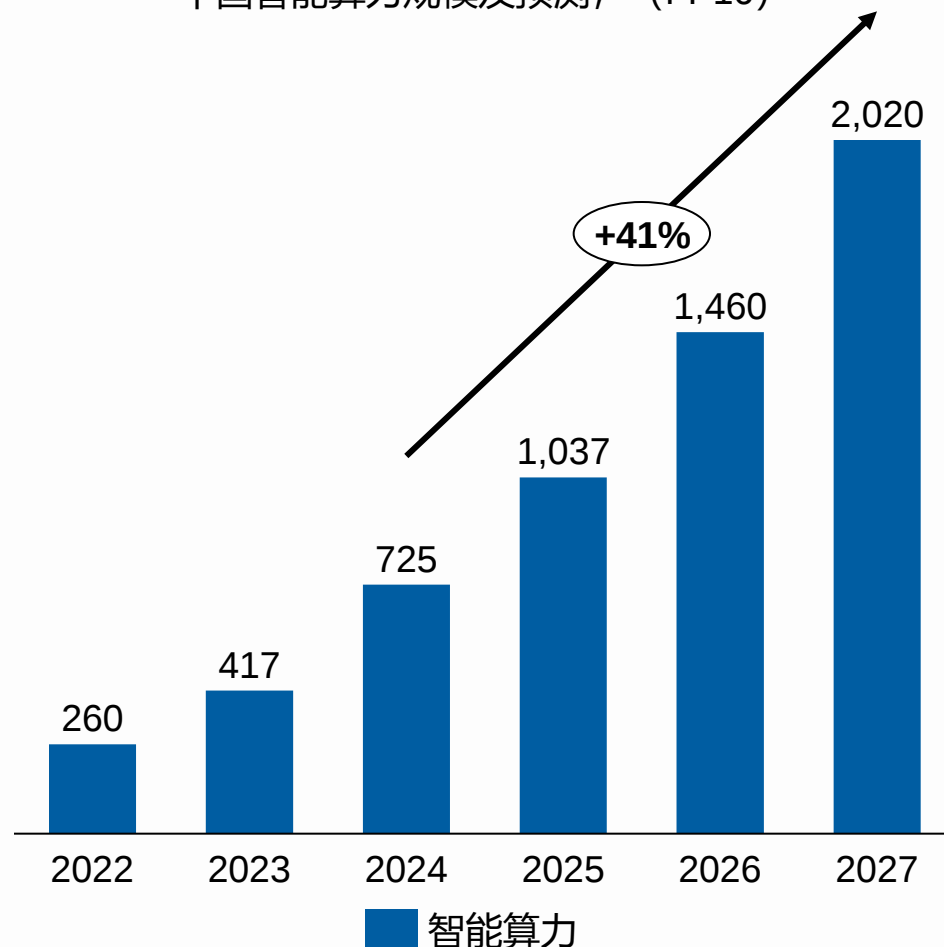
DeepSeek驱动算力需求变革

3

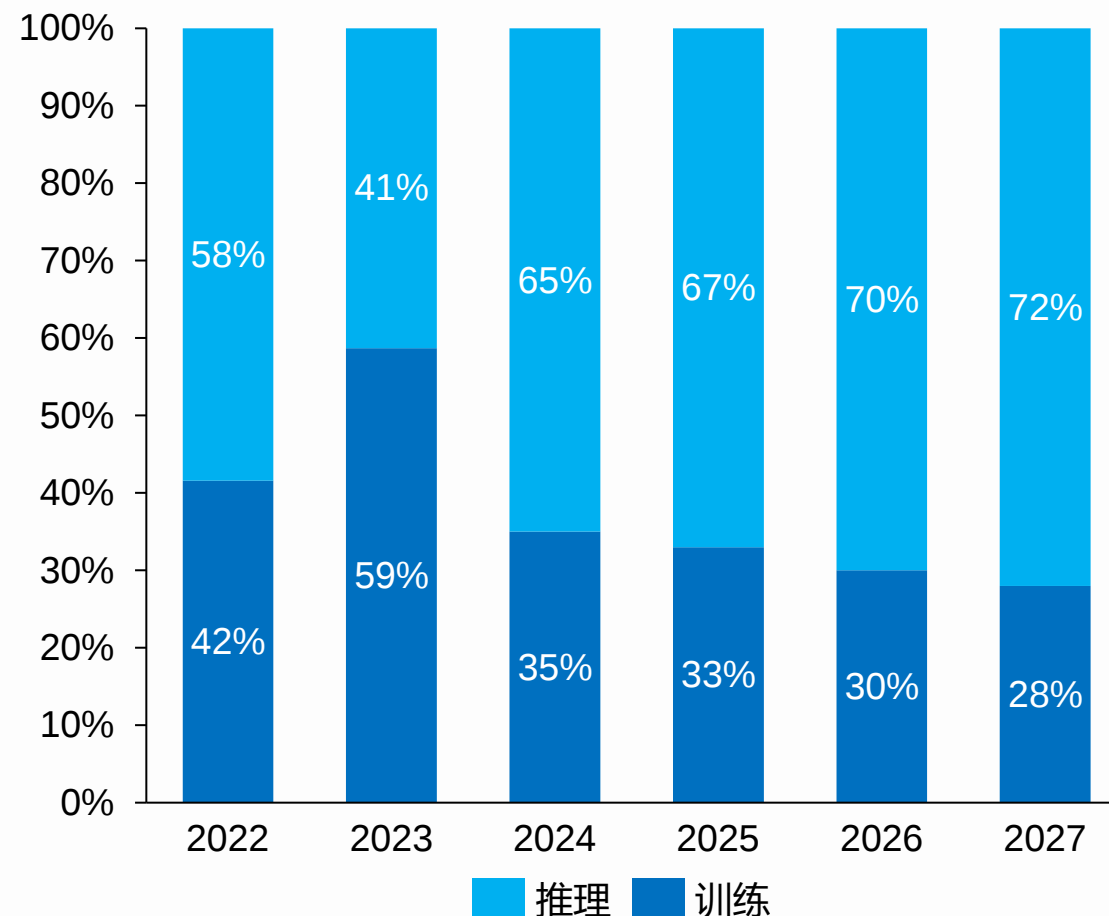
算力产业链的重构

中国智能算力市场规模持续增长，算力中心从训练侧向推理侧转移

中国智能算力规模及预测, (FP16)



中国人工智能服务器工作负载预测, 2022-2027



训练算力头部集中，推理算力爆发式增长

训练算力仍有空间和前景

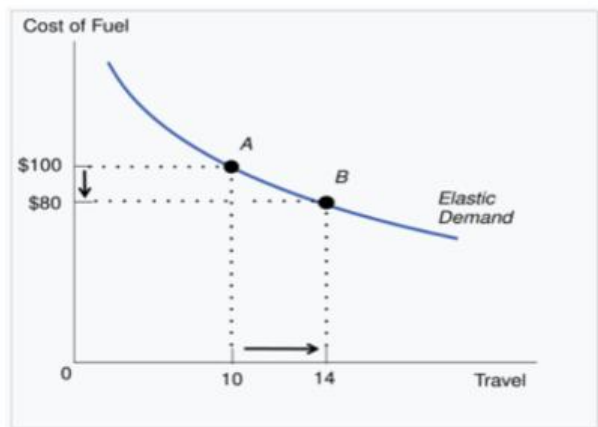
- ✓ 头部企业会持续进行教师模型的训练：模型蒸馏的前提是有一个好的教师模型，字节、阿里、百度等已经明确会持续加大投入；24年H2有些停滞的大模型训练近期已经重启
- ✓ 各模型厂商会借鉴deepseek的优化方法如FP8精度训练、跨节点通信等，与自身模型训练结合，探索更高效的模型训练方法
- ✓ 多模态的模型对算力的消耗会是近十倍的增长

推理算力爆发式增长：杰文斯悖论在推理侧上演，开源模型和较低的推理成本，有助于应用的繁荣，助推推理算力增长

头部企业仍持续加码大模型训练，追求更高性能的AGI目标。

- 阿里：未来3年的AI infra投资，超过过去10年的infra投资
- 字节：24 年资本开支 800 亿元，接近百度、阿里、腾讯三家的总和（约 1000 亿元）。25 年，字节资本开支有望达到 1600 亿元，其中约 900 亿元将用于 AI 算力的采购，700 亿元用于 IDC 基建以及网络设备。
- 百度：在2月11日的阿联酋迪拜World Governments Summit 2025峰会上，百度创始人李彦宏提到，百度需要继续在芯片、数据中心和云基础设施上加大投入，目的是为了开发下一代模型。
- 硅谷四大科技巨头（谷歌、微软、Meta、亚马逊）2025年合计资本开支超3,000亿美元，重点投向AI数据中心建设。

