

DeepSeek研究框架 ——计算机人工智能系列深度报告

评级：推荐(维持)

刘熹(证券分析师)
S0350523040001
liux10@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



相对沪深300表现

表现	1M	3M	12M
计算机	30.1%	3.3%	57.5%
沪深300	5.0%	-4.1%	16.5%

相关报告

《计算机行业点评报告：DeepSeek搅动了全球AI的“一池春水”（推荐）*计算机*刘熹》——2025-02-03

《美国对华AI限制加剧，自主可控大势所趋——AI算力“卖水人”系列（4）（推荐）*计算机*刘熹》——2025-01-24

《计算机行业事件点评：我国中部最大智算中心投产，国产算力景气上行（推荐）*计算机*刘熹》——2025-01-14

DeepSeek引领全球AI创新，一定程度上也影响了全球AI格局，并提振了国内AI产业信心。我们看好由DeepSeek带来的AI产业、尤其是国内AI产业的发展机遇，包括AI应用、端侧AI、算力等三个方向。

◆ DeepSeek（深度求索）专注大模型技术，V3和R1模型惊艳海内外

DeepSeek（深度求索）成立于2023年7月，由量化资管公司幻方量化创立，DeepSeek专注于开发先进的大语言模型（LLM）和相关技术。2024年1月5日，发布第一个大模型DeepSeek LLM；12月26日，上线DeepSeek-V3并同步开源，DeepSeek-V3采用FP8训练，性能对其世界顶尖的闭源模型GPT-4o以及Claude-3.5-Sonnet。2025年1月20日，发布DeepSeek-R1，DeepSeek-R1在数学、代码、自然语言推理等任务上，性能比肩OpenAI o1正式版。DeepSeek-R1推出后广受关注，据Appfigures、Sensor Tower报告，1月26日以来，深度求索（DeepSeek）发布的DeepSeek AI智能助手爆火，在全球140个市场的应用商店下载榜上排名第一。

DeepSeek V3和R1模型基于Transformer架构，采用了MLA和DeepSeek MoE两大核心技术，引入了多令牌预测、FP8混合精度训练等创新技术，显著提升了模型的训练效率和推理性能。DeepSeek创始人梁文锋表示“V2模型没有海外回来的人，都是本土的”。DeepSeek代表中国本土AI大模型，也代表开源AI走在了全球AI市场的前列。

◆ DeepSeek对全球AI行业影响颇深：激发创新、提振国产、推广开源

1) DeepSeek成为了全球AI的一条“鲶鱼”。DeepSeek发布或导致全球AI格局变化，中美AI形势生变，全球AI被“鲶鱼”激活。预计美系AI会不断反应，全球AI模型迭代和发布频率将提速，投入继续加大。自1月20日DeepSeek-R1发布以来，OpenAI连续发布了Agent operator, O3 mini、Deep Research等模型，OpenAI CEO表示GPT-5将是超级混合模型，计划把GPT和o系列模型整合在一起。

2) DeepSeek驱动国产AI估值重塑。我们认为：长期以来，算力和技术是制约国内AI估值的主要因素，DeepSeek在国内AI芯片受限的环境里，通过本土AI团队，探索出一条“算法创新+有限算力”的新路径，较大地提振了国内AI产业信心。DeepSeek-R1的推出或同时打破了抑制国产AI产业的技术和算力这两项天花板，将驱动国产AI软硬件迎估值重塑。

3) DeepSeek是开源AI的“ChatGPT时刻”。OpenAI CEO首次承认OpenAI的闭源策略“站在了历史错误的一边”。DeepSeek-R1开源将会吸引更多人参与到大模型研发中，并通过蒸馏等技术显著提升推理AI、小模型的性能，将大幅加速全球AI创新，加速AI推理进程，普惠AI、AI平权将驱动DeepSeek迅速推广，近期全球CSP大厂密集上架DeepSeek能力也验证了这点，我们预计Killer APP的诞生或将临近。

◆ DeepSeek推动AGI时代到来，关注AI应用、端侧AI、算力三大主线

1) **AI应用**：DeepSeek的创新带来成本极致优化，带来AI普惠、AI平权，将加速AI应用的创新，国内AI应用将受益于DeepSeek实现能力显著提升，应用上游的模型API的价格下降也将驱动应用厂商的商业模式快速成熟。

2) **AI端侧**：DeepSeek支持用户进行“模型蒸馏”，并通过DeepSeek-R1的输出，蒸馏了6个小模型开源给社区。端侧AI能力过去受限于端侧AI计算影响，DeepSeek将显著提升端侧小模型的能力，进而提升AI终端能力。

3) **算力**：杰文斯悖论指出当我们希望通过技术进步来提高资源效率时，可能会导致资源的消耗增加。我们预计DeepSeek带来的大模型推理成本的优化，将加速AI的普及推广，和下游应用的商业模式构建，并推动AI算力进入由终端用户需求驱动的长增长周期。

◆ 投资建议

DeepSeek探索出一条“算法创新+有限算力”的新路径，开源AI时代或已至，国产AI估值或将重塑，维持计算机行业“推荐”评级。

◆ 相关公司

1) **AI应用**：①**2G**：中科曙光、科大讯飞、中国软件、太极股份、深桑达A、中科星图、国投智能、云从科技、能科科技、拓尔思、航天信息、税友股份、金财互联、浪潮软件、数字政通；②**2B**：金蝶国际、卫宁健康、石基信息、明源云、新致软件、用友网络、广联达、莱斯信息、四川九洲、泛微网络、致远互联、新开普、东方财富、同花顺、恒生电子、宇信科技、当虹科技、万达信息、创业惠康、润和软件、彩讯股份、第四范式、焦点科技；③**2C**：金山办公、三六零、万兴科技、福昕软件、合合信息、萤石网络。

2) **算力**：①**云**：海光信息、寒武纪、浪潮信息、华勤技术、云赛智联、光环新网、中兴通讯、宝信软件、紫光股份、中国电信、优刻得-W、青云科技-U、首都在线、并行科技、润泽科技、中国软件国际、神州数码、深信服、新炬网络、天玑科技；②**边**：网宿科技、顺网科技、云天励飞；③**端**：软通动力、中科创达、乐鑫科技、移远通信。

◆ **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等、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。

一、DeepSeek背景介绍

- 1.1、DeepSeek股权结构及创始人背景
- 1.2、DeepSeek母公司幻方量化，早期确立AI战略为后续出圈埋下伏笔
- 1.3、DeepSeek重视年轻团队且兼具深厚技术底蕴，薪酬对标字节跳动研发岗
- 1.4、DeepSeek产品家族全梳理
- 1.5、DeepSeek日活远超同期ChatGPT，下载量霸榜全球140个市场移动应用榜首
- 1.6、DeepSeek获得海内外市场认可，中国AI产业首次步入引领位置

二、DeepSeek模型家族技术详解

- 2.1、DeepSeek模型家族技术创新框架总揽
- 2.2、DeepSeek v3：性能对齐海外领军闭源模型，DeepSeek2024年巅峰之作
- 2.3、DeepSeek R1 Zero核心创新点——RL（强化学习）替代SFT（有监督微调）
- 2.4、DeepSeek R1：高质量冷启动数据+多阶段训练，将强推理能力泛化
- 2.5、开源大模型：打破OpenAI等闭源模型生态，提升世界对中国AI大模型认知

三、DeepSeek对AI应用的影响？

- 3.1、DeepSeek打开低成本推理模型边界，加速AI应用布局进程
- 3.2、DeepSeek R1蒸馏赋予小模型高性能，端侧AI迎来奇点时刻

四、DeepSeek对算力影响？

- 4.1、DeepSeek V3训练中GPU成本558万美元，对比海外成本降低
- 4.2、DeepSeek或有约5万Hopper GPU，训练总成本或较高
- 4.3、推理化：推理算力需求占比提升，GenAI云厂商有望受益

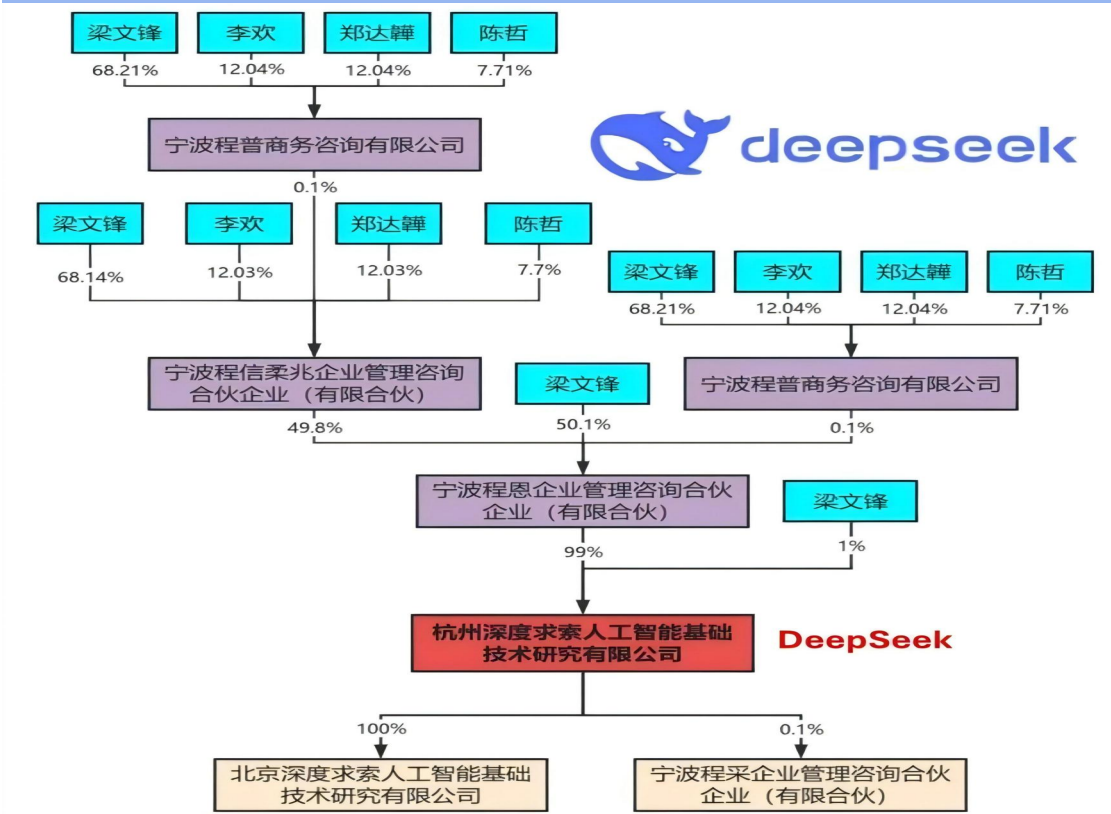
五、盈利预测及风险提示

一、DeepSeek背景介绍

1.1、DeepSeek股权结构及创始人背景

- 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。从股权结构图显示，DeepSeek由四名自然人通过五层控股掌握100%股份（其中梁文锋间接持股比例83.29%，直接持股1%，累计84.2945%）。
- 创始人梁文锋出生于广东湛江，浙江大学毕业，拥有信息与电子工程学系本科和硕士学位，2008年起开始带领团队使用机器学习等技术探索全自动量化交易，2015年幻方量化正式成立。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。

图：DeepSeek股权结构



图：幻方创始人梁文峰（图右）



1.2、DeepSeek母公司幻方量化，早期确立AI战略为后续出圈埋下伏笔

- 母公司幻方量化确立以AI为发展方向。2016年，幻方量化第一个由深度学习算法模型生成的股票仓位上线实盘交易，使用GPU进行计算。随后不久，该公司明确宣布AI为主要发展方向。
- 量化投资全面AI化驱动囤卡需求，为后续蜕变埋下伏笔。复杂的模型计算需求使得单机训练遭遇算力瓶颈，训练需求和有限的计算资源产生矛盾，幻方需要解决算力受限难题。于是幻方在2019年成立了一家AI基础研究公司，并推出自研的“萤火一号”AI集群，搭载500块显卡。2021年，幻方又斥资10亿元建设“萤火二号”，为AI研究提供算力支持。幻方在构建AI算力过程中的“囤卡”动作为它赢得了市场机会。作为国内早期的英伟达芯片大买家之一，2022年其用于科研支持的闲时算力高达1533万GPU时，大大超越了后来很多大模型公司。

图：幻方量化发展历程



图：幻方官网首页标语，以AI为核心发展方向



1.3、DeepSeek重视年轻团队且兼具深厚技术底蕴，薪酬水平对标字节跳动研发岗



- 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。早在2022年，幻方量化便着手为DeepSeek筹建AI团队，至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。技术报告的致谢栏揭示，参与DeepSeek V3研发的工程师阵容，已壮大至139人。
- 团队薪酬水平对标字节跳动研发岗位，且不限人才算力使用。据36氪资料显示，DeepSeek薪酬水平对标的字节研发，“根据人才能拿到的字节offer，再往上加价”；同时只要梁文锋判断技术提案有潜力，DeepSeek给人才的算力，“不限”。

图：DeepSeek公开招聘职位信息汇总

职位名称	面向群体	申请要求	薪酬水平
深度学习研究员	校招&实习	熟练掌握至少两种编程语言；在国际顶会或期刊发表相关论文；知名比赛成绩	8-11万元/月，一年14薪
资深ui设计师	经验不限，本科	优秀的艺术类教育背景；有互联网或科技公司UI设计工作经验；	4-7万元/月，一年14薪
深度学习研发工程师	在校/应届，本科	较强的工程能力；工程能力；知名比赛成绩	4-7万元/月，一年14薪
数据架构工程师	在校/应届，本科	有搜索、推荐、广告等业务数据的处理经验；有 规模中文网页数据收集和清洗经验者优先	4.5-6.5万元/月，一年14薪
全栈开发工程师	在校/应届，本科	对主流的开源软件有深入的了解，并且对此有做出贡献	2.5-5万元/月，一年14薪
客户端研发工程师	在校/应届，本科	计算机或相关专业优先；有独立开发App经验，有优秀开源项目者优先。	2-4万元/月，一年14薪
深度学习实习生	计算机及相关专业研究生，特别优秀的本科生；	具有扎实的编程功底；有顶级AI会议论文发表经验或开源项目贡献经验者优先	500元/天，4天一周，6个月；非北京地区学生来京实习有租房补助3000元/月

1.4、DeepSeek产品家族全梳理



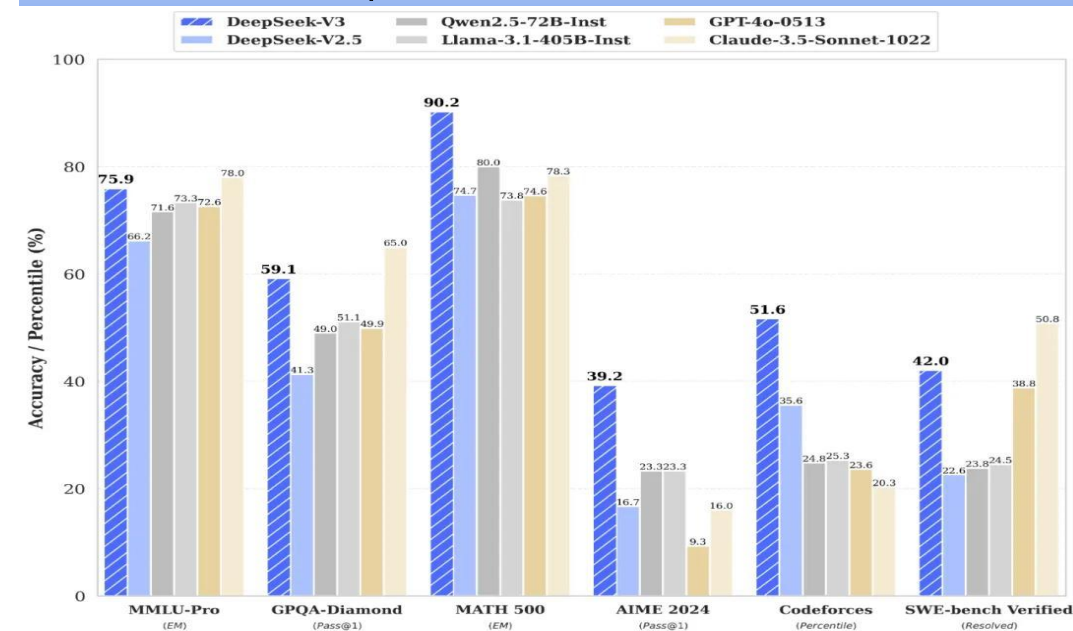
模型类别	日期	名称	内容	对标
LLM	2023年11月2日	DeepSeek Coder	模型包括 1B, 7B, 33B 多种尺寸, 开源内容包含 Base 模型和指令调优模型。	Meta的CodeLlama是业内标杆, 但DeepSeek Coder展示出多方位领先的架势。
	2024年6月17日	DeepSeek Coder V2	代码大模型, 提供了 236B 和 16B 两种版本。DeepSeek Coder V2 的 API 服务也同步上线, 价格依旧是「1元/百万输入, 2元/百万输出」。	能力超越了当时最先进的闭源模型 GPT-4-Turbo。
	2023年11月29日	DeepSeek 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。	
架构开源	2024年1月11日	DeepSeekMoE	开源了国内首个 MoE (混合专家架构) 大模型 DeepSeekMoE: 全新架构, 支持中英, 免费商用, 在 2B、16B、145B 等多个尺度上均领先	被普遍认为是 OpenAI GPT-4 性能突破的关键所在

