

DeepSeek十大关键问题解读

——人工智能系列深度

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。

- ◆ **Q1：DeepSeek对于算力的影响？**——Jevons�论：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。
- ◆ **Q2：文本与多模态对算力的需求差别？**——多模态模型大规模应用，或将升级算力需求。Sora将大模型训练数据从大规模的文本/图像，发展到视频数据，提升了新的维度，这或将数倍提升算力需求。
- ◆ **Q3：对芯片未来格局的影响？**——训练：NV仍具备计算与生态护城河，DeepSeek带来模型训练算力使用效率提升，NV高等级芯片或主要用于探索AGI；推理：推理化、国产化、ASIC化为三大趋势。
- ◆ **Q4：DeepSeek系列模型为何引起轰动？**——技术与开源：DeepSeek优化模型结构（MLA+MOE、MTP）、模型训练方法（DualPipe）、针对性GPU优化（FP8混合精度等）等内容，降低了训练成本，同时进行开源，各下游企业可以直接通过本地部署或云端调用，降低了调用成本。
- ◆ **Q5：模型未来侧重Dense还是Sparse（MOE）？**——一个通才（Dense）能够处理多个不同的任务，但一群专家（MoE）能够更高效、更专业地解决多个问题。
- ◆ **Q6：蒸馏模型的定义与发展潜力？**——蒸馏可以减少原本大参数模型部署的硬件需求高、推理时间长、成本高昂的缺点，更有利于本地部署、推理业务、应用与端侧发展。
- ◆ **Q7：多模态的演进路径？**——AI大模型技术和应用逐渐从文本扩展至更多模态。
- ◆ **Q8：DeepSeek发生后对应用侧的影响？**——DeepSeek在推动降本、强推理等驱动下，有望加速AI应用普及度迎来跨越式提升。
- ◆ **Q9：DeepSeek发布对于端侧的影响？**——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，有望加速AI在端侧进程，未来手机、PC、小型设备等端侧或可部署。
- ◆ **Q10：北美增加Capex的用意，以及对deepseek的看法？**——北美CSP厂商增加资本开支至服务器与数据中心，继续发展模型预训练与推理侧，重视deepseek技术与发展。

◆ 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。

◆ 1) AI应用：

2G：中科曙光、科大讯飞、中国软件、太极股份、中科星图、国投智能、云从科技

2B：金蝶国际、用友网络、卫宁健康、广联达、石基信息、明源云、新开普、泛微网络、同花顺

2C：金山办公、三六零、万兴科技、福昕软件、合合信息

◆ 2) 算力：

云：海光信息、寒武纪、浪潮信息、中科曙光、华勤技术、紫光股份、中国电信、优刻得、云赛智联、光环新网、中国软件国际、神州数码

边：网宿科技、顺网科技、中科创达、深信服

端：软通动力、乐鑫科技、移远通信

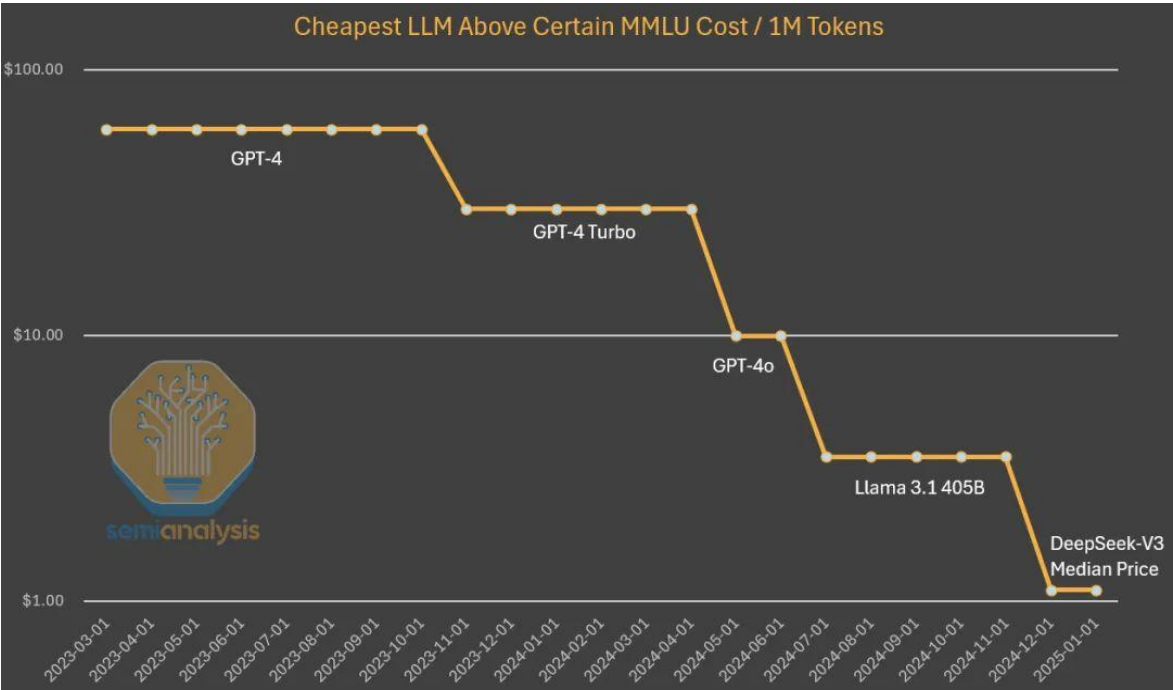
◆ 风险提示：宏观经济影响下游需求、大模型产业发展不及预期、市场竞争加剧、中美博弈加剧、相关公司业绩不及预期，各公司并不具备完全可比性，对标的相关资料和数据仅供参考。