“杰文斯悖论”指出成本下降将刺激资源需求更大增长



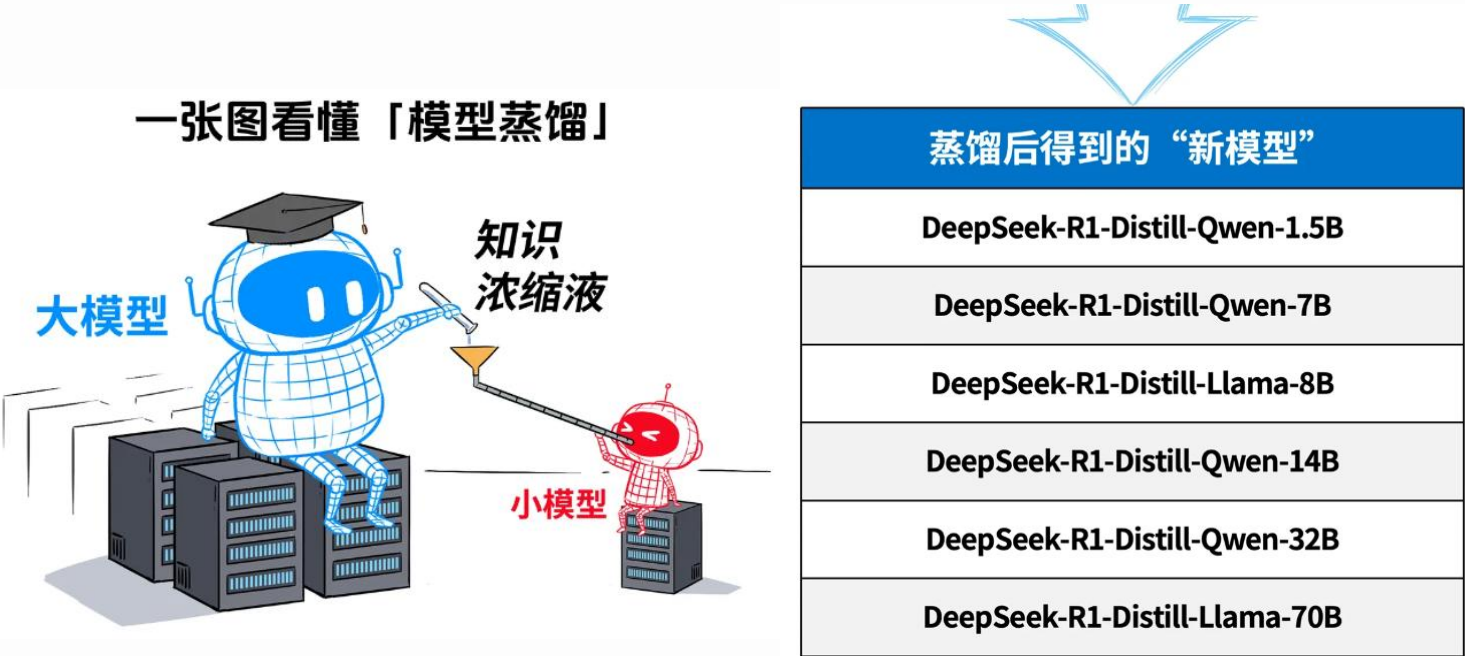
Elastic Demand: A 20% increase in efficiency causes a 40% increase in travel. Fuel consumption increases and the Jevons paradox occurs.

图表：海外大厂各年度资本开支（亿美元）

亿美元	FY21	FY22	FY23	FY24	FY25
谷歌	246	315	323	503	626
亚马逊	611	636	527	750	964
Meta	186	314	272	392	600-650
微软	232	244	352	557	800
苹果	111	107	110	94	100