1.4.1、DeepSeek V3性能位居全球领先水平，代码/数学/中文能力测试表现优异



- 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对比领域开源/闭源模型



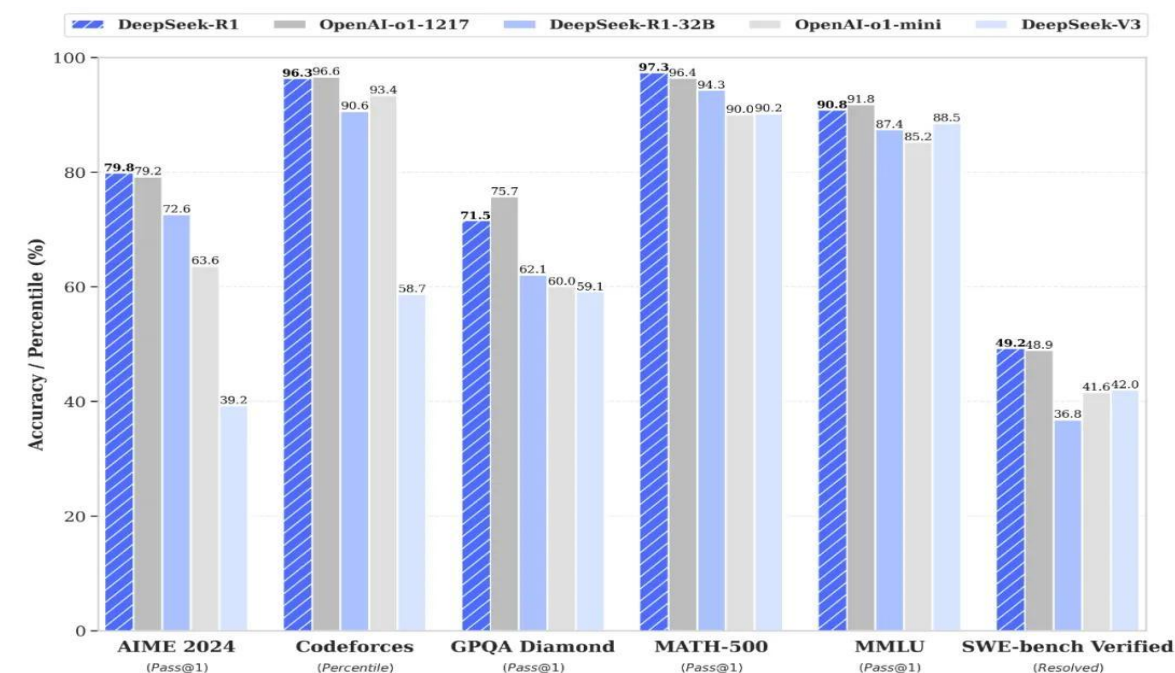
图：DeepSeek-V3在英文、代码、数学领域表现优异

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

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

- DeepSeek-R1性能比较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。DeepSeek R1同步登录DeepSeek官网或官方App。网页或者app端打开“深度思考”模式，即可调用最新版 DeepSeek-R1 完成各类推理任务。
- 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

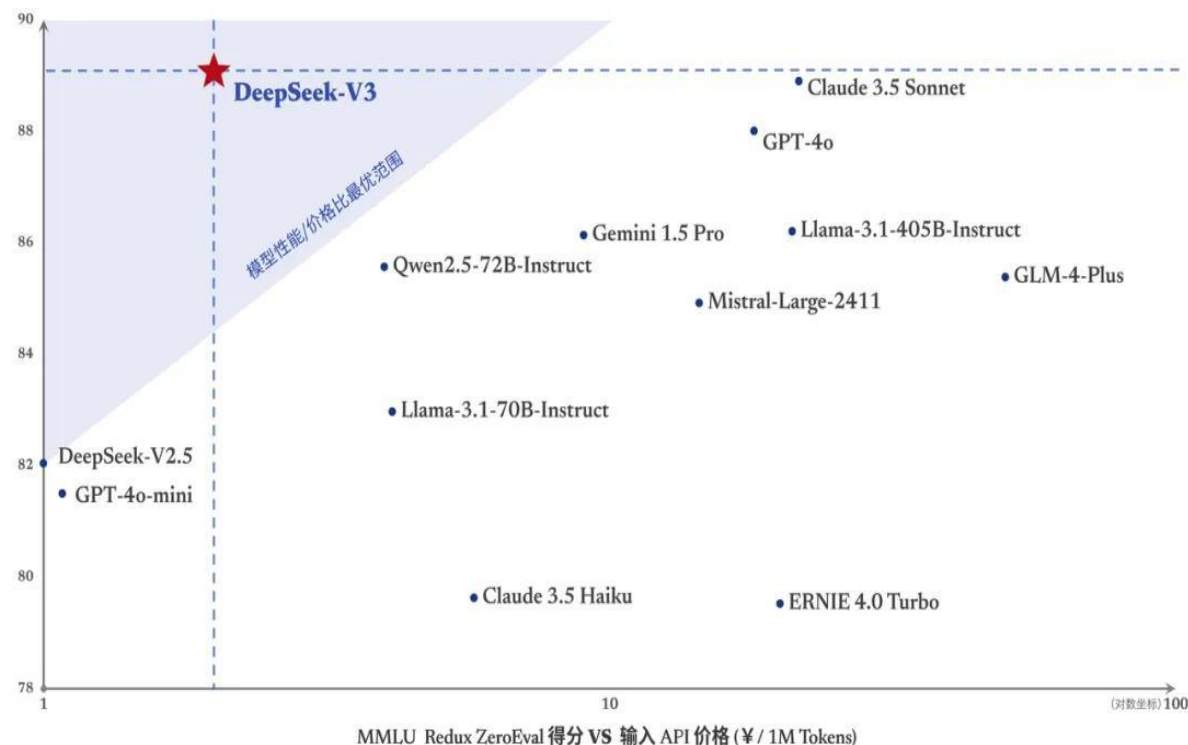


1.4.2、DeepSeek-V3/R1均具备领先的性价比优势

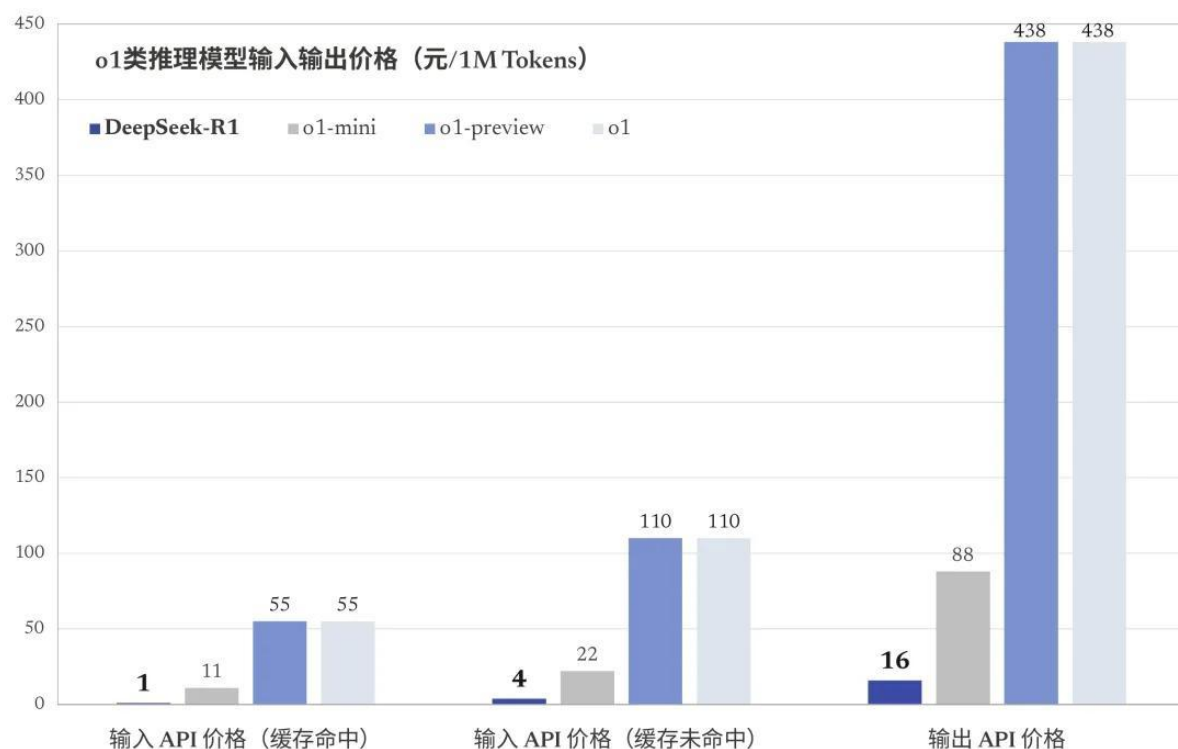
✓ DeepSeek 系列模型均极具定价优势。

- ✓ 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



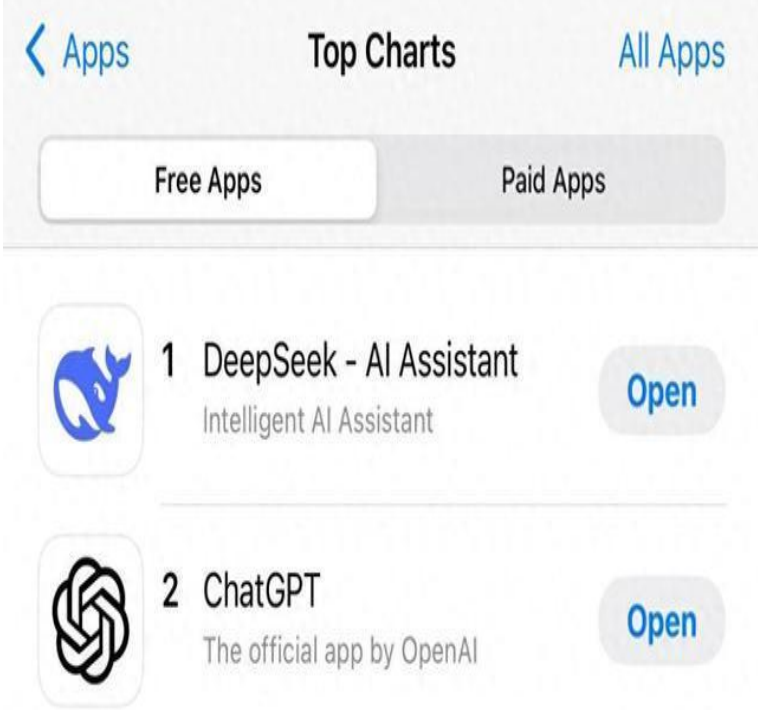
1.5、DeepSeek日活远超同期ChatGPT，下载量霸榜全球140个市场移动应用榜首

DeepSeek远超同期ChatGPT，AI格局或迎来重塑。2025年1月15日，DeepSeek 官方 App 正式发布，并在 iOS/Android 各大应用市场全面上线。数据显示，DeepSeek在上线18天内达到日活跃用户1500万的成就，相较之下，同期ChatGPT则耗费244天才实现相同日活；2月4日，上线20天后日活突破2000万，创下又一个新纪录。DeepSeek在发布的前18天内累计下载量达到1600万次，峰值日下载量高达500万次，几乎是ChatGPT同期900万次下载量的两倍。此外，DeepSeek在全球140个市场中的移动应用下载量排行榜上位居榜首。

图：DeepSeek对话助手

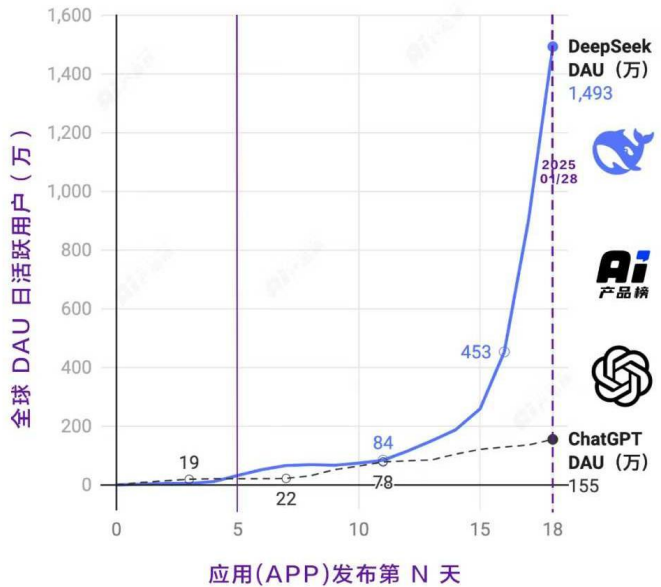


图：DeepSeek霸榜下载榜榜首



图：DeepSeek成全球增速最快的AI应用

仅上线18天日活1500万，增速是ChatGPT的13倍

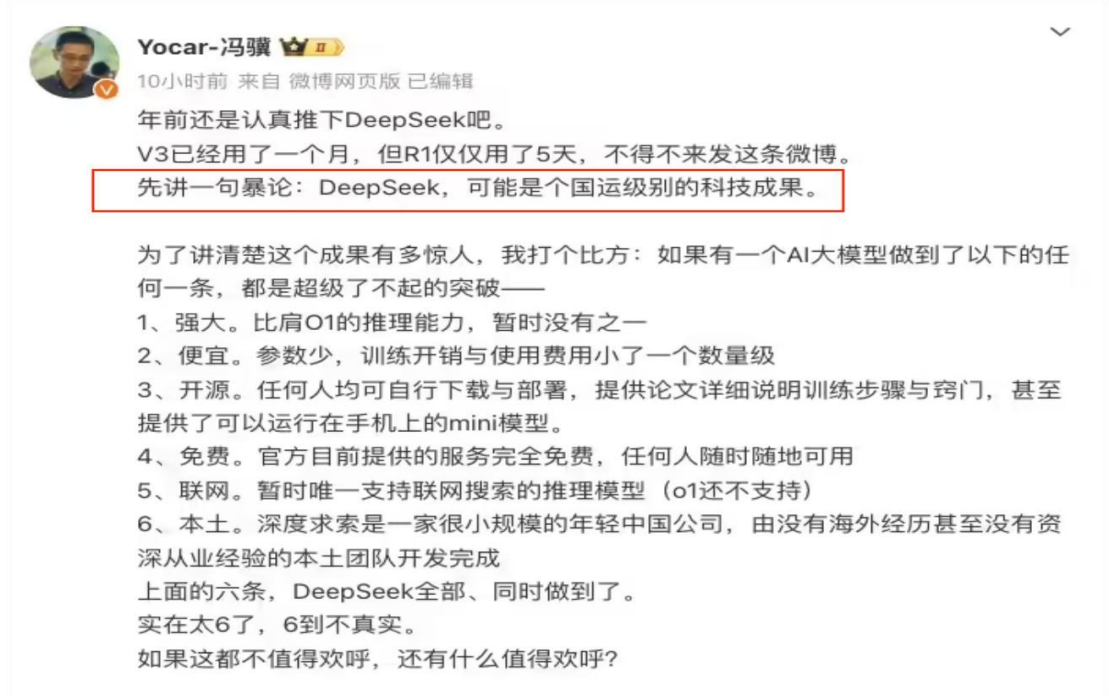


- DeepSeek惊艳海外市场，中国AI产业首次步入引领位置。
- 英伟达表示，DeepSeek为推理、数学和编码等任务提供了“最先进的推理能力”“高推理效率”以及“领先的准确性”。
- Meta首席AI科学家Yann Lecun表示“DeepSeek-R1面世与其说意味着中国公司在AI领域正在超越美国公司，不如说意味着开源大模型正在超越闭源。”
- OpenAI首席执行官Sam Altman首次承认OpenAI的闭源策略“站在了历史错误的一边”。
- 微软COE纳德拉表示，公司的DeepSeekR1模型展现了"真正的创新力"。
- 国内黑神话制作人悟空冯冀表示，DeepSeek 可能是个国运级别的科技成果。