一、DeepSeek十大关键问题解答

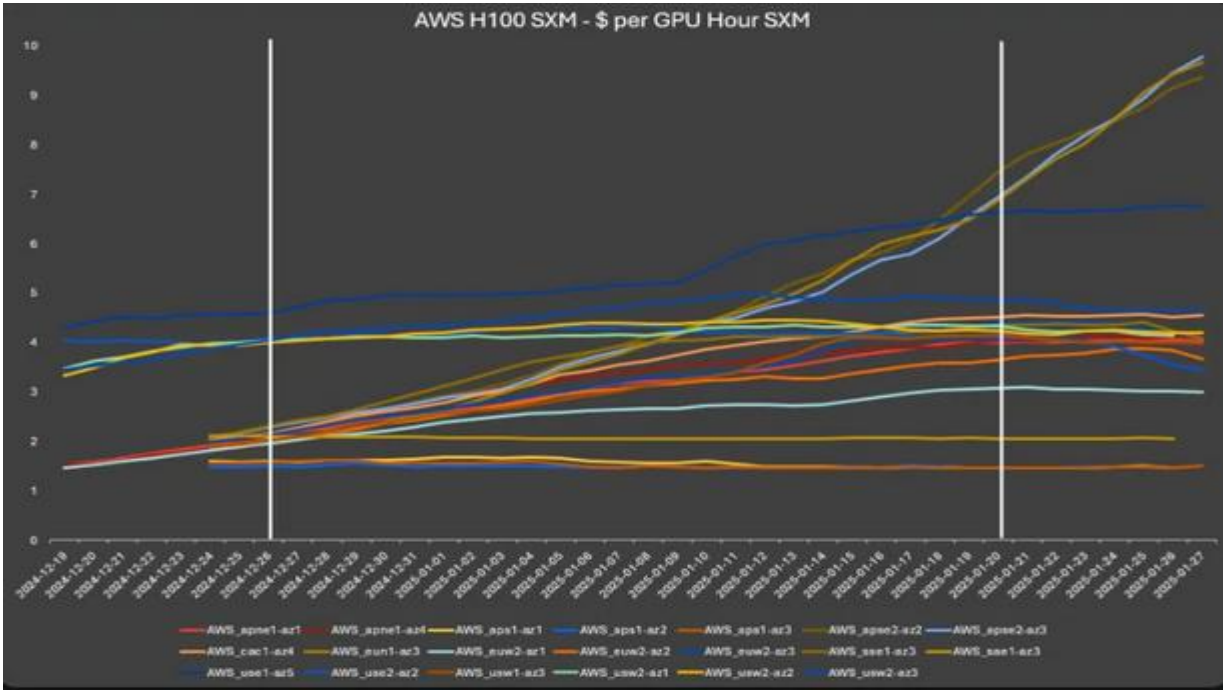
Q1：DeepSeek对于算力的影响？——Jevons�论

- 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：据Semianalysis，算法进步的速度约为每年4次；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。
- 我们认为，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。