模型轻量化催生端侧算力的崛起

DeepSeek通过知识蒸馏技术，将大模型压缩至轻量化版本，使其能够在端侧设备上高效运行。



数据安全与隐私计算刚需

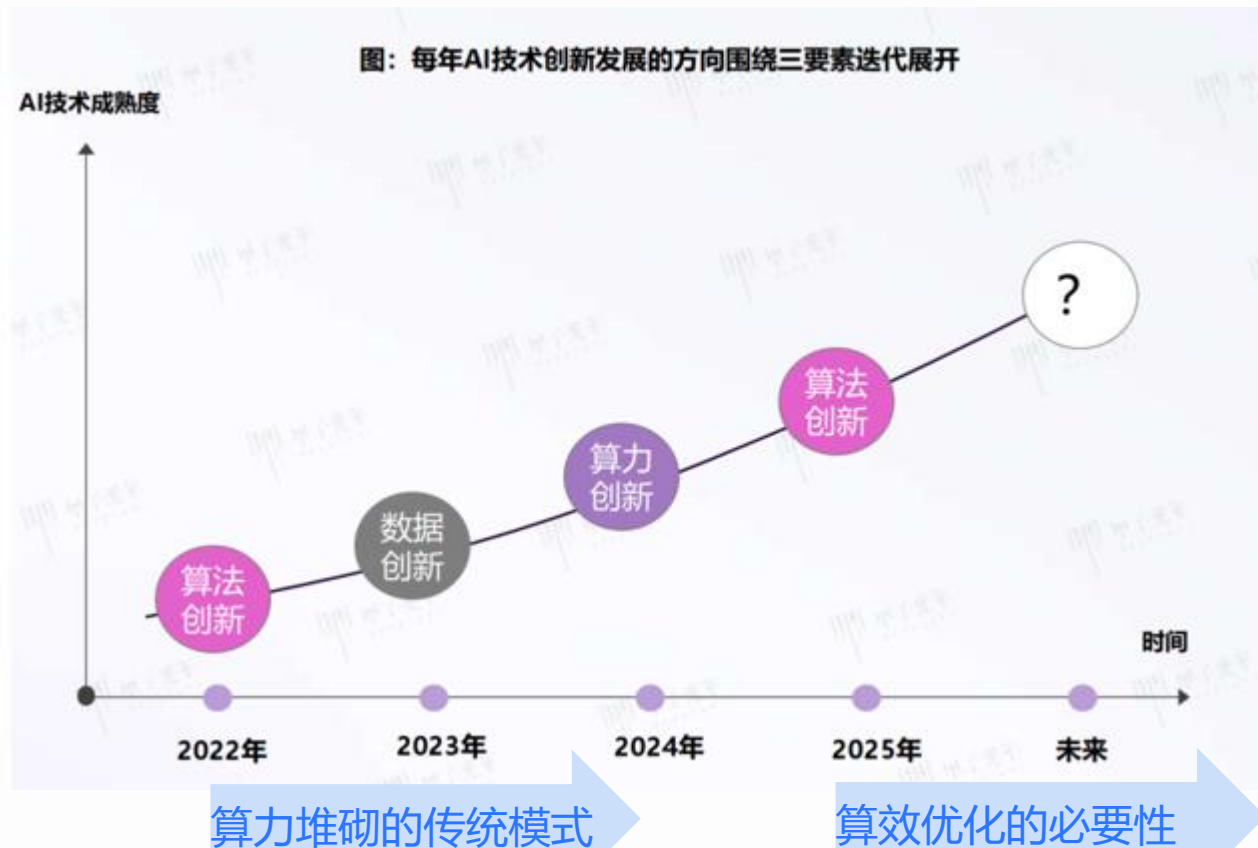
- 本地化部署需求（如医疗数据脱敏处理）
推动隐私计算技术发展，2024年数据治理市场规模超50亿元。

一体机等端侧算力市场扩容

- 国产deepseek一体机疯狂上新：三大电信运营商、浪潮、壁仞、京东云、联想、优刻得、宝德、华鲲振宇、超聚变等均推出基于不同国产芯片的deepseek一体机
- 工业质检、自动驾驶等场景需求推动边缘AI服务器出货量增长，2025年市场规模预计突破200亿元。

从“算力堆砌”到“算效优化”的产业转型

DeepSeek提出的“四两拨千斤”的技术路径推翻了统治了2023年-2024年的全球大模型产业的“暴力美学”逻辑，2025年再次进入算法创新阶段



大模型训练和应用依赖于大规模的算力堆砌，通过不断增加硬件资源来满足算力需求

随着技术发展和应用深化，算效优化成为产业转型的关键方向，通过提高算力的利用效率，降低能耗和成本

算力、数据、算法的三角创新体系，在动态循环中再次进入算法创新阶段：

- 2022年：算法创新为主，ChatGPT发布，引发Transformer架构的风潮迭起
- 2023年：数据创新为主，数据合成、数据标注等成为高质量数据集建设的热点方向
- 2024年：算力创新为主，算力迈向超万卡时代，算力运营商等产业新物种诞生
- 2025年：再次进入算法创新阶段