图：Sam Altman评价DeepSeek



图：黑悟空神话制作人评价DeepSeek

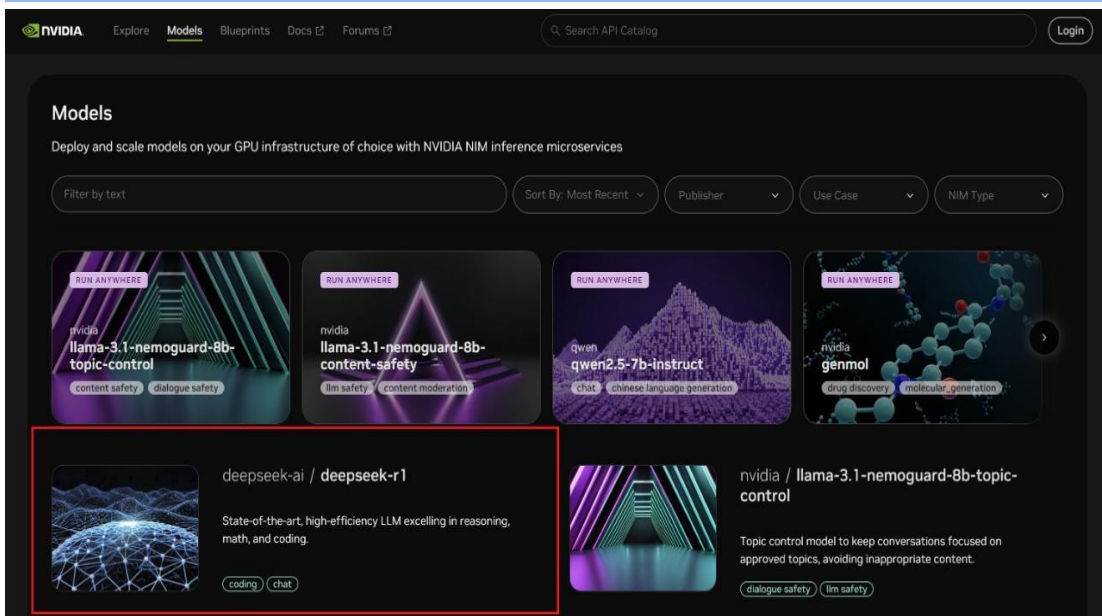


1.7、微软/英伟达/亚马逊/腾讯/华为等一众科技巨头拥抱DeepSeek

✓ 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AWS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



二、DeepSeek模型家族技术详解

2.1、DeepSeek模型家族技术创新框架总揽

DeepSeek V3

MoE架构模型

核心创新

- 1、多头潜在注意力（MLA）
使用低秩联合压缩方法减少注意力计算的缓存需求，同时保持多头注意力的性能。
- 2、混合专家架构（DeepSeekMoE）
 - ① 细粒度专家分割
 - ② 共享专家隔离
 - ③ 辅助损失优化的专家负载平衡策略。
- 3、多 Token 预测目标（MTP）
扩展模型在每个位置预测多个未来 token 的能力，提高训练数据效率。
- 4、DualPipe算法。
- 5、支持 FP8 混合精度训练。

DeepSeek R1 Zero

以V3作为基础模型，纯强化学习替代有监督微调

核心创新

- 1、强化学习算法：使用 GRPO 框架，通过群体奖励优化策略模型。奖励设计包括准确性奖励和格式奖励。
- 2、自我演化与顿悟时刻：模型通过 RL 自动学习复杂的推理行为，如自我验证和反思。随着训练过程的深入，模型逐步提升了复杂任务的解答能力，并在推理任务上显现突破性的性能提升。

DeepSeek R1

以V3作为基础模型，结合冷启动数据的多阶段训练

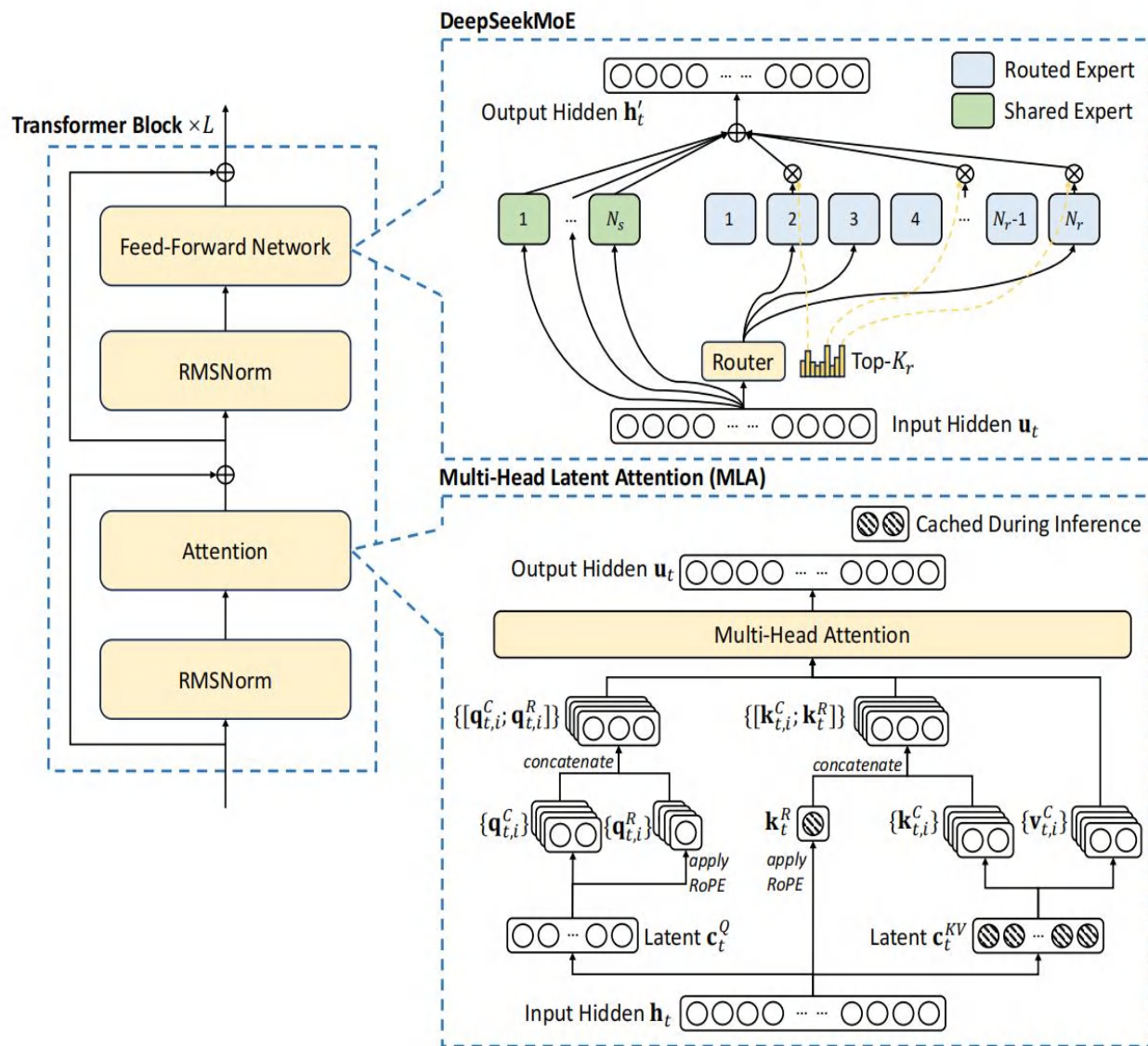
核心创新

- 1、冷启动数据引入：从零开始的 RL 容易导致初期性能不稳定，为此设计了包含高质量推理链的冷启动数据集。该数据提高了模型的可读性和训练初期的稳定性。
- 2、推理导向的强化学习：通过多轮 RL，进一步优化模型在数学、编程等推理密集型任务中的表现。
- 3、监督微调与拒绝采样：使用 RL 检查点生成额外的推理和非推理任务数据，进一步微调模型。
- 4、全场景强化学习：在最终阶段结合多种奖励信号，提升模型的有效性和安全性。

2.2.1、MLA（多头潜在注意力机制）：显著节省计算资源及内存占用

- MLA从传统的MHA（多头注意力机制）出发，MHA通过并行运行多个Self-Attention层并综合其结果，能够同时捕捉输入序列在不同子空间中的信息，从而增强模型的表达能力。通过将输入的查询、键和价值矩阵分割成多个头，并在每个头中独立计算注意力，再将这些头的输出拼接线性变换，从而实现在不同表示子空间中同时捕获和整合多种交互信息，提升模型的表达能力。
- 处理长序列时MHA会面临计算和内存效率上的局限性，MLA显著降低计算及内存占用问题。MLA的核心思想则是使用低秩分解（LoRA）来近似Key和Value的投影，以在推理期间减少键值缓存（KV cache），显著降低计算和内存占用的复杂度。

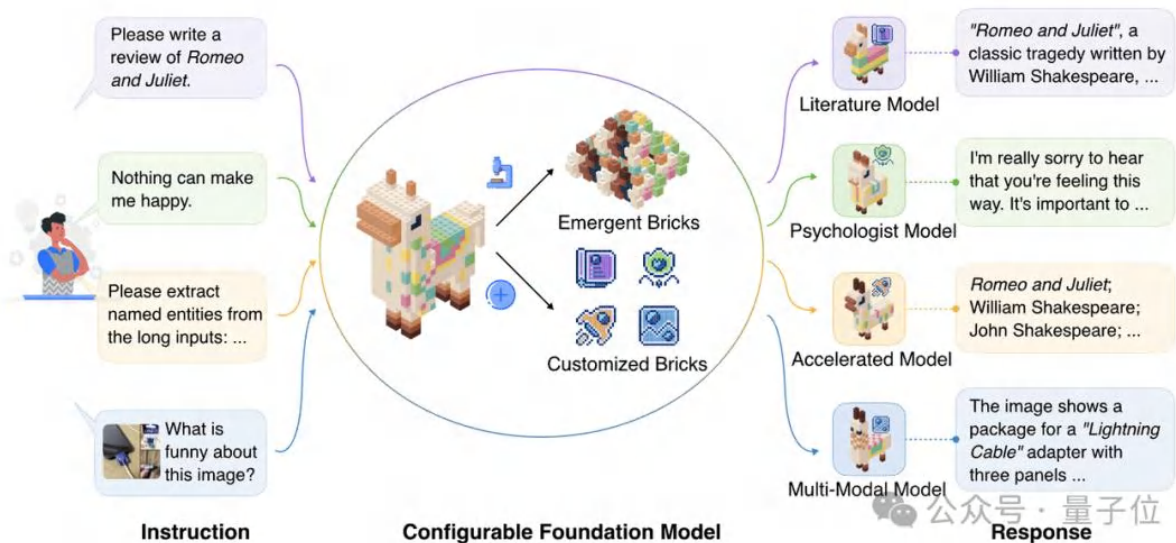
图：DeepSeek V3的创新架构



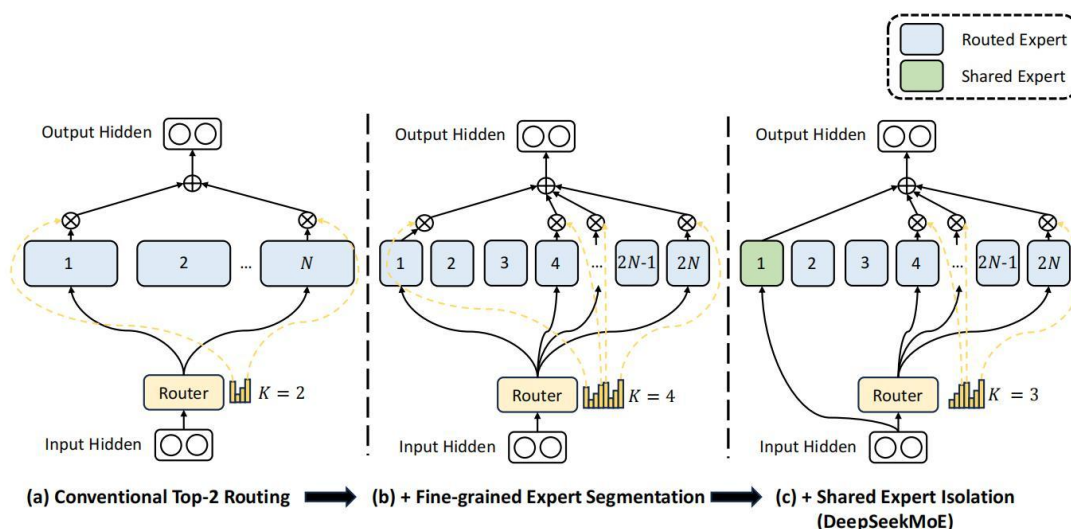
2.2.2、DeepSeekMoE架构以及创新性负载均衡策略

- ❖ **MoE架构**：传统MoE架构的主要优势是利用稀疏激活的性质，将大模型拆解成若干功能模块，每次计算仅激活其中一小部分，而保持其余模块不被使用，从而大大降低了模型的计算与学习成本，能够在同等计算量的情况下产生性能优势。
- ❖ **DeepSeekMoE在传统MoE架构之上，更新了两个主要的策略**：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



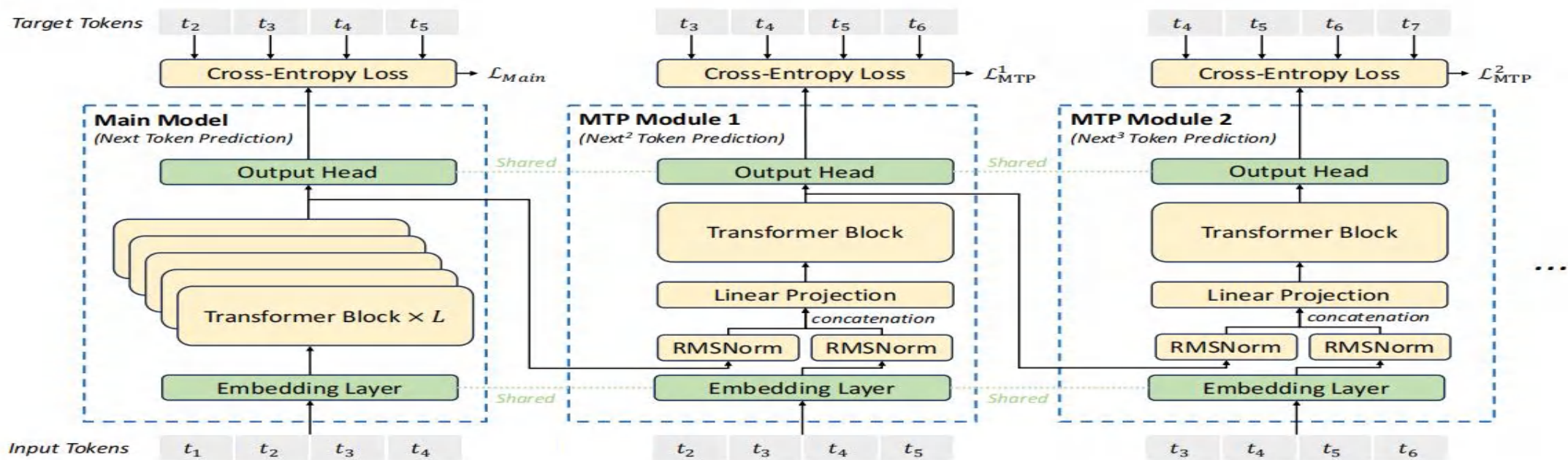
图：DeepSeekMoE对比传统MoE架构



2.2.3、MTP（多token预测）大幅提升模型性能

- MTP显著提升模型性能。**
 - 训练过程：**传统语言模型一次只预测一个token的范式。它就像是让模型从"一字一句"地朗读，进化为"整句整段"地理解和生成。在训练过程中，模型不再局限于预测序列中的下一个token，而是学会同时预测多个连续位置的token。这种并行预测机制不仅提高了训练效率，还让模型能够更好地捕捉token之间的依赖关系。在保持输出质量的同时，模型整体性能提升2-3%。
 - 推理阶段：**MTP的优势更加明显。传统模型生成文本时就像是在"一笔一划"地写字，而MTP则像是"提前打草稿"，可以同时生成多个token。通过创新的推测解码机制，模型能够基于当前上下文同时预测多个可能的token序列。即使某些预测不准确需要回退，整体效率仍然显著提升。这种并行生成机制使推理速度提升了1.8倍，还显著降低了计算开销。