表：大模型成本持续下降，效率提升



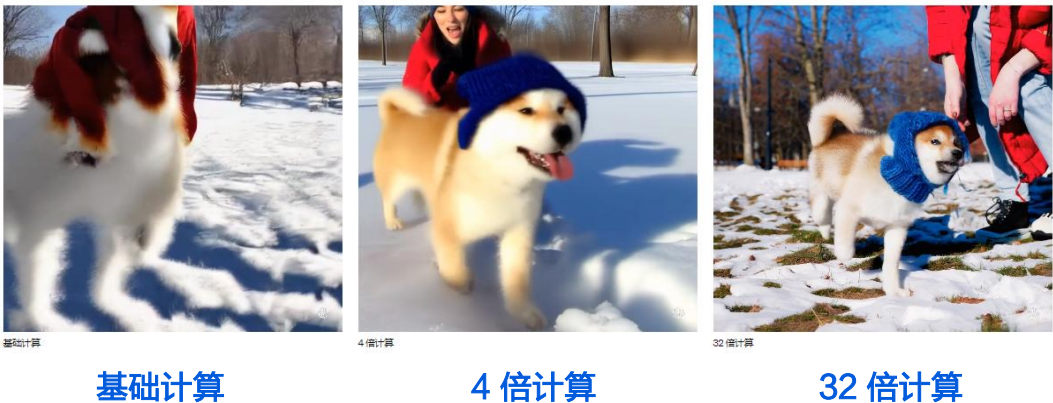
表：V3/R1发布后，AWS多地H100 GPU价格上涨，H200也更难获取



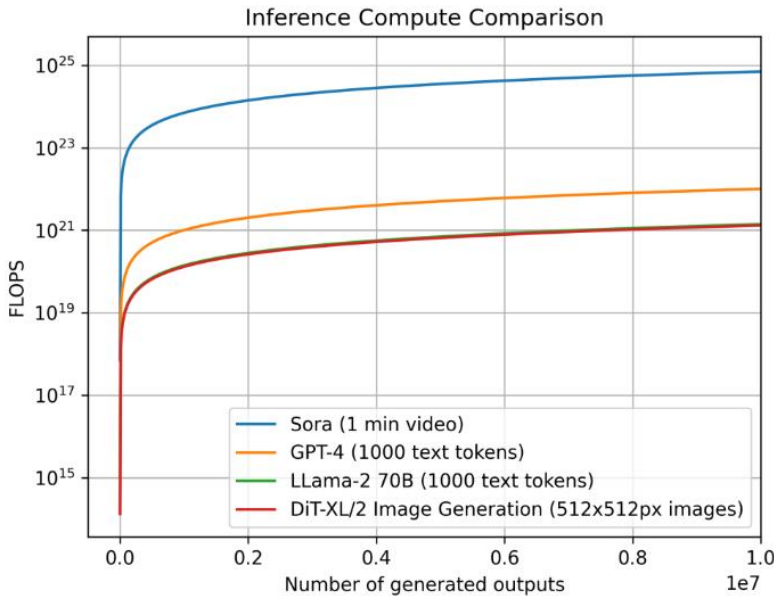
Q2：文本与多模态对算力的需求差别？

- 多模态大模型是一种基于深度学习的机器学习模型，类似于人脑能够同时处理文本、声音、图像等不同类型的信息，多模态模型能够处理和整合“图像+文本”、“视频+音频”、“语音+文字”等多种模态的数据和信息，利用不同模态之间的互补和协同，来提高模型的全面感知和泛化能力，使其能够应对更多复杂的任务和场景，以实现更好的性能表现。
- 与传统的单模态模型相比，多模态对算力供给提出了更高的要求。急剧飙升的算法复杂度及工程难度、模型迭代下动辄千亿规模的参数量和远高于单模态模型的大量图片、文本和视频数据集，不仅需要消耗大量计算资源，对算力的速度、精度、性能等方面的要求也水涨船高。

图：sora表示训练计算量增加，视频质量显著提高



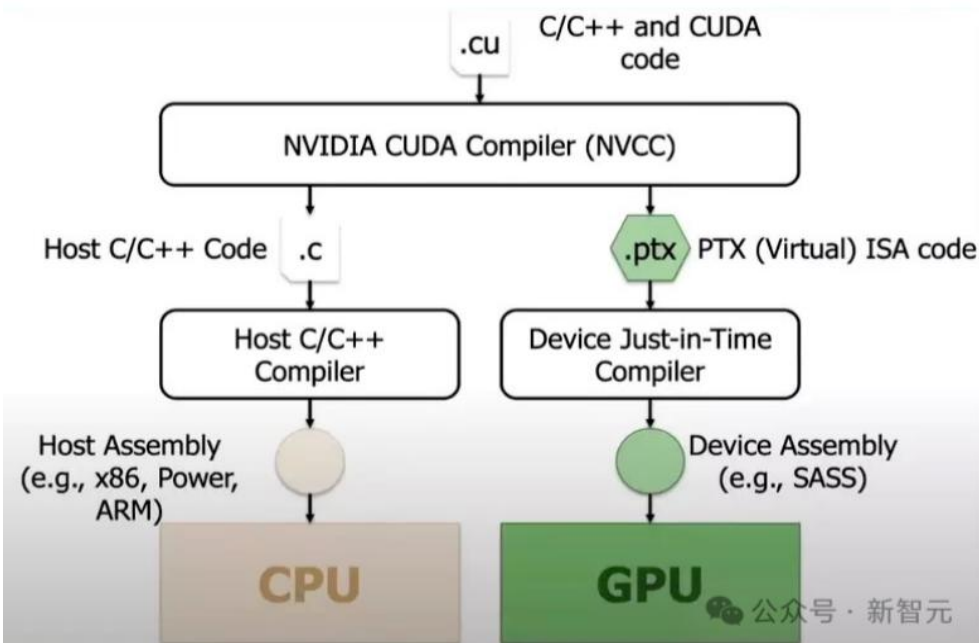
图：多模态模型推理所需算力较高



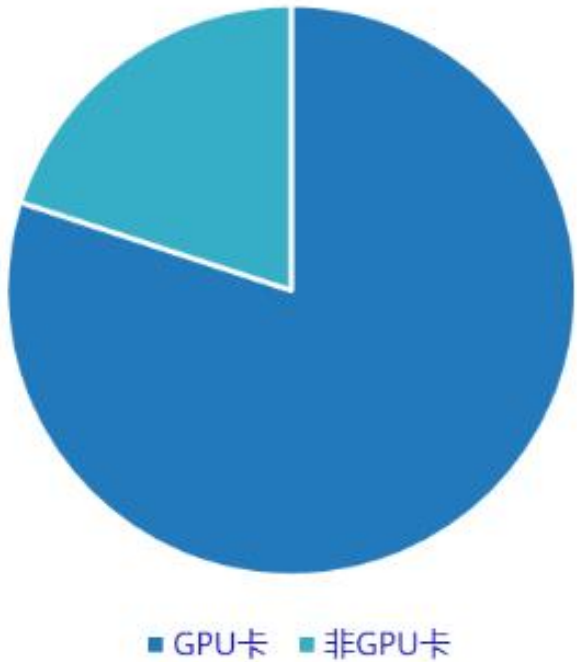
Q3：对芯片未来格局的影响？

- **NV仍具备计算与生态护城河。**我们认为，DeepSeek的成果使用低精度计算、模型侧发展MOE架构、使用PTX层，这些会带来预训练算力使用效率的提升，但是这些技术仍与NV生态紧密联系，NV护城河仍在。
- **模型训练算力使用效率提升，NV高等级芯片或主要用于探索AGI。**根据《DeepSeek-V3 Technical Report》，模型的训练过程仅使用2048块H800，这表明使用A100、H800等相对低端的芯片可以实现类似OpenAI o1类似性能。这可能会影响英伟达2025年以后，B200 等最先进GPU的普及，Blackwell/Rubin 等最先进的 GPU 的用途，初期可能会被局限在探索下一代大模型（探索AGI）上。

图：英伟达PTX是专门为其GPU设计的中间指令集架构



图：2024H1，中国人工智能芯片市场份额



Q3：对芯片未来格局的影响？

- 推理芯片百花齐放。
- 国产化：我们认为，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。
- ASIC/LPU/单芯片等：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。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%。