目录

CONTENT

1

DeepSeek的技术突破与市场定位

2

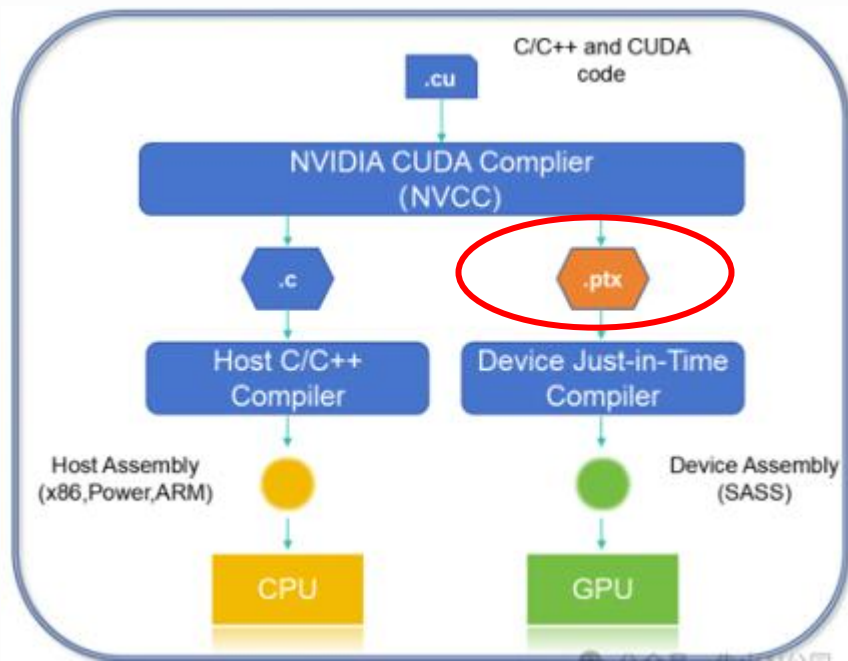
DeepSeek驱动算力需求变革

3

算力产业链的重构

DeepSeek通过PTX优化等创新技术，降低了模型训练对NV芯片的依赖，推动国产算力的应用落地

DeepSeek通过PTX手动优化跨芯片通信



- 英伟达 H800 芯片互联带宽相比 H100 被阉割，为弥补这一缺陷，DeepSeek 借助 PTX 手动优化跨芯片通信，保障数据传输效率。
- PTX 是CUDA编译的中间代码，处于高级编程语言（如 CUDA C/C++）和底层机器码（SASS）之间，起到在 CUDA 和最终机器码之间的桥梁作用。
- 借助 PTX，开发者能够直接对 GPU 的寄存器分配、线程调度等硬件级操作进行控制，实现细粒度的性能优化。在多 GPU 协同训练场景中，可通过 PTX 手动调整跨芯片通信效率，提升整体训练效能。

CUDA 生态的封闭性导致其跨硬件平台兼容性差，对国产 GPU 的适配存在较大困难。PTX 算力优化经验大幅降低了对高端 GPU 的依赖，对国产 GPU 的底层接口适配有一定帮助（需要重新设计工具链，短期内难以实现无缝迁移）

截至 2025 年 2 月 18 日，DeepSeek 已与 18 家国产 AI 芯片企业完成适配，包括华为昇腾、沐曦、天数智芯、摩尔线程、海光信息、壁仞科技、太初元基、云天励飞、燧原科技、昆仑芯、灵汐科技、鲲云科技、希姆计算、算能、清微智能和芯动力等

私有化部署及端侧小模型大量涌现，为国产芯片在实际场景的应用及试错提供了大量机会，为国产芯片的设计、性能提升等提供空间

DeepSeek采用FP8混合精度训练取得较好效果，国内芯片企业亟待增强对原生FP8支持

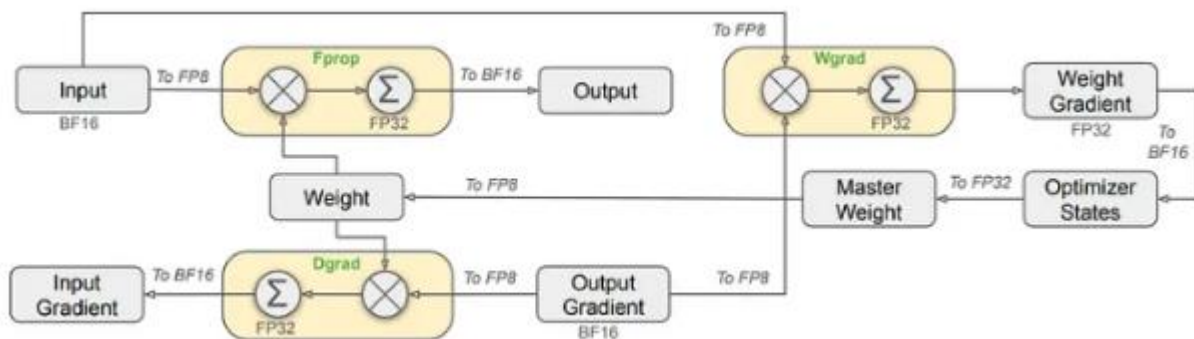
DeepSeek采用FP8混合精度训练取得较好效果：

- GPU训练时间减少40%
- 预训练成本降至仅278.8万H800 GPU小时
- 训练总费用为557.6万美元，比同类模式便宜约10倍

目前DS原生训练和推理用的是FP32、BF16和FP8，三种格式，也是DS团队探索出来效率最高的计算方式。如果不是原生支持FP8，而是需要通过其他计算精度的转译，至少损失30%性能和20%的智商

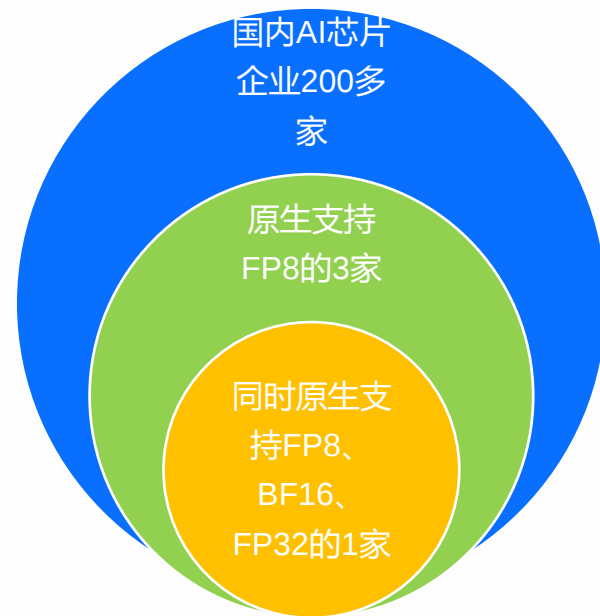
目前国内有200多家AI芯片公司，原生支持FP8计算格式的AI芯片只有3款，同时支持三种计算格式的国产AI芯片公司只有1款。