图：MTP架构

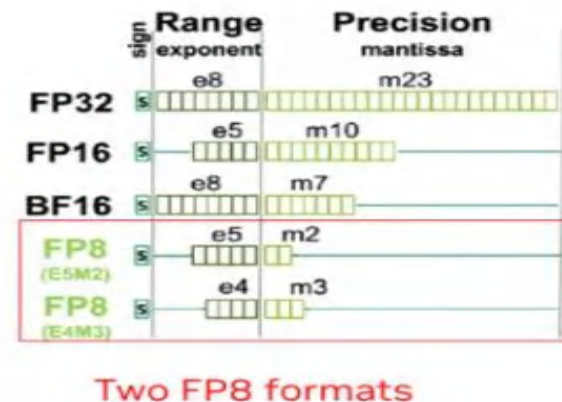


2.2.4、DeepSeek-FP8混合精度训练：实现更高的计算效率

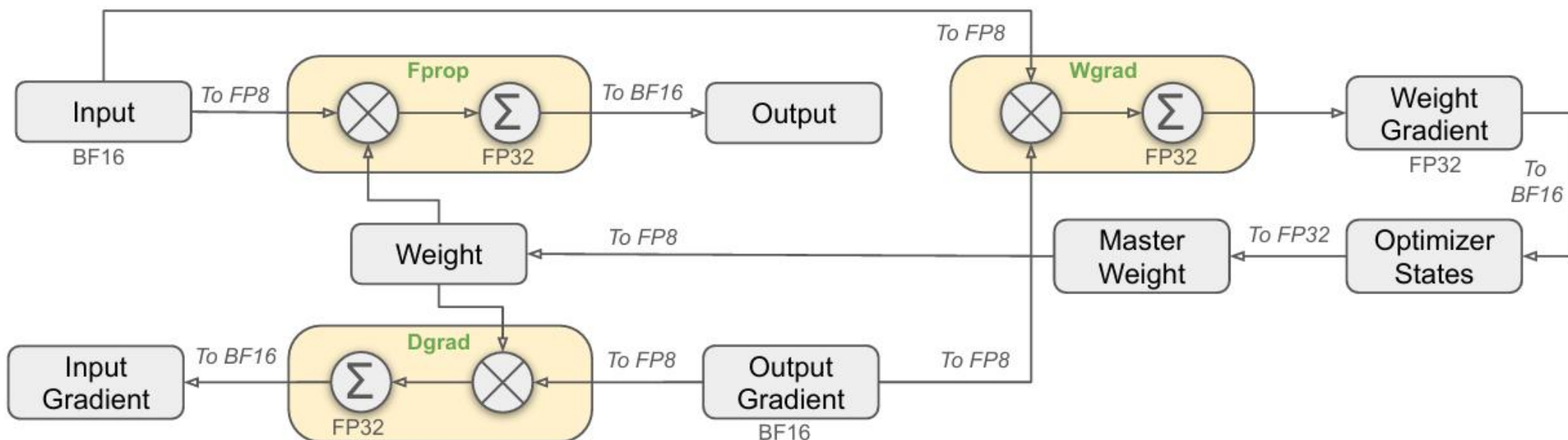
DeepSeek V3采用了FP8混合精度训练框架。在训练过程中，大部分核心计算内核均采用FP8精度实现。例如，在前向传播、激活反向传播和权重反向传播中，输入数据均使用FP8格式，而输出结果则使用BF16或FP32格式。这种设计使得计算速度相较于原始BF16方法提升一倍。

FP8格式是一种低精度的数据格式，具有较小的存储空间和计算开销。通过使用FP8格式，DeepSeek能够在有限的计算资源下，实现更高的计算效率。例如，在处理大规模数据集时，FP8格式可以显著减少显存的占用，从而提高模型的训练速度。

图：多种精度数据类型结构



图：具有 FP8 数据格式的整体混合精度框架



2.2.5、DeepSeek-DualPipe算法：减少流水线气泡，提升GPU利用率

- DeepSeek-V3 采用了一种名为 DualPipe 的创新流水线并行策略。与传统的单向流水线 (如 1F1B) 不同，DualPipe 采用双向流水线设计，即同时从流水线的两端馈送 micro-batch。这种设计可以显著减少流水线气泡 (Pipeline Bubble)，提高 GPU 利用率。
- DualPipe 还将每个 micro-batch 进一步划分为更小的 chunk，并对每个 chunk 的计算和通信进行精细的调度。随后将一个 chunk 划分为 attention、all-to-all dispatch、MLP 和 all-to-all combine 等四个组成部分，并通过精细的调度策略，使得计算和通信可以高度重叠。

图：DualPipe性能优越

Method	Bubble	Parameter	Activation
1F1B	$(PP-1)(F+B)$	1x	PP
ZB1P	$(PP-1)(F+B-2W)$	1x	PP
DualPipe (Ours)	$(\frac{PP}{2}-1)(F+B+B-3W)$	2x	PP+1

DualPipe 在流水线气泡数量和激活内存开销方面均优于 1F1B 和 ZeroBubble 等现有方法

图：DualPipe示意图

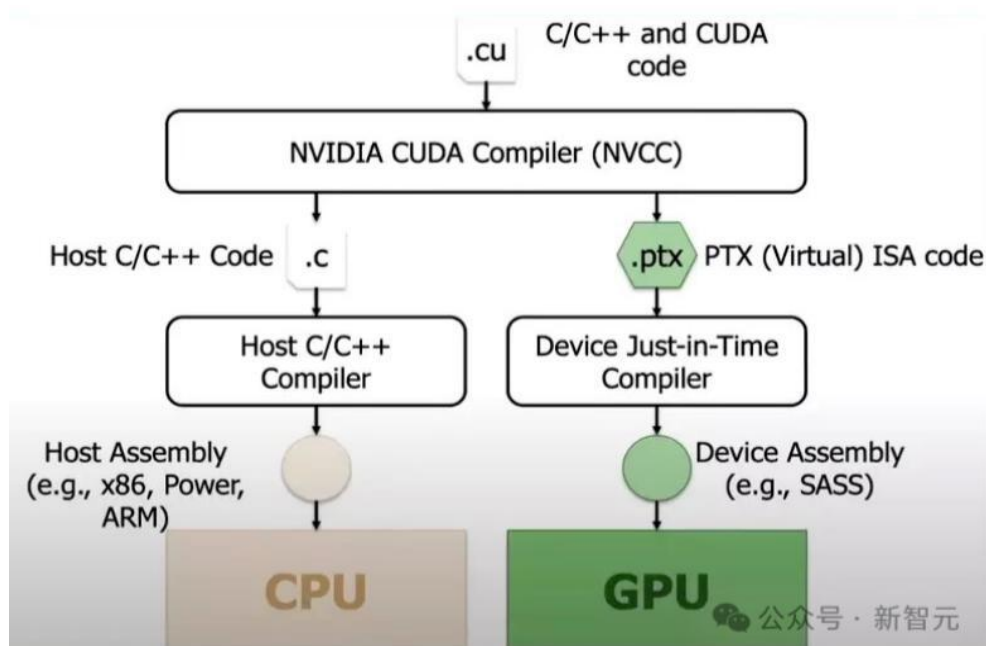


2.2.6、英伟达PTX：位于CUDA与机器代码之间，实现细粒度控制与性能优化

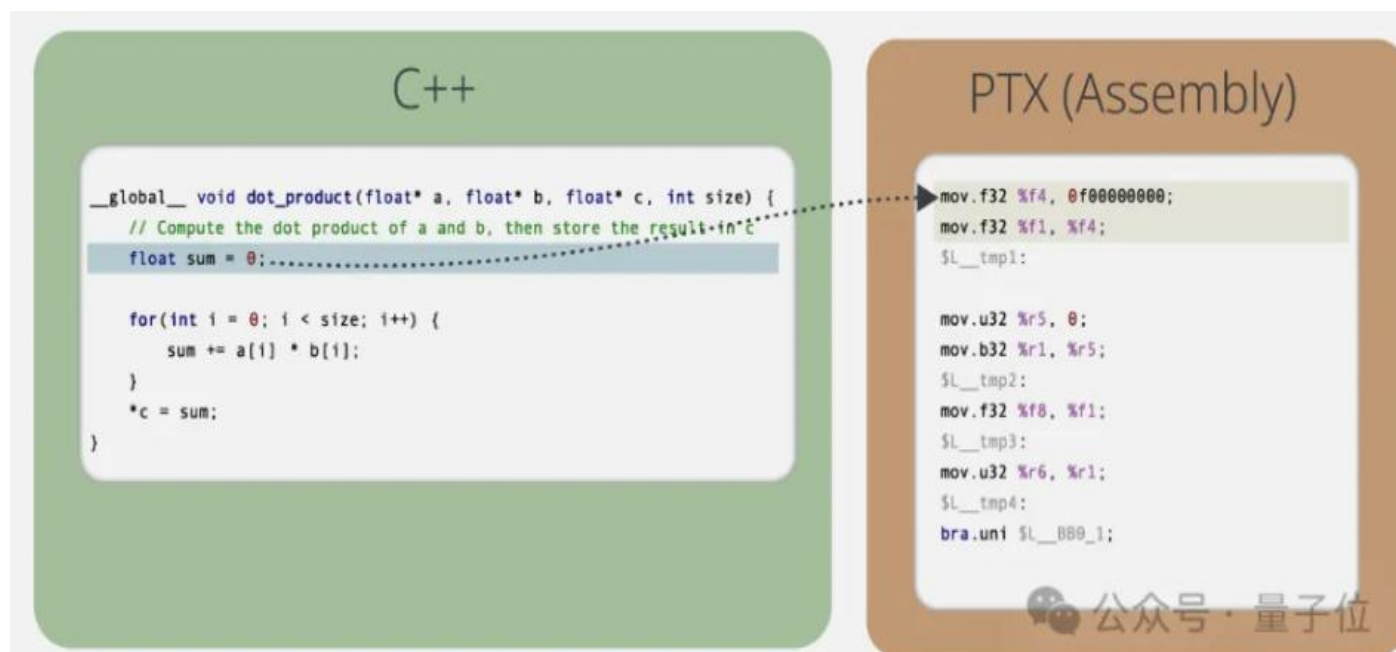
英伟达PTX（并行线程执行）是专门为其GPU设计的中间指令集架构，位于高级GPU编程语言（如CUDA C/C++或其他语言前端）和低级机器代码（流处理汇编或SASS）之间。PTX是一种接近底层的指令集架构，将GPU呈现为数据并行计算设备，因此能够实现寄存器分配、线程/线程束级别调整等细粒度优化，这些是CUDA C/C++等语言无法实现的。

DeepSeek V3采用定制的PTX（并行线程执行）指令并自动调整通信块大小，这大大减少了L2缓存的使用和对其他SM的干扰。PTX允许对GPU硬件进行细粒度控制，这在特定场景下可以带来更好的性能。

图：英伟达PTX是专门为其GPU设计的中间指令集架构



图：C++与PTX代码的区别



2.3、DeepSeek R1 Zero核心创新点——RL（强化学习）替代SFT（有监督微调）

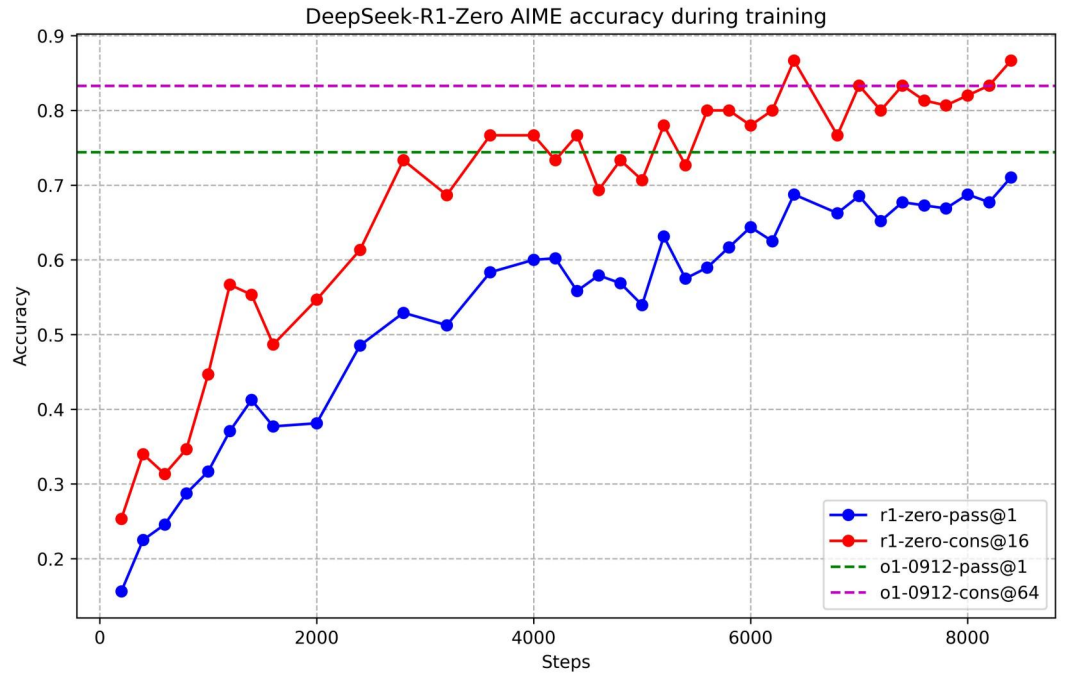


- DeepSeek探索LLM在没有任何监督数据的情况下发力推理能力的潜力，通过纯RL（强化学习）的过程实现自我进化。具体来说，DS使用 DeepSeek-V3-Base 作为基础模型，并使用GRPO（群体相对策略优化）作为RL框架来提高模型在推理中的性能。在训练过程中，DeepSeek-R1-Zero自然而然地出现了许多强大而有趣的推理行为。
- 经过数千次 RL 步骤后，DeepSeek-R1-Zero 在推理基准测试中表现出卓越的性能。例如，AIME 2024 的 pass@1 分数从15.6%增加到 71.0%，在多数投票的情况下，分数进一步提高到86.7%，与OpenAI-o1-0912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

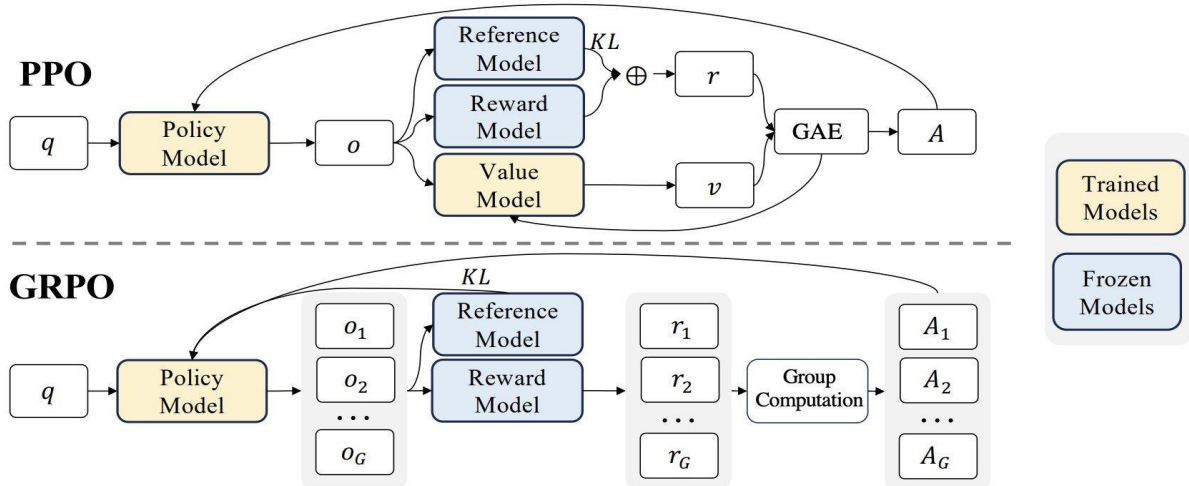


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

图：GRPO核心方法详解

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:

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

2.3.2、强化学习框架下的“顿悟”时刻，模型思维能力跃升的关键

- 模型训练过程中会迎来顿悟时刻。引用DeepSeek相关论文中的表述，GRPO强化学习框架下，不是人类教模型如何解决问题，而仅仅为其提供正确的激励措施，模型会自主开发高级问题解决策略，模型训练过程中会出现“Aha moment”，这表示模型意识到了更好的解决办法并重新评估，为问题分配更多思考时间。
- 这些顿悟时刻是模型思维能力跃升的关键。据DeepSeek的研究，模型的进步并非均匀渐进的。在强化学习过程中，响应长度会出现突然的显著增长，这些“跳跃点”往往伴随着解题策略的质变。这种模式酷似人类在长期思考后的突然顿悟，暗示着某种深层的认知突破。