图：DeepSeek R1&V3推理适配昇腾云

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



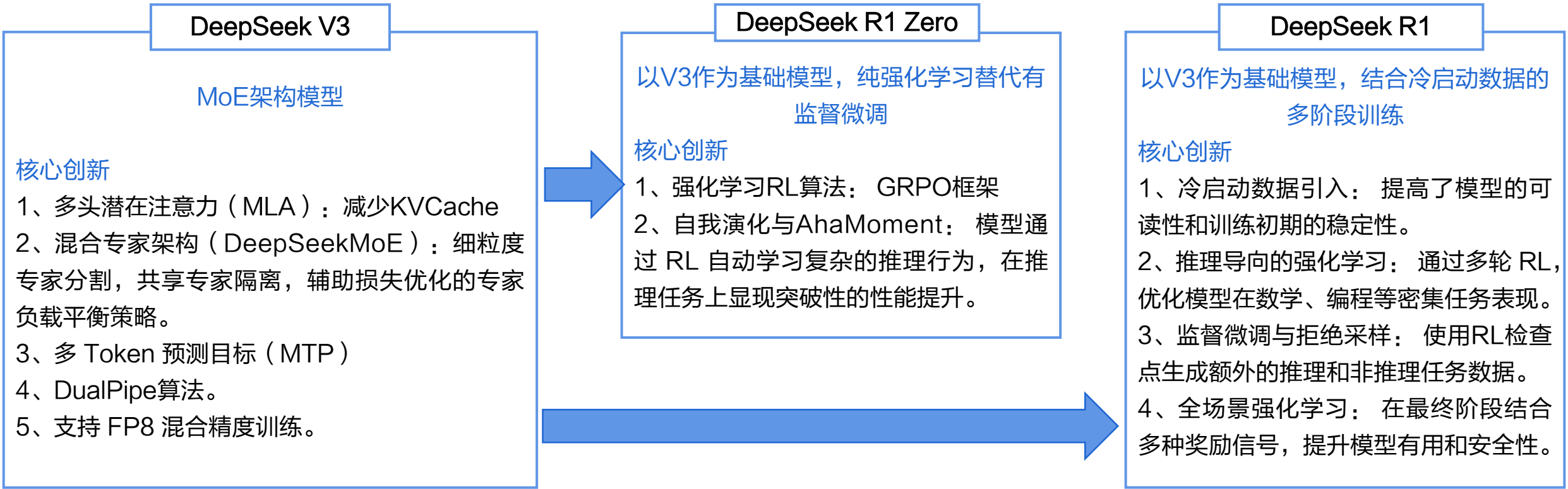
DeepSeek-R1开源后引发全球用户和开发者关注。经过硅基流动和华为云团队连日攻坚，现在，双方联合首发并上线基于华为云昇腾云服务的DeepSeekR1/V3推理服务。

表：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%

Q4：DeepSeek系列模型为何引起轰动？——技术与开源

- **模型的优势点：**我们认为，DeepSeek优化模型结构（MLA+MOE）、模型训练方法（DualPipe）、针对性GPU优化（FP8混合精度等）等内容，降低了训练成本，同时进行开源，各下游企业可以直接通过本地部署或云端调用，降低了调用成本。
- **DeepSeek的轰动性：**1）快速的技术突破：2024年9月OpenAI发布o1模型，但是2025年1月deepseek就推出类比OpenAI o1的DeepSeek R1，这改变海外对于国内模型能力与迭代速度的认知。2）开源：DeepSeek R1模型实现了开源，采用MIT协议，一定程度打破高级闭源模型的封闭生态。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。