采用FP8数据格式的整体混合精度框架

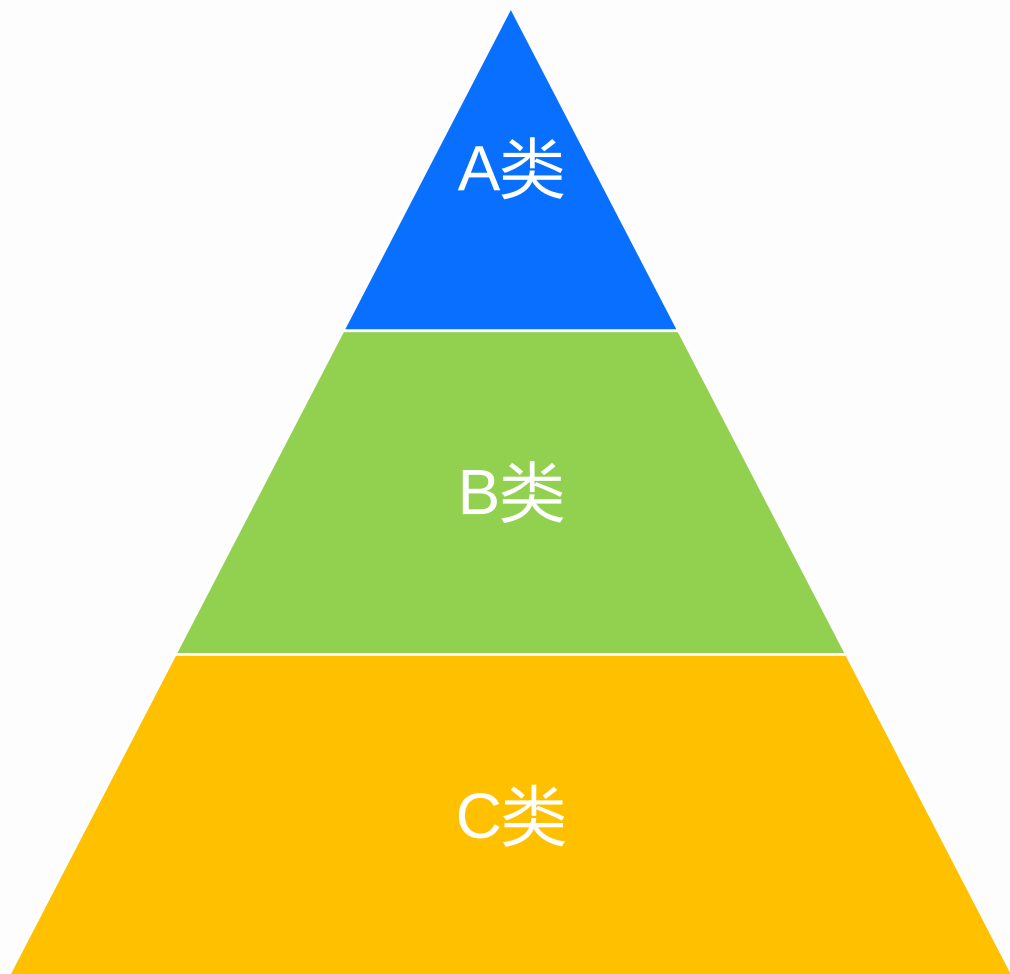


- 在DeepSeek的训练过程中，绝大多数**核心计算核**（即通用矩阵乘法GEMM操作）**均以FP8精度实现**。这些GEMM操作接受FP8张量作为输入，并输出BF16或FP32格式的结果。如下图所示，与线性算子（Linear operator）相关的三个GEMM运算——前向传播（Fprop）、激活梯度反向传播（Dgrad）和权重梯度反向传播（Wgrad）——都采用FP8精度执行。
- 对以下模块维持原有精度（如BF16或FP32）：嵌入模块（embedding module）、输出头（output head）、混合专家门控模块（MoE gating modules）、标准化算子（normalization operators）以及注意力算子（attention operators）。（尽管FP8格式具有计算效率优势，但由于部分算子对低精度计算较为敏感，仍需保持更高计算精度）