图：GRPO相对传统PPO强化学习方式对比

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a} - \sqrt{a+x} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a} - \sqrt{a+x} = x$, let's start by squaring both ...

$$(\sqrt{a} - \sqrt{a+x})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a} - \sqrt{a+x} = x$$

First, let's square both sides:

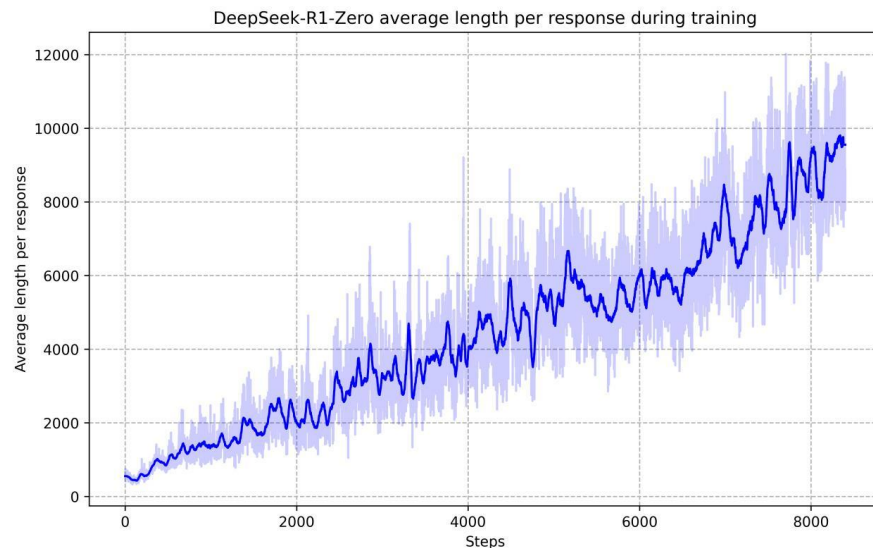
$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

例如模型在解决这个 $\sqrt{a} - \sqrt{a+x} = x$ 的问题时，模型突然停下来说“Wait, wait. Wait. That's an aha moment I can flag here”（等等、等等、这是个值得标记的啊哈时刻），随后重新审视了整个解题过程。这种类似人类顿悟的行为完全是自发产生的，而不是预先设定的。

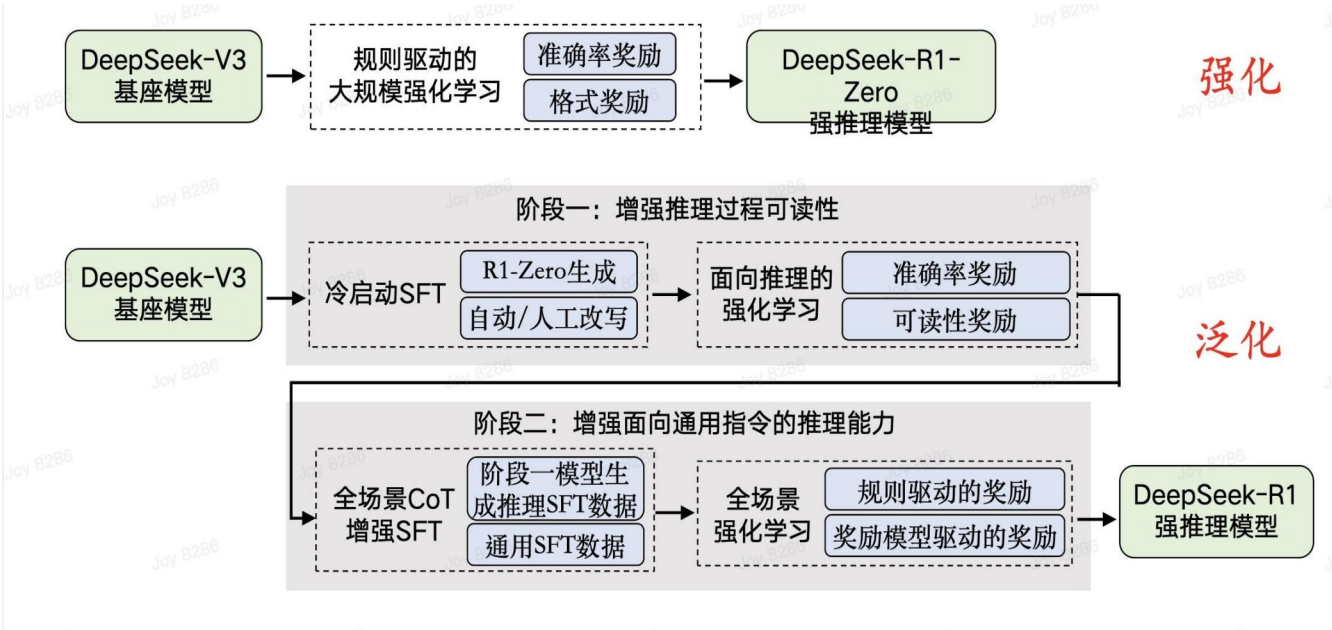
图：强化学习过程中，模型会出现跳跃点，这就是顿悟时刻



2.4、DeepSeek R1：高质量冷启动数据+多阶段训练，将强推理能力泛化

- 纯强化学习后出来的DeepSeek R1 zero存在可读性差以及语言混乱等问题，主要因其全通过奖惩信号来优化其行为，没有任何人类示范的"标准答案"作为参考，因此DeepSeek团队使用冷启动+多阶段训练推出DeepSeek R1模型。
- 具体训练步骤：
1) 高质量冷启动数据：与DeepSeek R1 zero同理，以DeepSeek v3 base作为强化学习的起点，但为了克服可读性差的问题，选择可读性更强的cot（长思维链）数据作为冷启动数据，包括以可读格式收集DeepSeek-R1 Zero输出，并通过人工注释者进行后处理来提炼结果。
2) 面向推理的强化学习，这与DeepSeek R1 zero的强化学习过程相同，但是在RL期间引入语言一致性奖励，虽然语言对齐可能会造成一定的性能损失，但是提高了可读性。
3) 抑制采样和监督微调，拒绝采用指模型训练过程中生成的一些不符合特定标准或质量要求的样本数据进行舍弃，同时选取了v3的SFT数据集一部分作为微调数据。
4) 全场景强化学习，属于一个二级强化学习阶段，目的是与人类偏好保持一致。

图：DeepSeek R1 模型训练过程



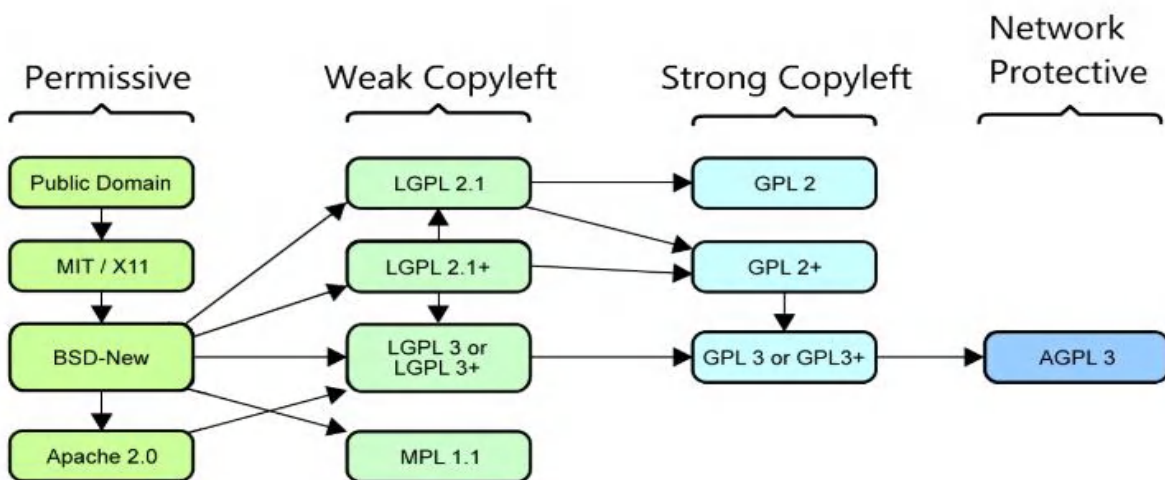
图：DeepSeek R1与其他模型的性能对比

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	-

2.5、开源大模型：打破OpenAI等闭源模型生态，提升世界对中国AI大模型认知

- ✓ 开源即代码层面开源，可以调用与进行二次开发。开源免费调用有助于先行占据市场份额，成为规则制定者，率先拓展生态粘性。如，谷歌将安卓开源，获得了全球80%的移动手机端市场份额，同时也覆盖电视、汽车等使用场景。
- ✓ 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，即使用者针对自身需求蒸馏，或用自有数据训练，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。

图：开源许可证协议标准



图：DeepSeekMoE对比传统MoE架构

DeepSeek-R1 已发布并开源，性能对标 OpenAI o1 正式版，在网页端、APP 和 API 全面上线，点击查看详情。

deepseek

探索未至之境

开始对话

免费与 DeepSeek-V3 对话
使用全新旗舰模型

获取手机 App

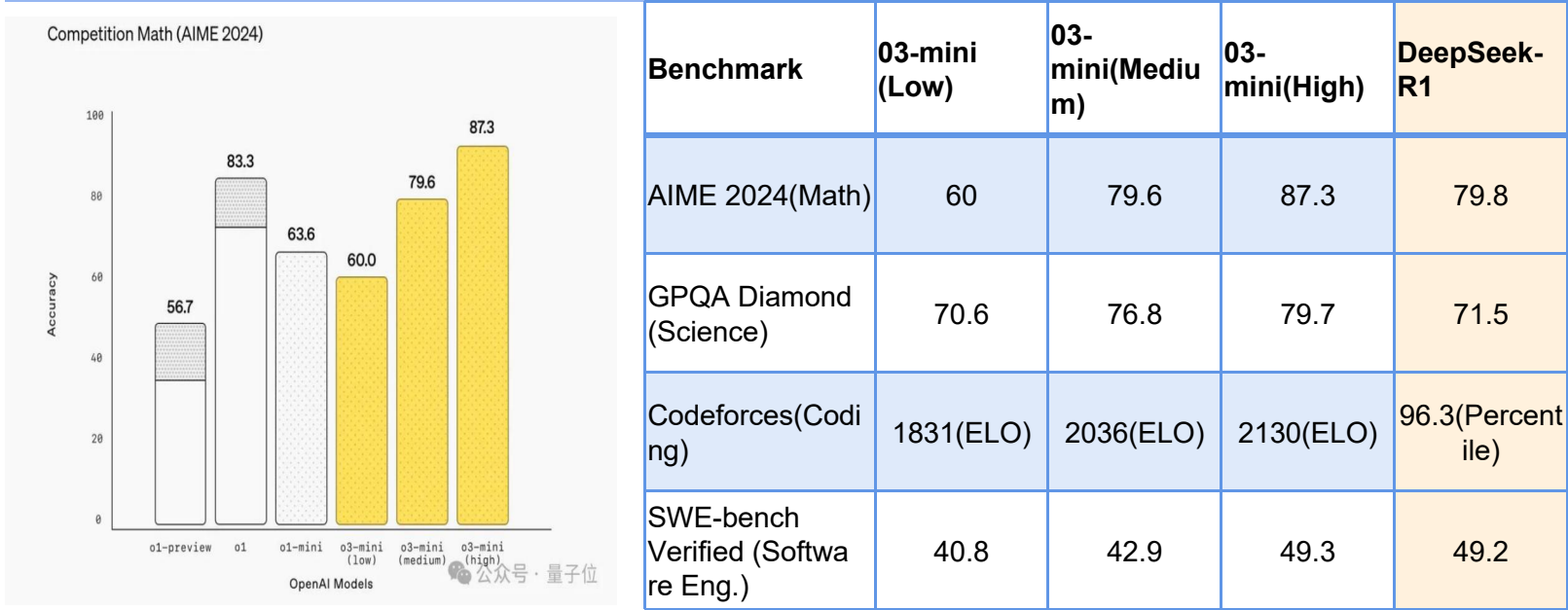
DeepSeek 官方推出的免费 AI 助手
搜索写作阅读解题翻译工具

三、DeepSeek对AI 应用的影响？

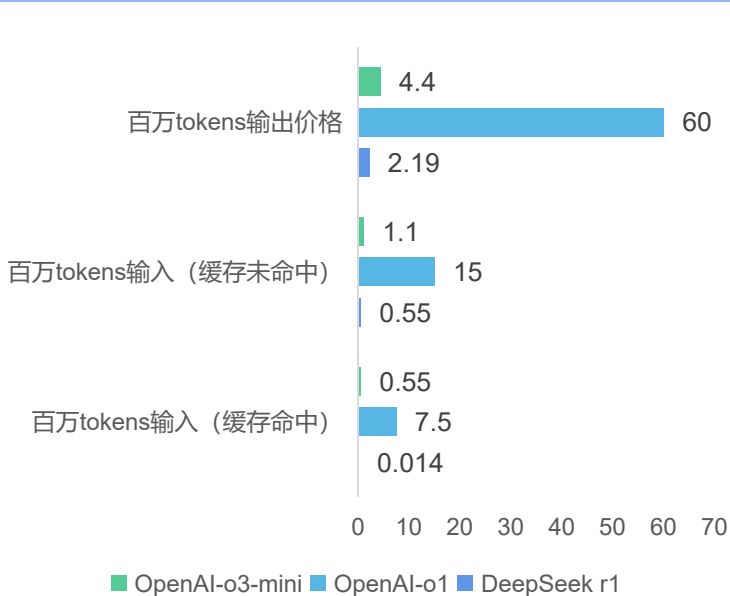
3.1、DeepSeek打开低成本推理模型边界，加速AI应用布局进程

- 核心观点：DeepSeek在推动降本、强推理三大层面驱动下，有望加速AI应用普及度迎来跨越式提升。
- OpenAI上线性价比模型o3-mini，加速低成本推理模型边界。2025年2月1日，OpenAI深夜上线o3-mini系列推理模型，其也是OpenAI系列推理模型中最具性价比的模型。性能方面，o3-mini在数学、编程、科学等领域表现优异，以数学能力为例，o3-mini（low）达到了与o1-mini相当的水平；o3-mini（medium）能力媲美满血版o1；o3-mini（high）表现超越o1系列一众模型。对比DeepSeek-R1在数学能力、编程能力上的测试结果，DeepSeek R1处于OpenAI o3-mini（medium）水平。
- DeepSeek价格优势仍大幅领先于OpenAI系列推理模型。DeepSeek定价为百万tokens输入0.014美元（缓存命中，未命中则0.55美元），百万tokens输出价格2.19美元；o3-mini百万tokens输入价格0.55美元（缓存命中，未命中则1.1美元），百万tokens输出价格为4.4美元。