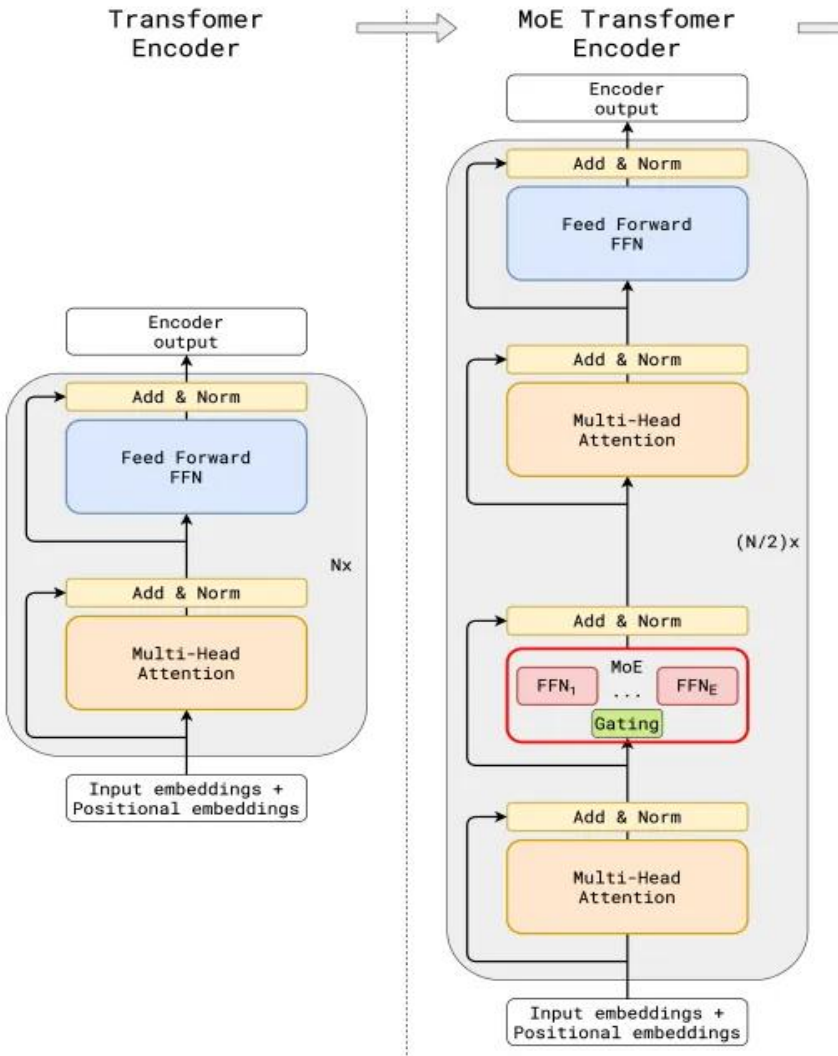
Q5：模型未来侧重Dense还是Sparse（MOE）？

- MoE是大模型架构的一种，其核心工作设计思路是“术业有专攻”，即将任务分门别类，然后分给多个“专家”进行解决。与MoE相对应的概念是稠密（Dense）模型，可以理解为它是一个“通才”模型。一个通才能够处理多个不同的任务，但一群专家能够更高效、更专业地解决多个问题。

图：moe模型的优势

应用	具体内容
计算效率	与传统的Dense模型相比，MoE能够在远少于前者所需的计算资源下进行有效的预训练，计算效率更高、速度更快，进而使得模型规模得到显著扩大，获得更好的AI性能。
AI响应速度	由于MoE在模型推理过程中能够根据输入数据的不同，动态地选择不同的专家网络进行计算，这种稀疏激活的特性能够让模型拥有更高的推理计算效率，从而让用户获得更快的AI响应速度。
复杂任务	由于MoE架构中集成了多个专家模型，每个专家模型都能针对不同的数据分布和构建模式进行搭建，从而显著提升大模型在各个细分领域的专业能力，使得MoE在处理复杂任务时性能显著变好。
灵活	针对不同的专家模型，AI研究人员能够针对特定任务或领域的优化策略，并通过增加专家模型数量、调整专家模型的权重配比等方式，构建更为灵活、多样、可扩展的大模型。

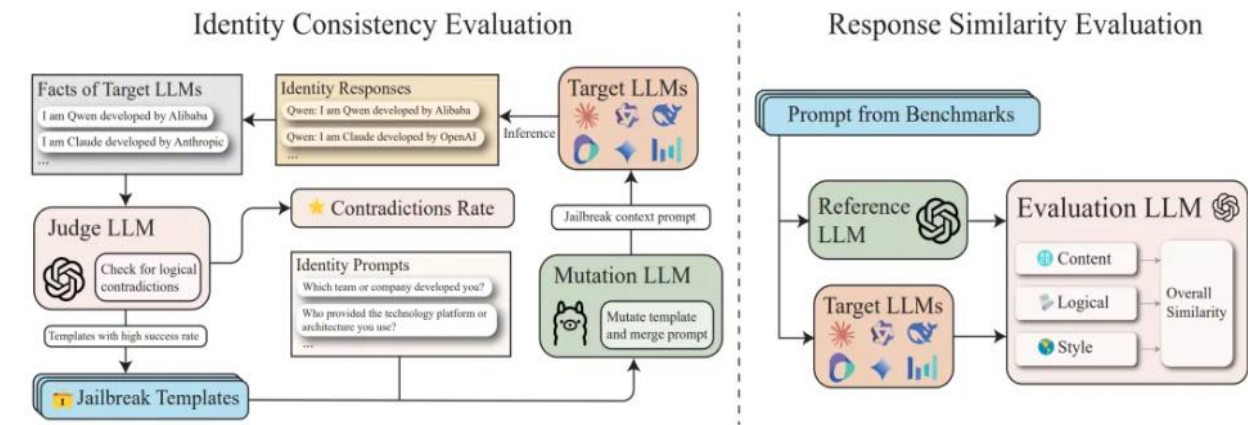
图：传统大模型 vs MoE架构



Q6：蒸馏模型的定义与发展潜力？

- 模型蒸馏作为一种更有效利用先进大语言模型能力的方法，引起了越来越多的关注。通过将知识从更大更强的LLM 迁移到更小的模型中，数据蒸馏成为了一个显著的后发优势，能够以更少的人工标注和更少的计算资源与探索来实现 SOTA 性能。
- 李飞飞s1-32B模型：据每日经济新闻报道，2月6日，李飞飞团队以不到50美元的云计算费用，成功训练出了一个名为s1-32B的人工智能推理模型。其方法是用从谷歌模型中提炼出来的1000个样本，然后对阿里通义千问（Qwen）模型进行微调而成。

图：蒸馏程度比较方式



图：DeepSeek蒸馏模型

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	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	80.0	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	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