国内芯片对三种计算精度的支持情况



智算中心分为三类



定位

功能

规模

芯片

预训练	用于训练超大参数量的原创教师大模型，如移动的九天大模型、阿里的通义千问等	万卡以上	H200、B200等最先进的芯片或国产高端芯片（针对有强信创需求的企业）
后训练	用于学生大模型的调优，训练行业化、客制化大模型	几十台到几百台为主	A100/A800、H100/H800，或者采购部分高端国产卡
推理	用于推理的算力中心，针对模型在企业端现实场景的实际应用	大小不等	利旧原有设备或者经营不善的B类3090/4090或910A、910B及其他国产卡

推理类智算中心爆发增长，超大规模智算中心建设加快

智算中心	市场影响
A类	建设速度不减： 头部科技大厂仍计划大量投资；超前建设的需要；下一轮AI技术的涌现（如多模态等）仍需要十倍左右的算力支撑；中美博弈（美国“星际之门”、欧洲“Invest AI计划”等）
B类	结构性过剩，建设减缓： 规模小、位置偏僻、型号旧&性价比低、国产算力等类型的智算中心闲置状况严重 新建要看是不是有强主体包销，如果有强主题的3-5年包销合同，依然可以正常建设；如果是弱主体的客户，甚至没有客户的前提下，建设可能会暂缓或者停滞
C类	爆发式增长： 推理算力需求大幅增加，端侧、边缘侧分布式算力部署快速增长，私有化部署1~10台GPU服务器的小规模集群需求在内部部署中爆发。

算力包销合同主题分强、弱两类，市场上更多的是弱主体：

- 强主体：各个参与方资质和信誉主题都很强。央国企、A股上市公司承建、金融机构垫资、互联网大厂包销，这种主要是H系列为主，风险相对可控，互联网大厂可以用来做B类或C类。
- 弱主体：相对于强主体，出资方、承建方、包销方相对来说资质弱一些，比如包销方是一些AI大模型创业公司、创新型实验室或者，这类24年下半年已经开始毁约，风险非常高，H系列租金每个月6~8w/台。这类的算力中心风险非常大，需要注意，这类算力使用方，未来多转向C类算力租赁。

中国大模型主要有两类玩家

企业类型

典型玩家

科技大厂



大模型玩家

AI创业公司



Deepseek发布后科技大厂：拥抱DeepSeek，同时跟进类似的自研产品

- 投资加大：前文已论述
- 产品：科技大厂一方面拥抱DeepSeek，一方面跟进类似的自研产品

国际大厂也加快了产品的推陈出新

公司	时间	动作概况
阿里	1月29日	• 阿里云 PAI Model Gallery 支持一键部署 DeepSeek-V3 和 DeepSeek-R1 • 阿里云发布开源的通用千问Qwen 2.5-Max MoE（混合专家模型），它使用了与DeepSeek-R1类似的技术路线
百度	2月14日	• 百度搜索全面接入 DeepSeek。 • 百度宣布文心一言4月1日起开源免费，并计划推出文心大模型 4.5 系列，于 6 月 30 日起正式开源
腾讯	/	• 从云平台腾讯云、腾讯云旗下大模型知识应用开发平台知识引擎、国民应用微信、AI智能工作台ima、主力AI应用元宝全方位拥抱DeepSeek，纷纷宣布接入R1模型
华为	/	• 鸿蒙系统的小艺助手接入DeepSeek-R1；与硅基流动合作，基于昇腾云服务推出DeepSeek-R1/V3推理服务

OpenAI连续官宣GPT-4.5在几周内上线，GPT-5在几个月内上线，及模型路线规模的调整



OpenAI @OpenAI · Feb 13

Two updates you'll like—

OpenAI o1 and o3-mini now support both file & image uploads in ChatGPT

We raised o3-mini-high limits by 7x for Plus users to up to 50 per day

Deepseek发布后AI创业企业：从参数竞争到进入理性期

AI六小虎的策略变化

- 仍坚守大模型预训练，但技术路线分化：
 - ✓ 智谱（引入强化学习和多模态，注重B端市场）
 - ✓ 月之暗面（长文本）
 - ✓ 阶跃星辰（多模态）
- 转向细分领域：
 - ✓ 零一万物与阿里合作产业实验室，放弃超级大模型研发
 - ✓ 百川智能专注医疗赛道
 - ✓ MiniMax布局海外市场