图：DeepSeek和OpenAI能力对比



图：DeepSeek和OpenAI推理模型定价对比

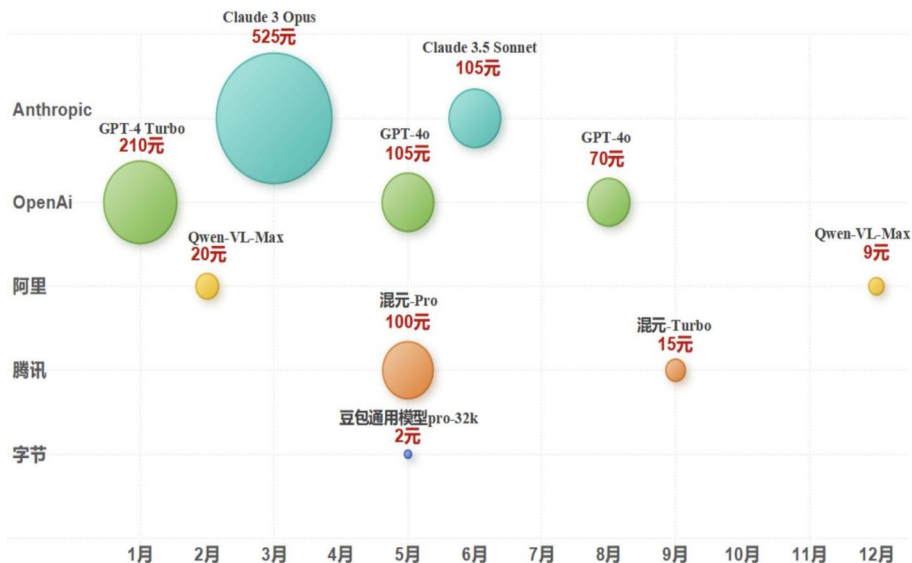


3.1.1、模型成本下降+性能第一梯队+开源，国内AI应用商业模式有望加速跑通

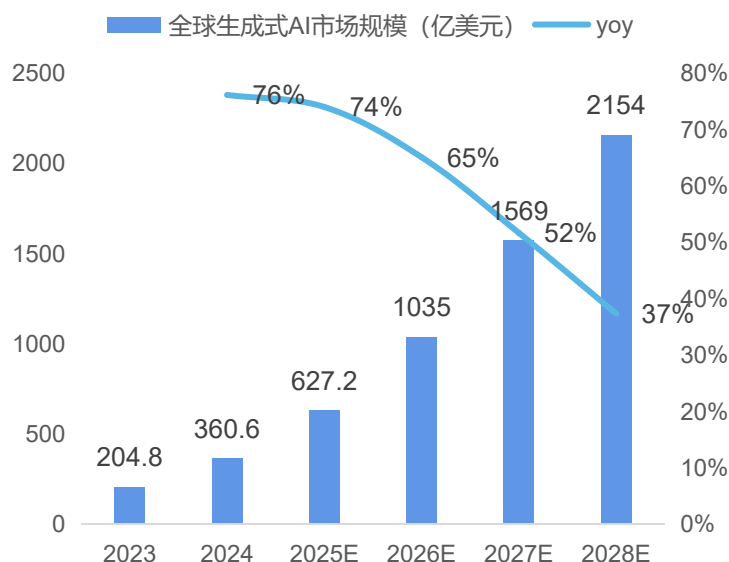
我们认为DeekSeek或推动AI投资回报率加速提升，AI应用商业模式加速跑通。据中国工业互联网研究院数据，2024年以字节火山引擎、阿里云、百度云为代表的云厂商掀起了大模型价格战，降价幅度普遍达到90%以上。海外以OpenAI为例，5月发布GPT-4o，模型性能升级且价格较GPT-4-Turbo下降50%;8月上线GPT-4o新版本，更强更便宜，但输出价格节省33%。国内以阿里为例，12月31日阿里云宣布2024年度第三轮大模型降价，通义千问视觉理解模型全线降价超80%。

全球及中国AI应用市场规模加速提升。据IDC数据，全球生成式AI市场规模在2024年达到360.6亿美元，同比+76%，预计在2028年达到2154亿美元；中国AI软件市场规模在2024年达到5.7亿美元，预计2028年达到35.4亿美元。

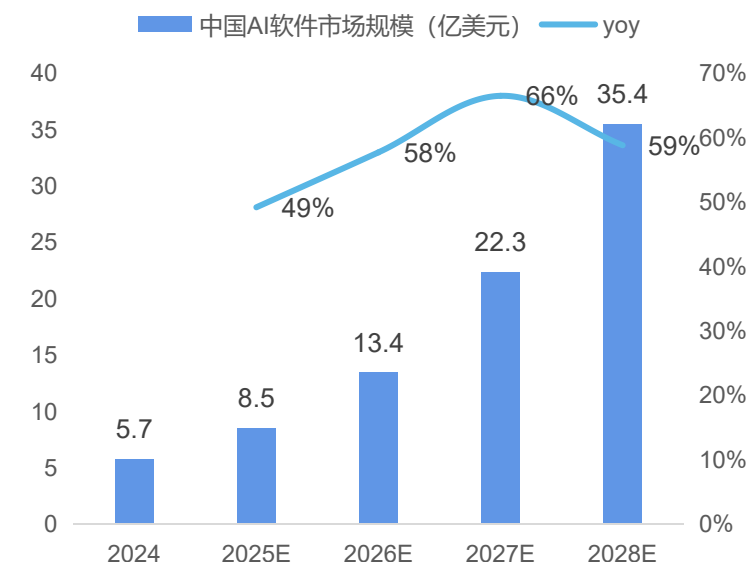
图：大模型降本趋势明确



图：全球生成式AI市场规模



图：中国AI软件市场规模



国海证券
SEALAND SECURITIES

🔴 **DeepSeek最终目标：AGI。**传统的AI训练方法可能一直在重复于让AI模仿人类的思维方式。通过纯粹的强化学习，AI系统似乎能够发展出更原生的问题解决能力，而不是被限制在预设的解决方案框架内。虽然R1-Zero在输出可读性上存在明显缺陷，但这个"缺陷"本身可能恰恰印证了其思维方式的独特性。就像一个天才儿童发明了自己的解题方法，却难以用常规语言解释一样。这提示我们：真正的通用人工智能可能需要完全不同于人类的认知方式。

图：我国AGI产业图谱

The diagram illustrates the AGI industry landscape in China, categorized into four main application scenarios:

- 零售应用场景 (Retail Application Scenarios):**
 - AI商拍: ArcSoft 虹软, ZMO.AI, meitu, PhotoMagic, Nolibox, WeShop唯象, 街匠科技.
 - 营销物料生成: Alibaba, 京东, meitu, Nolibox, 奇智, 佐糖, 时趣.
 - 门店管理: 多点DMALL, QMQI企迈科技, 实在智能, 沃丰科技.
 - 智能投放: 阿里妈妈, 百度营销, 磁力引擎, 京准通, 巨量引擎, 腾讯广告.
 - 平台商家助手: Alibaba, Bai百度, 京东.
 - 智能客服: Alibaba, Bai百度, 沃丰科技, BetterYeah, 光云, 智齿科技, 晓多科技.
 - 数字人导购/主播: Tabcut 特看, 百度智能云曦灵, XMOV, 拟仁智能, 风平智能, 相芯科技, 拟仁智能, 晓多科技.
- 金融应用场景 (Financial Application Scenarios):**
 - 传统金融机构: ICBC, 中国工商银行, 中国农业银行, 中国银行, 中国建设银行, 交通银行, 广发银行, 阳光保险集团, 中信银行.
 - 消费金融: 马上消费金融, 招联.
 - 金融科技公司: 蚂蚁集团, 众安科技, CIFU, 网商银行, 东方财富.
 - 互联网金融服务机构: 蚂蚁集团, 众安科技, CIFU, 网商银行, 东方财富.
 - 科技公司及互联网金融数据服务商: 商汤, HUAWEI, 科大讯飞, 功夫量化, 智谱·AI, 科大讯飞, MEGV旷视, 百度智能云, 网易数帆, 易道博识, 海致集团, 易道博识, 易道博识.
- 企业服务应用场景 (Enterprise Services Application Scenarios):**
 - 协同办公: Bai百度, 钉钉, 飞书, 企业微信, 钉钉, 钉钉, 钉钉.
 - 辅助编程: amazon, 百度智能云, 华为云, KUNLUN, 钉钉, 钉钉, 钉钉.
 - 自动化流程: SAIT, CYCLOPE, K, LAIYE, 实在智能, UiPath.
 - 数据分析: Fabarta, 百度智能云, 网易数帆, SMARTBI.
 - 企业资源规划: 律己AI, Beisen北森, Kingdee金蝶, OKKI, Tencent腾讯, 用友.
- 教育应用场景 (Education Application Scenarios):**
 - 学科辅导: 学而思, 高途, 三乐, 作业帮, 猿力科技, 有道 youdao.
 - 职业教育: 极客时间, 智云智训, 知学云, 全美在线 ATA, 中公教育.
 - 教育硬件: Bai百度, 科大讯飞, 希沃, 希沃, 希沃.
 - 智慧教研/校园: 科大讯飞, 希沃, 希沃, 希沃.

营销应用场景

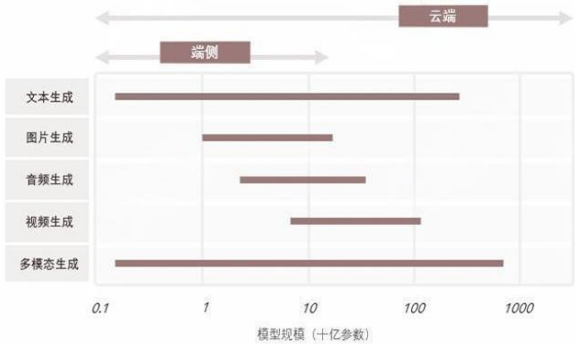
- 内容生产及创意**
 - 明略科技[®] (Minglue Technology)
 - SoundAI 声智
 - meitu
 - 倒映有声
 - 阿里云
 - FancyTech
 - 相芯科技
 - 元分身
 - 街述科技
 - Nolibox
 - VCG
 - 百度营销
 - 火山引擎
 - 360 智绘
- 广告投放及运营**
 - 科大讯飞 (iFLYTEK)
 - eclicktech 易点天下
 - 商汤 (sensetime)
 - 中电金信 (GienTech)
 - 百度营销
 - BlueFocus
 - 归一智能
 - LEODigital 利欧数字
 - 华院计算 (www.lin21.com)
- 销售及客服**
 - 阿里云
 - 京东云
 - 容联云 (CLIPOPEN)
 - 卫瓠科技
 - 网易云商
 - 腾讯云
 - Neocrm 销售易
 - 玄武云 (XWUWULCLOUD)
 - 螳螂科技
 - 亿量科技
- 策略与研究**
 - 归一智能
 - e-brands 云积天舒
 - 明略科技[®] (Minglue Technology)
 - 百度营销
 - Convertlab
 - 腾讯企点

3.2、DeepSeek R1蒸馏赋予小模型高性能，端侧AI迎来奇点时刻

☑️蒸馏法具有强大的潜力，端侧小模型迎来发展契机。如下表所示，只需提取 DeepSeek-R1 的输出即可使高效的DeepSeekR1-7B全面优于GPT-4o-0513等非推理模型，DeepSeek-R1-14B在所有评估指标上都超过了QwQ-32BPreview，而 DeepSeek-R1-32B和DeepSeek-R1-70B在大多数基准测试中明显超过了 o1-mini。此外，我们发现将 RL 应用于这些蒸馏模型会产生显著的进一步收益。我们认为这值得进一步探索，因此在这里只提供简单的 SFT 蒸馏模型的结果。

☑️DeepSeek产品协议明确可“模型蒸馏”。DeepSeek决定支持用户进行“模型蒸馏”，已更新线上产品的用户协议，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：端侧与云端部署AI的规模区别



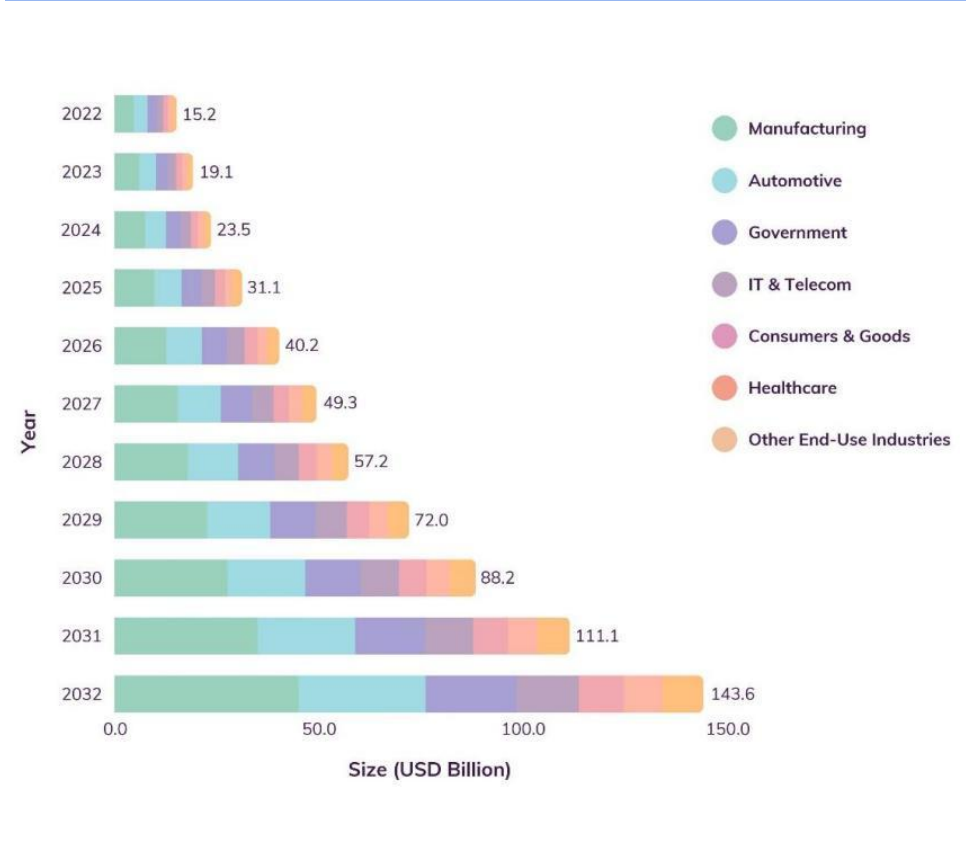
图：DeepSeek R1蒸馏小模型性能对比

	<u>AIME2024pass</u> <u>@1</u>	<u>AIME2024cons</u> <u>@64</u>	<u>MATH-</u> <u>500pass@1</u>	<u>GPQADiamond</u> <u>pass@1</u>	<u>LiveCodeBenc</u> <u>hpass@1</u>	<u>CodeForcesratin</u> <u>g</u>
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16	26.7	78.3	65	38.9	717
o1-mini	63.6	80	90	60	53.8	1820
QwQ-32B	44	60	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80	89.1	49	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70	86.7	94.5	65.2	57.5	1633

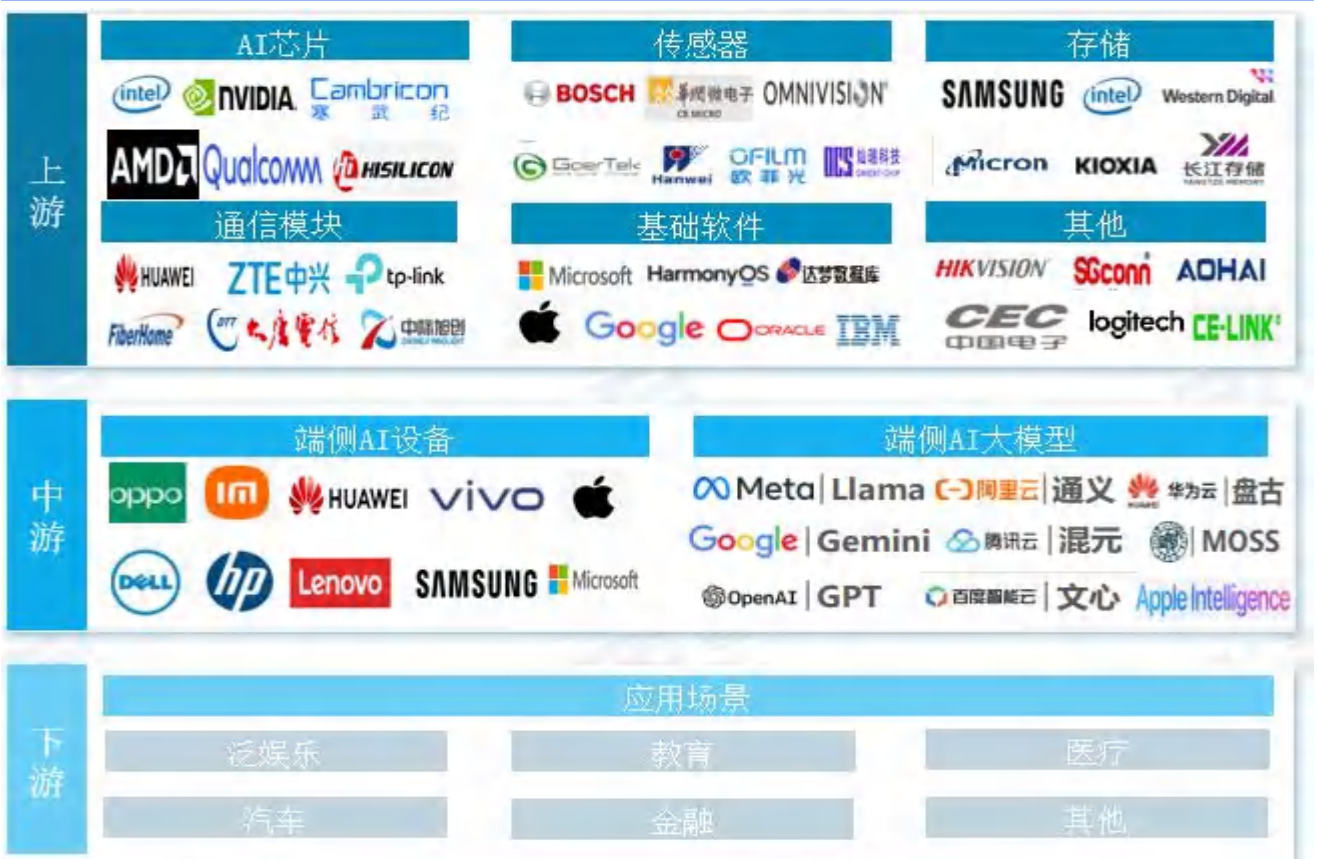
3.2、DeepSeek R1蒸馏赋予小模型高性能，端侧AI迎来奇点时刻

- 全球端侧AI市场规模预计从2022 年的152亿美元增长到2032年的1436亿美元。这一近十倍的增长不仅反映了市场对边缘 AI 解决方案的迫切需求，也预示着在制造、汽车、消费品等多个行业中，边缘 AI 技术将发挥越来越重要的作用。
- 在资源受限的设备上部署性能强大的模型，必须面对内存与计算能力的双重挑战，自2023年起，随着参数量低于 10B 的模型系列如 Meta 的 LLaMA、Microsoft 的 Phi 系列等的涌现，LLMs 在边缘设备上运行的可行性逐步明朗。