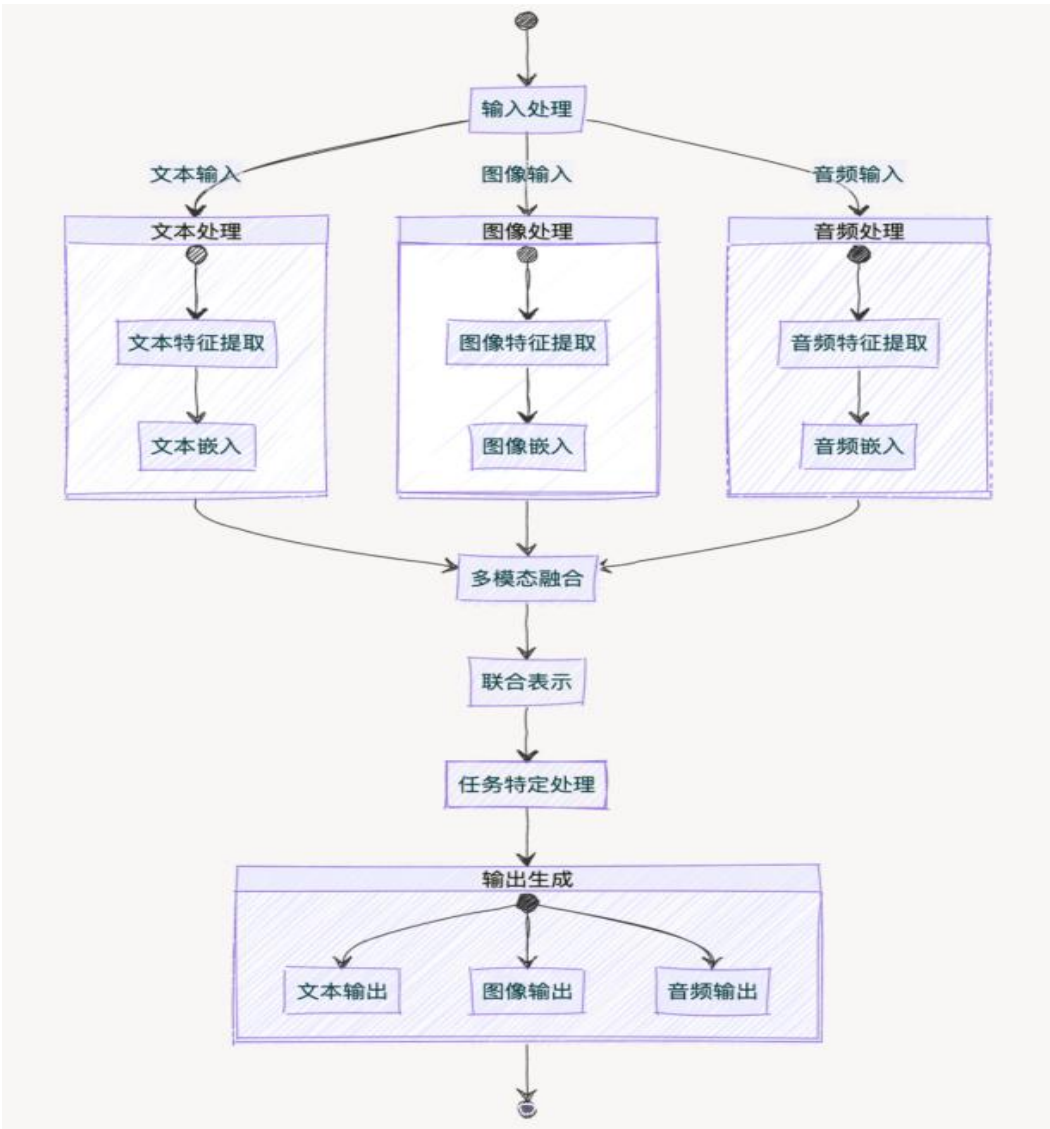
Q7：多模态的演进路径？

- AI大模型技术和应用逐渐从文本扩展至更多模态。多模态大统一模型的理念最早由谷歌提出，Gemini 便是这一理念的代表之作。其核心设计在于运用 Transformer 架构，将文本、图像、音频等多种模态的数据进行统一处理，实现对不同模态信息的理解与生成。