初创企业大模型六小虎的动作概况

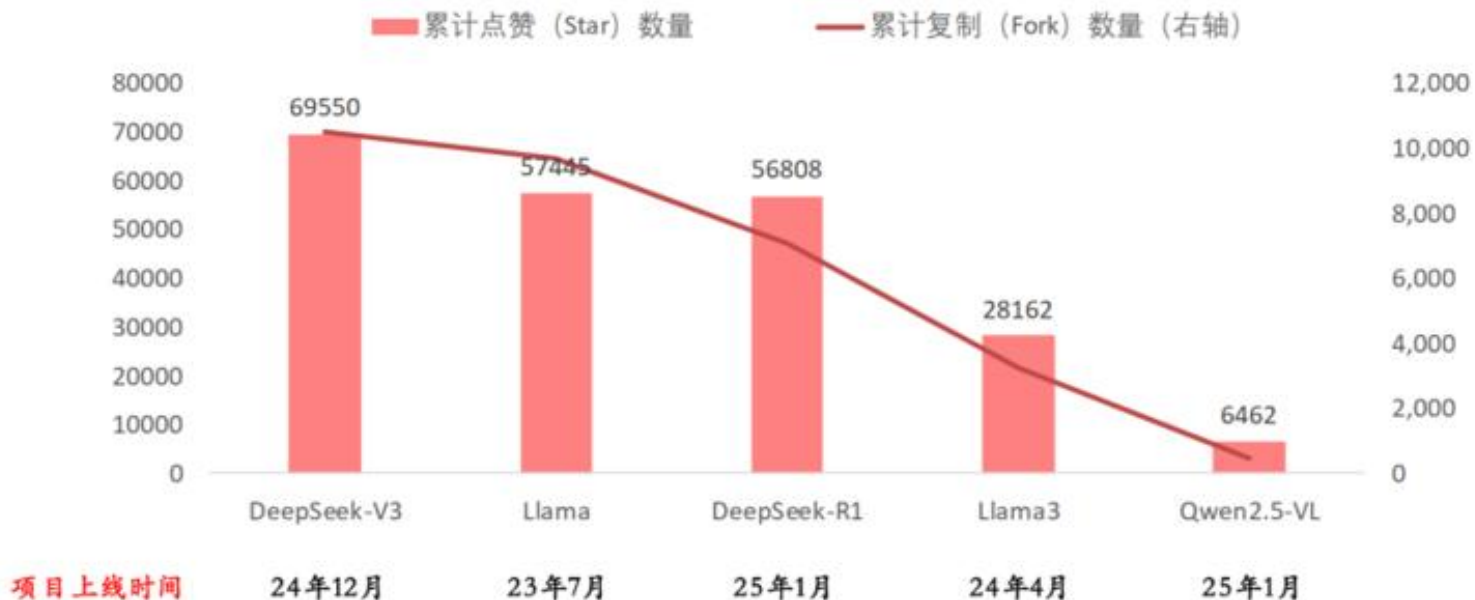
公司	时间	动作概况
零一万物	2月14日	与苏州高新区联合成立的产业大模型基地正式授牌
百川智能	1月25日	发布新模型Baichuan-M1-preview
阶跃星辰	2月13日	联合研发的「AI儿科医生」在北京儿童医院上岗
	1月20日	发布新语言大模型Step-2-mini和Step-2 文学大师版
	1月21日	升级语音模型Step-lo Audio，上新多模态理解大模型Step-lo Vision
	1月22日	发布视频生成模型Step-Video V2版本
	1月24日	应用端「跃问」推出「跃问AI创意板」功能
	/	「跃问」接入DeepSeek-R1
	2月21日	举办首届 “Step Up 生态开放日”
智谱华章	2月11日	创立发起人唐杰出席第三届人工智能行动峰会边会 “人工智能技术进步与应用” 并发言
	2月11日	Agentic GLM登陆三星最新款Galaxy S25系列手机
	/	和AI画图捏角色的应用软件「捏ta」展开系列合作
月之暗面	1月20日	发布Kimi k1.5多模态思考模型
MiniMax	1月20日	升级发布T2A-01系列语音模型，并上线海螺语音产品

大模型领域迎来“安卓时刻”，大量AI应用将爆发式出现

回顾安卓与iOS应用的发展，安卓系统发布一年后，大量的安卓应用开始出现。现在的deepseek类似当初的安卓系统

- GitHub 的 Stars 是项目在社区中受欢迎程度的直接指标，Fork 则表示项目累计被用户拷贝的数量，两个指标均代表项目上线至今的关注度和用户喜爱度。DeepSeek V3 和 R1 两个项目上线至今均不足 2 个月，但它们的累计 Star 和 Fork 均与上线时间更早的 Llama 接近，显著高于 24 年 4 月发布的 Llama3，直接反映了开发者对 DeepSeek 开源模型的高认可度。
- 根据GitHub、Hugging Face社区上的开发者实测，经过R1微调的80亿参数小模型可以在个人笔记本中运行，本地化部署门槛显著下降，应用的开发将迎来百花齐放。

图：DeepSeek累计关注度高于更早发布的Llama（根据GitHub统计）



两个关键词：端侧AI、AI Agent

- 端侧AI

- AI Agent

感谢您的聆听!

中科智道(北京)科技股份有限公司