图：全球端侧AI市场规模



图：端侧AI产业链图谱



四、DeepSeek对算力影响？

4.1、DeepSeek V3训练中GPU成本558万美元，对比海外成本降低

- DeepSeek V3模型训练成本达278.8万H800小时，共花费557.6万美元。对比OpenAI、Anthropic、LlaMA3等模型，DeepSeek V3单次训练成本显著降低，主要系DeepSeek公司通过优化模型结构、模型训练方法、针对性GPU优化等部分，提升了模型训练过程中的算力使用效率。

表：DeepSeek V3训练成本（假设H800租赁价格为 2 美元/每GPU小时）

训练成本	预训练	上下文扩展	后训练	总计
H800 GPU小时 (万小时)	266.4	11.9	0.5	278.8
美元 (万元)	532.8	23.8	10	557.6

图：DeepSeek V3节省训练成本的方法，包括调整模型结构、训练方法、GPU优化等

模型结构 Architecture	模型训练方法 Pre-Train	针对性GPU优化
专家模型 MOE+ 多头潜在自注意力 MLA	Dual Pipe	低精度FP8训练
用于负载均衡的辅助无损策略	All To ALL 通信内核 IB+NVLlink	PTX语言
多标记预测 (MTP)	无张量并行 TP	带宽限制

4.2 、DeepSeek或有约5万Hopper GPU，训练总成本或较高

- 据Semianalysis，DeepSeek大致拥有10000张H800 GPU芯片、10000张H100 GPU芯片以及大量H20 GPU芯片，用于模型训练/推理、研究等任务。其估计，DeepSeek的总服务器资本支出（CapEx）约为13亿美元（约90亿元人民币），其中仅集群运营成本就高达7.15亿美元。
- DeepSeek V3论文中557.6万美元成本，仅为预训练中消耗的GPU计算成本，但模型完整训练成本包括研发、数据清洗、人员薪资、硬件总拥有成本TCO（服务器、电力、冷却系统、数据中心维护）等，会带来训练总成本体量更高。作为对比，Anthropic训练Claude 3.5 Sonnet的成本就高达数千万美元。

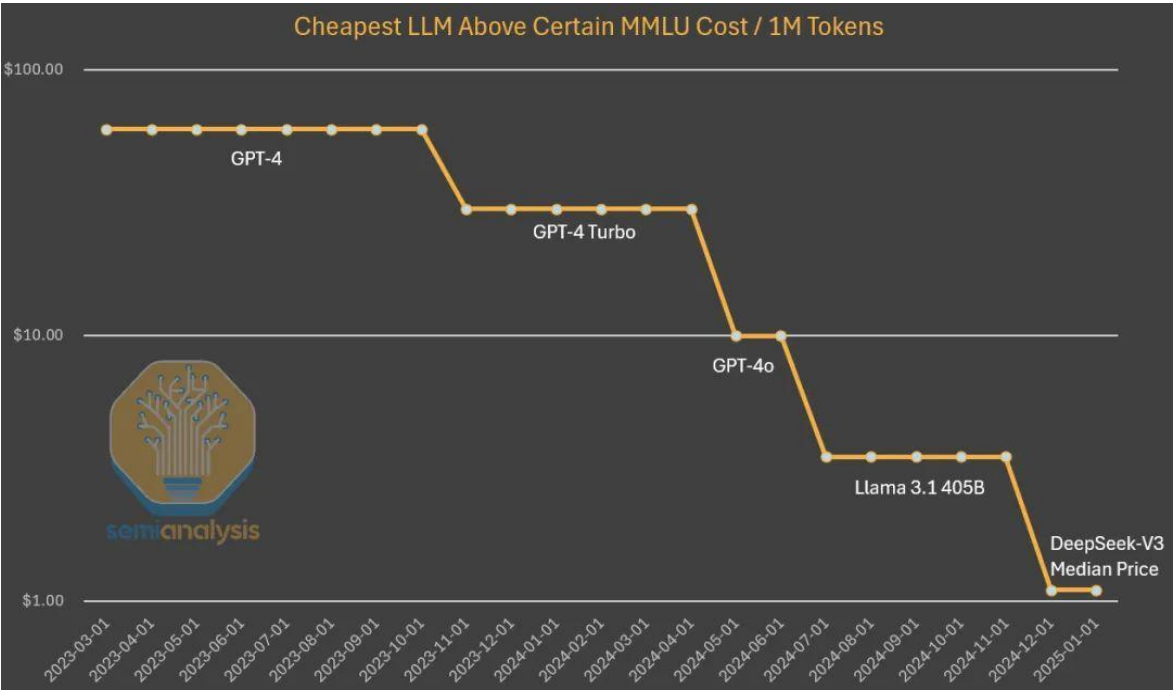
表：DeepSeek AI TCO（总拥有成本）

Chip	Unit	A100	H20	H800	H100	Total
Years	#	4	4	4	4	
# of GPUs	#	10,000	30,000	10,000	10,000	60,000
NVDA \$ ASP	\$	13,500	12,500	20,000	23,000	46,000
Server CapEx / GPU	\$	23,716	24,228	31,728	34,728	79,672
Total Server CapEx	\$m	237	727	317	347	1,281
Cost to Operation	\$m	157	387	170	230	715
Total TCO (4y Ownership)	\$m / hr	395	1,114	487	577	1,996

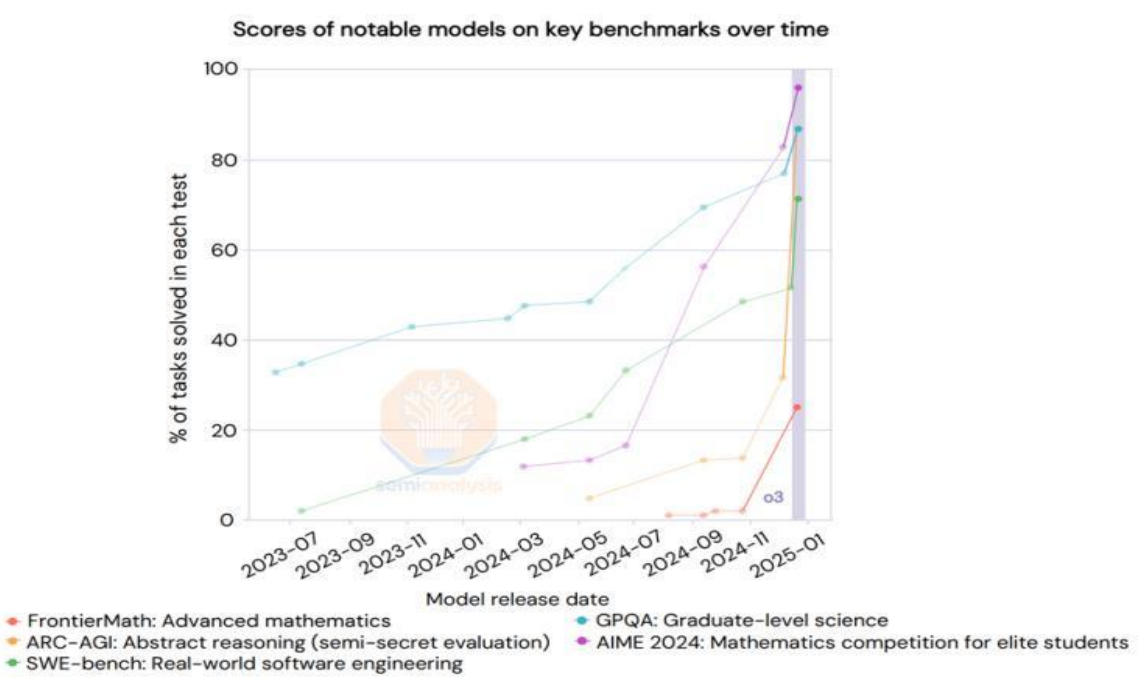
4.2 、Jevons�论：AI计算提效引总需求提升，NV H100租赁价格上涨

- AI的演进路径中，推理成本不断下降，计算效率持续提高是长期趋势。例如：算法进步的速度约为每年4倍，即每年达到相同能力所需的计算量减少到四分之一；Anthropic CEO Dario甚至认为，这一进步速度可能达到10倍。
- Jevons 悖论**：技术进步提高了资源使用效率，效率提高降低了资源使用成本，成本下降刺激了资源需求的增长，需求增长可能超过效率提升带来的节约，最终导致资源总消耗增加。
- 短期训练侧算力需求或受影响，但DeepSeek推理价格下降吸引更多用户调用模型，带来英伟达H100 GPU的租赁价格提升，故表明算力需求量短期仍呈提升趋势，中长期推理算力需求有望持续增长。

图：大模型成本持续下降，效率提升



图：关键模型的测试情况



4.3、推理化：推理算力需求占比提升，GenAI云厂商有望受益

- **DeepSeek降低推理成本，引算力需求结构变化。**模型算法改进提升了训练算力使用效率、降低了训练成本，促进了模型商品化和更便宜的推理。据Semianalysis，DeepSeek推理服务可能以成本价运营，以此抢占市场份额，还在推理端优化英伟达H20 GPU的使用（H20内存与带宽容量高于H100，推理效率更具优势）。
- **推理占比持续提升。**更低的推理成本有望提升下游应用与端侧对大模型推理使用需求，推理算力需求占比有望增长。2024H1，用于推理的人工智能芯片市占率为61%。据IDC，预期2023-2027年，推理AI服务器工作负载占比从41%提升至73%左右。
- **集合多种模型的云服务厂商有望受益。**无论是开源还是闭源模型，计算资源都很重要，如果云厂商基于计算资源打造上层服务或产品，那么计算资源的价值就有可能提升，这意味着更多的Capex流向硬件领域，软件也有望受益。

表：DeepSeek V3性能优越，推理价格较低

Model	Price / 1M Input Tokens	Price / 1M Output Tokens	MMLU (Pass@1)	SWE Verified (Resolved)	AIME 2024	MATH - 500
Claude-3.5-Sonnet- 1022	\$3.00	\$15.00	88.3	50.8	16	78.3
GPT - 4o - 0513	\$2.50	\$10.00	87.2	38.8	9.3	74.6
DeepSeek - V3 (TogetherAI)	\$1.25	\$1.25	88.5	42.0	39.2	90.2
DeepSeek - V3 Median Provider	\$0.90	\$1.10				
DeepSeek - V3 (Normal Price)	\$0.27	\$1.10				
DeepSeek - V3 (Discount Price)	\$0.14	\$0.28				
Gemini 1.5 Pro	\$1.25	\$5.00	86		20	88
GPT - 4o - mini	\$0.15	\$0.60	82	33.2	6.7	79
Llama 3.1 405B	\$3.50	\$3.50	88.6	24.5	23.3	73.8
Llama 3.2 70B	\$0.59	\$0.73	86		20	64

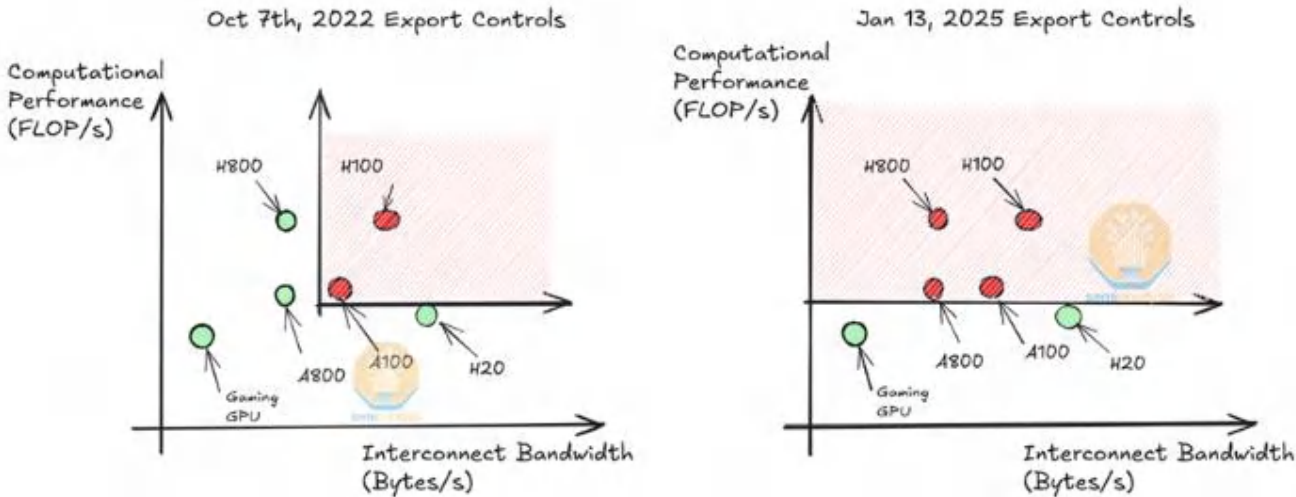
图：2024H1，中国 Top5 GenAI IaaS服务厂商市场份额



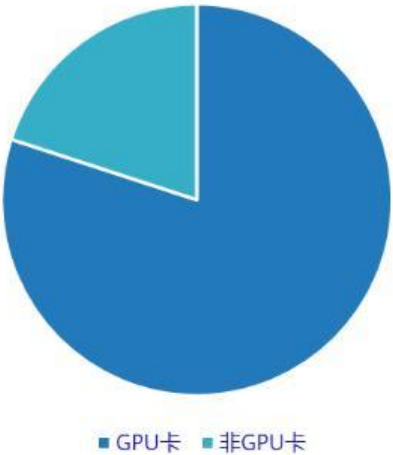
4.3.1、国产化：推理化+中美博弈加剧，国产AI芯片需求有望提升

- 模型推理对大型集群要求弱于训练，这与目前国产算力单卡实力较强、互联能力不足的情况匹配，并考虑到目前中美半导体博弈加剧，DeepSeek积极适配昇腾、海光等国产芯片，国产化推理算力需求有望持续增长。

图：美国限制高端NV GPU出口中国



图：2024H1，中国人工智能芯片市场份额



图：DeepSeek R1&V3推理服务适配昇腾云

首发！硅基流动×华为云联合推出基于昇腾云的DeepSeek R1&V3推理服务！

华为云 2025年02月01日 12:58 广东



DeepSeek-R1开源后引发全球用户和开发者关注。经过硅基流动和华为云团队连日攻坚，现在，双方联合首发并上线基于华为云昇腾云服务的DeepSeekR1/V3推理服务。

图：DeepSeek R1&V3推理服务适配海光DCU

官宣：DeepSeek V3和R1模型完成海光DCU适配并正式上线

光合组织 2025年02月02日 20:34 北京

近日，海光信息技术团队成功完成DeepSeek V3和R1模型与海光DCU（深度计算单元）的国产化适配，并正式上线！

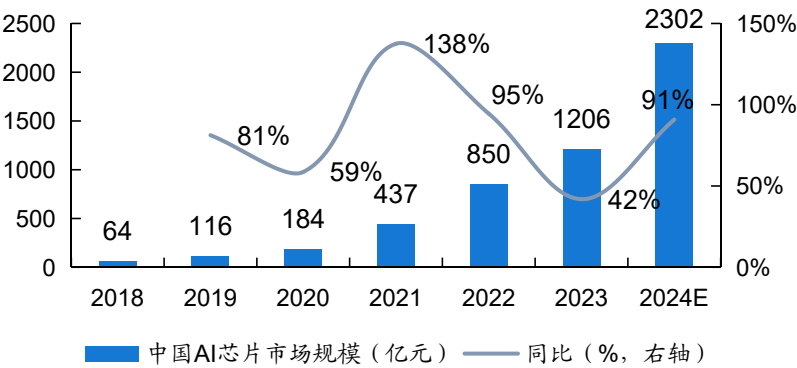
用户现可通过“光合开发者社区”中的“光源”板块访问并下载相关模型，或直接登录[www.sourcefind.cn]搜索“DeepSeek”，即可基于DCU平台快速部署和使用相关模型。