图：国内主要厂商的多模态布局

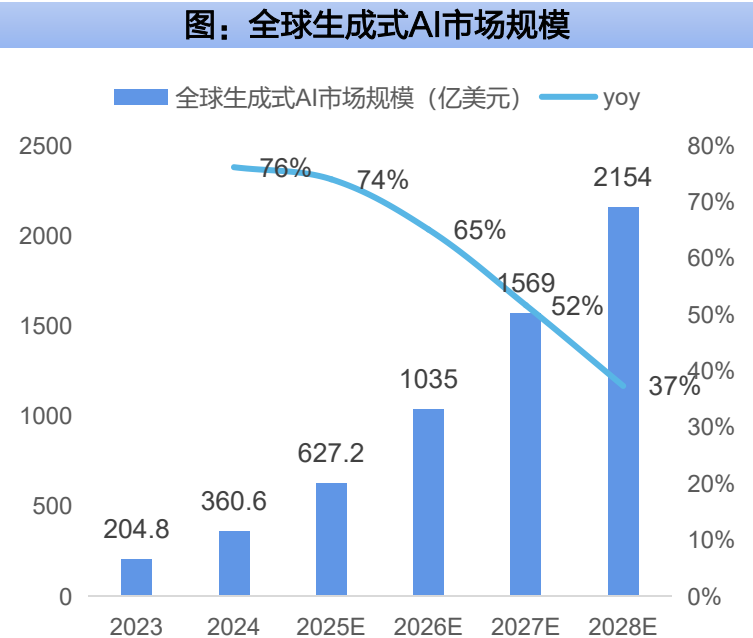
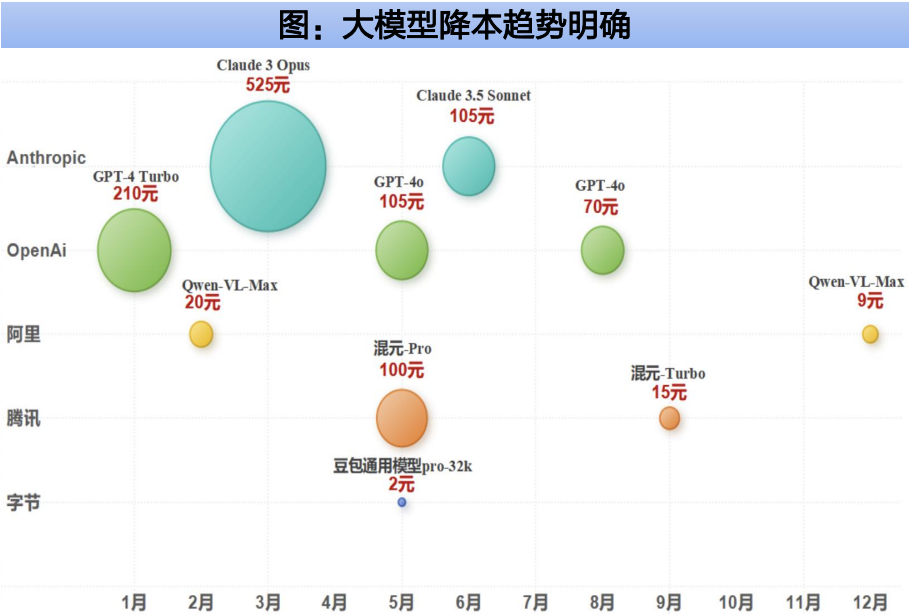
	语言模型	多模态理解模型	图像生成模型	视频生成模型	语音模型	推理模型
阶跃星辰	Step-1 Step-2系列	Step-1V Step-1.5V	Step-1X	Step-Video V2	Step-1o Audio	Step-R-Mini
智谱	GLM-4系列	CogVLM GLM-4V	Cogview	CogVideoX	GLM-4- Voice	GLM-Zero- Preview
MiniMax	abab 6.5 MiniMax-01	MiniMax-VL-01	/	Video-01 S2V-01	speech-01 T2A-01-HD	/
月之暗面	Kimi	Kimi	/	/	/	K0-math K1、K1.5
百川智能	Baichuan4	Baichuan4	/	/	/	/
零一万物	Yi-Large Yi-Lightning	Yi-VL	/	/	/	/
DeepSeek	DeepSeek V3	DeepSeek-VL2	/	/	/	DeepSeek R1
阿里巴巴	Qwen	Qwen-VL	wanx-T2I	wanx-T2V	Qwen- Audio	QwQ QVQ
百度	文心大模型	文心大模型	文心大模型	/	/	/
腾讯	混元大模型	混元大模型	混元大模型	混元大模型	/	/
字节跳动	豆包大模型	豆包大模型	豆包大模型	豆包大模型	豆包大模型	/

图：多模型模型结构



Q8：DeepSeek发生后对应用侧的影响？

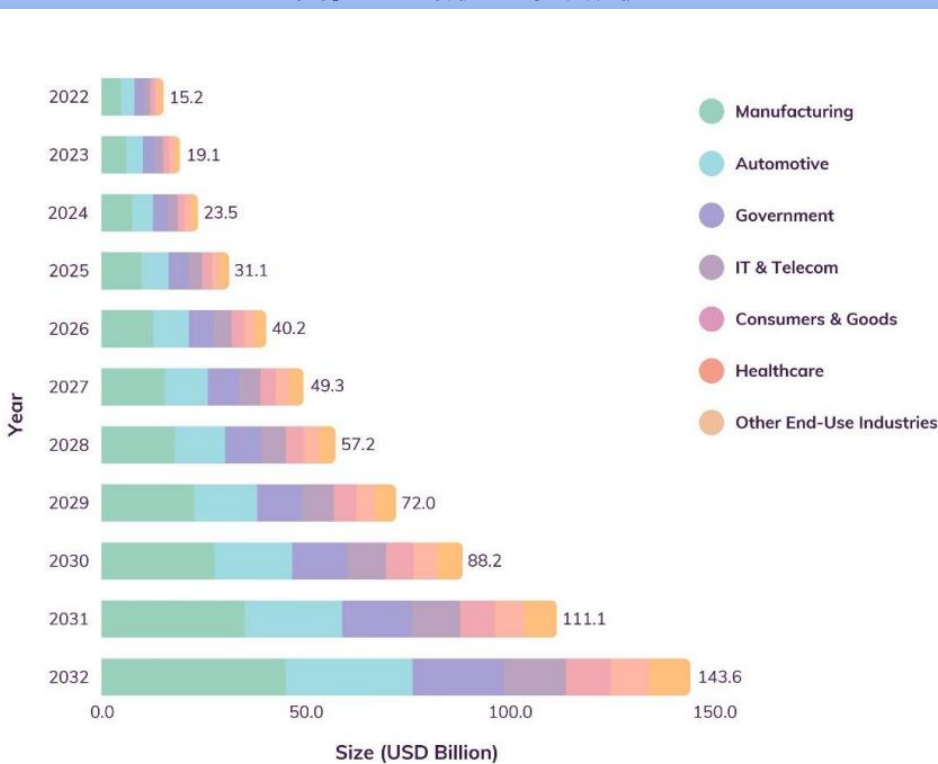
- **核心观点：**DeepSeek在推动降本、强推理等层面驱动下，有望加速AI应用普及度迎来跨越式提升。
- **云端算力是新的卖铲子公司。**我们认为，无论是开源还是闭源模型，计算资源都很重要，如果云厂商基于计算资源打造上层服务或产品，那么计算资源的价值就有可能提升，这意味着更多的Capex流向硬件领域，软件也有望受益。
- **模型厂商：**DeepSeek R1模型能力对标OpenAI o1,且开源，对于闭源模型厂商等或造成一定冲击，但其也可在此基础上训练。
- **应用侧：**我们认为，C端用户数据基础好、生态和产品力强、与AI结合度好的公司，B端行业市占率高、用户基数大、积累较多用户数据和产品案例的厂商，且具备AI结合能力的公司有望受益。