4.3.1 、国产化：国产AI芯片硬件性能提升，市占率持续提升

- 2024H1，全国AI芯片出货中，国产化比例达20%。2024H1，中国加速芯片的市场规模达超过90万张。GPU卡占据80%的市场份额；中国本土人工智能芯片品牌出货量已接近20万张，约占整个市场份额的20%。在加速卡入口受限之后，由于数质化转型大趋势对于算力的持续需求，中国本土品牌加速卡持续优化硬件能力，市场份额存在一定程度的增长。

图：2018-2024年中国AI芯片市场规模预测



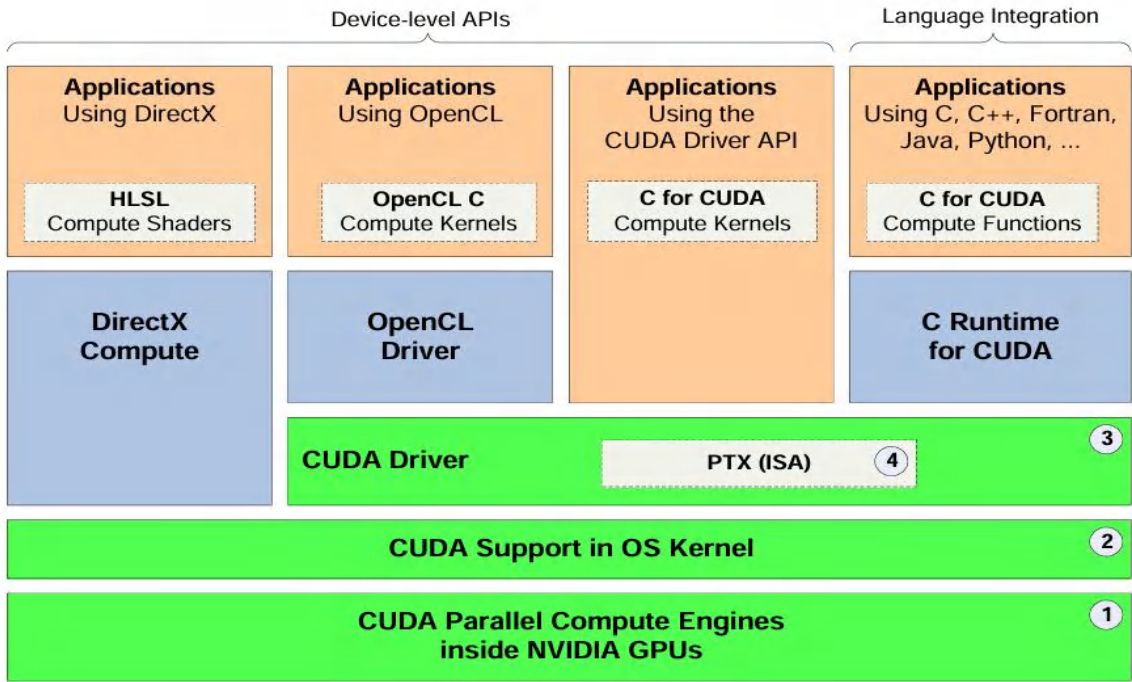
图：国内外主流人工智能芯片性能对比情况

	华为	华为	NVIDIA	NVIDIA	海光	寒武纪	壁仞科技	AMD
型号	昇腾310	昇腾910	A100 80GB PCIe	H100 SXM	深算一号	MLU370-X4	壁砺 104P	MI100
GPU架构	HUAWEI Da Vinci	HUAWEI Da Vinci	Ampere	Hopper	GPGPU	Cambricon LUarch03	壁立仞架构	CDNA
峰值 INT8 计算性能	16 TOPS	640 TOPS	624 TOPS	3958 TOPS	-	256 TOPS	1024 TOPS	184.6TOPS
峰值半精度 (FP16) 性能	8 TOPS	320 TFLOPS	312 TFLOPS	1979 teraFLOPS	-	96 TFLOPS	-	184.6 TFLOPs
峰值双精度 (FP64) 性能	-	-	19.5 TFLOPS	34 teraFLOPS	-	-	-	11.5 TFLOPs
显存容量	-	-	80GB HBM2	80GB	32GB HBM2	24GB LPDDR5	32GB HBM2e	32GB HBM2
最大功耗	8W	310W	300 W	700 W	350W	75 W	300 W	300 W
工艺制程	12nm	7nm	7nm	4nm	7nm	7nm	7nm	7nm
发布时间	2018Q4	2019Q3	2020Q2	2022Q2	2021H2	2021Q4	2022Q3	2020Q4

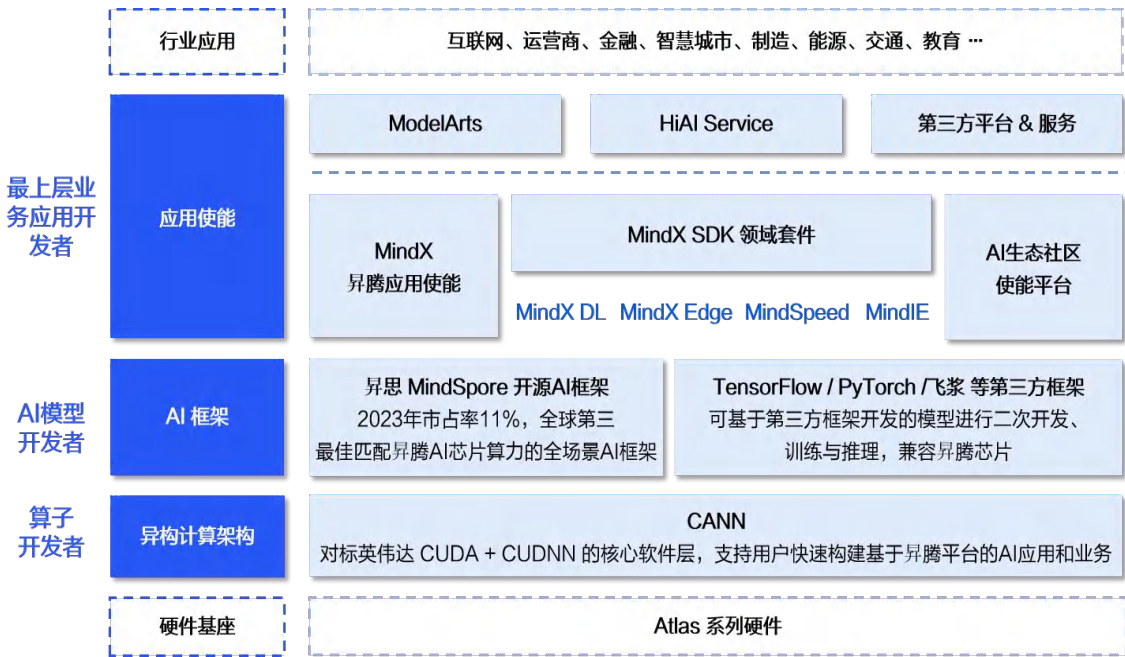
4.3.1、国产化：软件调用能力提升，国产AI芯片生态蓬勃发展

- 使用GPU过程中，通常需经过多个层级才能调用到底层硬件。从上到下依次是高层编程语言（如 Python、C++）、硬件接口（如 CUDA、OpenCL 等）、驱动程序，最后才是底层硬件。在这个过程中，CUDA 作为一个相对高层的接口，为用户提供编程接口，而 PTX 则隐藏在驱动背后。
- DeepSeek-V3模型在多节点通信时绕过了 CUDA 直接使用 PTX（Parallel Thread Execution），有望实现以算法的方式来高效利用硬件层面的加速。PTX 与底层硬件直接交互，编写和调用 PTX 代码能更精确地控制底层硬件，实现更高效的计算。
- 国内AI工作者在AI芯片的底层软件能力增强，为国产AI芯片的性能提升指明了新的方向，有利于国产AI芯片发展。例如，海光持续拓展软件栈DTK（DCU ToolKit）、寒武纪自建软件生态、华为昇腾发展AI框架CANN8.0版。

图：NV GPU CUDA结构图



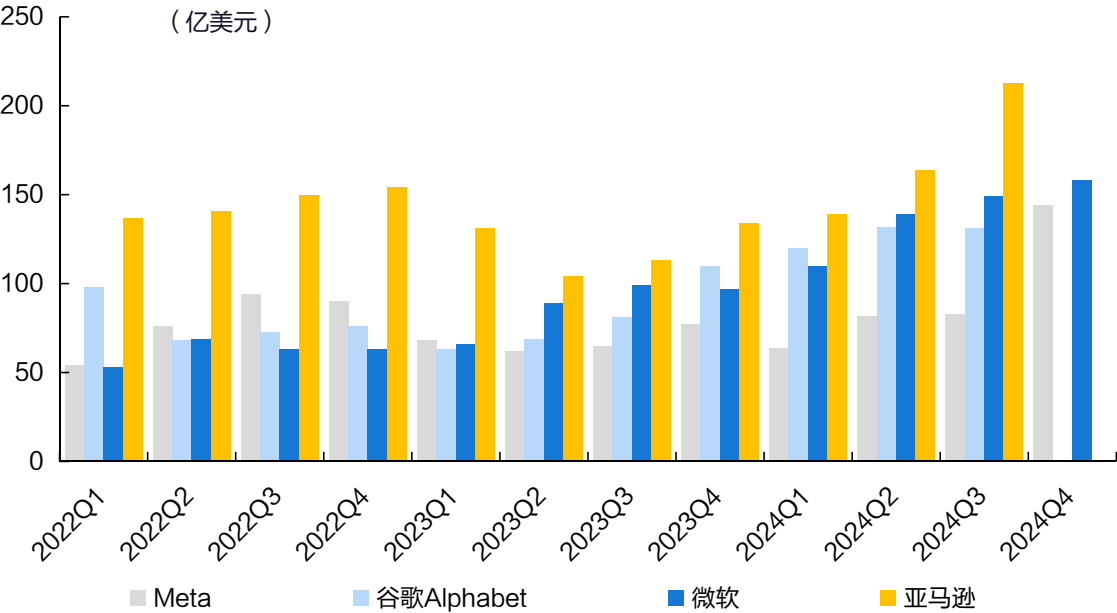
图：昇腾全栈 AI 软硬件平台，赋能昇腾生态不断发展



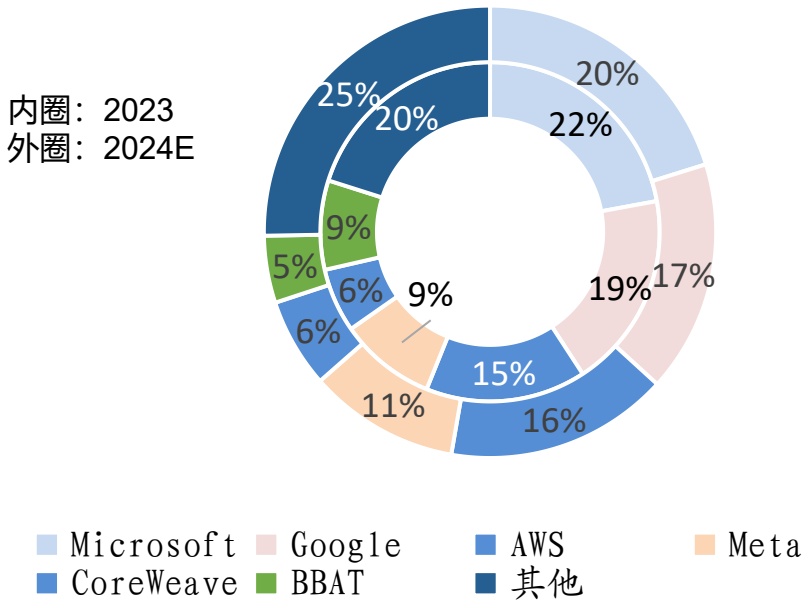
4.3.2 、ASIC：互联网厂商资本开支指引提升，ASIC服务器采购占比增长

公司	2024年互联网大厂资本开支预期情况
Microsoft	2024Q4（FY2025Q2），含融资租赁资本支出226亿美元，同比增长96.5%，环比增长13%，同比和环比增速均扩大，其中购买PP&E现金支出为158亿美元（高于一致预期1.2%）。与AI和云相关的支出中，超过一半用于15年折旧的长期基础设施资产，服务器CPU和GPU的占比有所下降。预计第三季度和第四季度的季度支出将与第二季度的支出保持相似水平。
Alphabet（谷歌）	2024Q3公司资本开支达到131亿美元。公司预计全年每季度资本支出将大致维持第一季度120亿美元或略高。
Meta	2024Q4公司资本支出（包括融资租赁本金支付）为 148 亿美元，主要用于服务器、数据中心和网络基础设施的投资。公司预计2025年的资本支出将在600–650亿美元之间，服务器仍将是最大的支出增长驱动力，非人工智能计算能力需求也会增长。
亚马逊	2024Q3资本开支为213亿美元。公司预计下半年的资本投资将更高，大部分支出将用于支持对AWS 基础设施日益增长的需求

图：2022–2024Q4 各厂商资本性开支



图：2023–2024年全球CSP对高阶AI服务器需求占比



4.3.2 、ASIC：互联网厂商资本开支指引提升，ASIC服务器采购占比增长

- ASIC芯片在性能、能效以及成本上优于标准GPU等芯片，更加契合AI推理场景的需求。
- CSP资本开支持续投向AI服务器采购。据TrendForce预估，2024年北美CSPs业者（如AWS、Meta等）持续扩大自研ASIC，以及中国的阿里巴巴、百度、华为等积极扩大自主ASIC 方案，促ASIC服务器占整体AI服务器的比重在2024年将升至26%，而主流搭载GPU的AI服务器占比则约71%。

表1：2024年搭载ASIC芯片AI服务器出货占比将逾2.5成

公司	2022	2023	2024E
NVIDIA	67.6%	65.5%	63.6%
AMD（包括Xilinx）	5.7%	7.3%	8.1%
Intel（包括Altera）	3.1%	3.0%	2.9%
Others	23.6%	24.1%	25.3%
全部	100%	100%	100%

4.3.3 、重塑价值链，机柜/铜缆/液冷/HBM占比提升

- GB200 NVL系列的发布，有望带来机柜、HBM、铜缆、液冷等市场的价值量占比提升。
 - ✓ 整机柜：机柜采用MGX架构，由计算托盘与交换托盘组成，提升组装复杂度，带来ODM整机厂的加工价值量提升。
 - ✓ HBM：H100采用5颗HBM3，Blackwell Ultra预期采用8颗HBM3e，单颗GPU采用HBM数量与单价均实现提升。
 - ✓ 铜连接：GB200 NVL72采用NVLink铜缆链接。
 - ✓ 散热：B200 GPU功耗约1200W，GB200功耗约2700W，或达到风冷上线，有望推动液冷组件价值量提升。

	GB200		H100
 整机柜	\$ 12k 每GPU 服务器ODM毛利情况	2X	\$ 6k
 HBM	\$ 3k 每GPU HBM花费	3X	\$ 1k
 铜连接	\$ 3k 每GPU	10X	\$ 0.3k
 散热	\$ 1.4k 每GPU BOM	3X	\$ 0.4k

五、投资建议及风险提示

◆ 投资建议

DeepSeek探索出一条“算法创新+有限算力”的新路径，开源AI时代或已至，国产AI估值或将重塑，维持计算机行业“推荐”评级。

◆ 相关公司

1) AI应用：

①2G：中科曙光、科大讯飞、中国软件、太极股份、深桑达A、中科星图、国投智能、云从科技、能科科技、拓尔思、航天信息、税友股份、金财互联、浪潮软件、数字政通；

②2B：金蝶国际、卫宁健康、石基信息、明源云、新致软件、用友网络、广联达、莱斯信息、四川九洲、泛微网络、致远互联、新开普、东方财富、同花顺、恒生电子、宇信科技、当虹科技、万达信息、创业惠康、润和软件、彩讯股份、第四范式、焦点科技；

③2C：金山办公、三六零、万兴科技、福昕软件、合合信息、萤石网络。

2) 算力：

①云：海光信息、寒武纪、浪潮信息、华勤技术、云赛智联、光环新网、中兴通讯、宝信软件、紫光股份、中国电信、优刻得-W、青云科技-U、首都在线、并行科技、润泽科技、中国软件国际、神州数码、深信服、新炬网络、天玑科技；

②边：网宿科技、顺网科技、云天励飞；

③端：软通动力、中科创达、乐鑫科技、移远通信。

- 1) 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；
- 2) 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；
- 3) 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；
- 4) 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；
- 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。

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，主题与价值并重，致力于进行前瞻重磅推荐。

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