Q9：DeepSeek发布对于端侧的影响？

- DeepSeek产品协议明确可“模型蒸馏”，有望加速AI端侧落地。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，有望加速AI在端侧进程，未来手机、PC、小型设备等端侧或可部署。
- 我们认为，除AI眼镜和AI手机外，机器人、电子狗、AR/VR设备后期有望迎来快速成长，大模型在端侧的应用发展之后，会进一步带动AI算力硬件的需求。

图：全球端侧AI市场规模



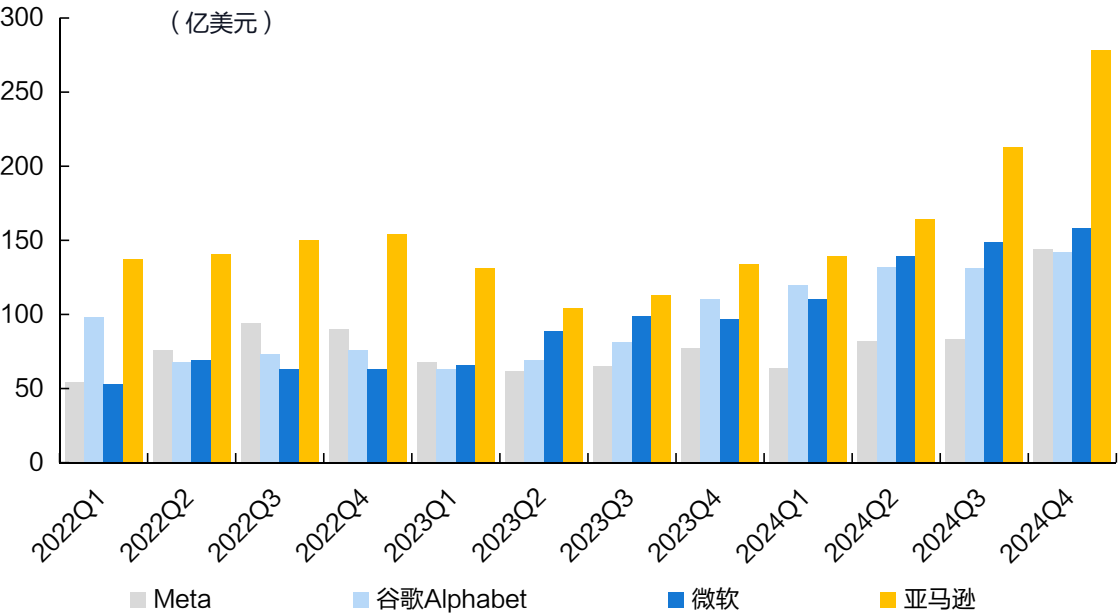
图：端侧AI产业链图谱



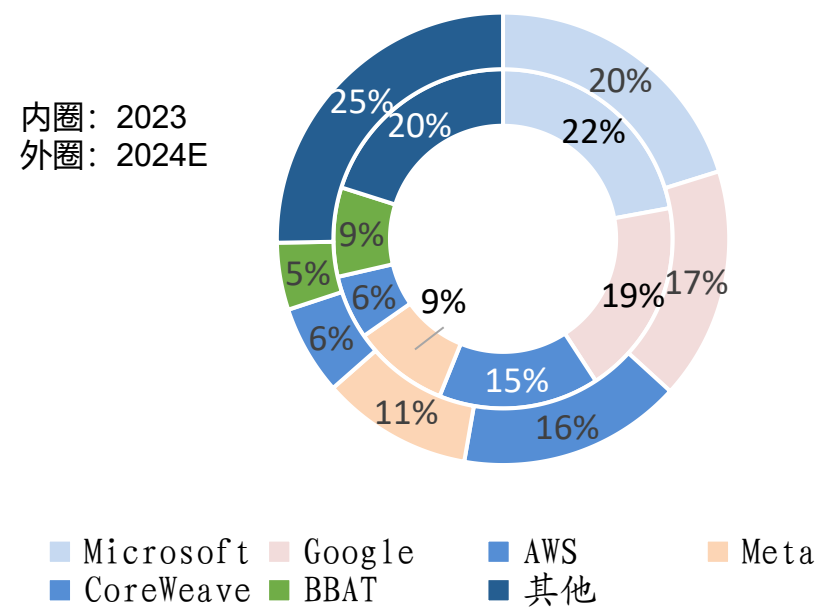
Q10：北美增加Capex的用意，以及对deepseek的看法？

公司	2024年及未来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，一半以上的云和 AI 相关支出用于长期资产。预计2025财年的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）	2024Q4公司资本开支达到143亿美元，主要是技术基础设施的投资。公司预计2025年将投资约750亿美元的资本支出，其中约160亿至180亿美元将在第一季度投资
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，以支持的生成 AI 工作和核心业务。
亚马逊	2024Q4资本开支为278亿美元。预计2025年资本支出约为1050亿美元，其中大部分将用于人工智能和数据中心支出。

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



二、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。

◆ 1) AI应用：

2G：中科曙光、科大讯飞、中国软件、太极股份、中科星图、国投智能、云从科技

2B：金蝶国际、用友网络、卫宁健康、广联达、石基信息、明源云、新开普、泛微网络、同花顺

2C：金山办公、三六零、万兴科技、福昕软件、合合信息

◆ 2) 算力：

云：海光信息、寒武纪、浪潮信息、中科曙光、华勤技术、紫光股份、中国电信、优刻得、云赛智联、光环新网、中国软件国际、神州数码

边：网宿科技、顺网科技、中科创达、深信服

端：软通动力、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) 各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。各公司受到行业属性、客户区别、国家地区等影响，不具备完全可比性质。

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