计算机行业2025年1月投资策略

国产deepseek+豆包发力，海内外大模型刺激推理算力

优于大市

核心观点

海外资本支出呈现持续上扬的态势，而国内资本支出近期承压。全球云计算厂商资本开支进入新一轮增长浪潮，AI基础设施成为核心驱动力。国外厂商中，微软24Q3资本开支同比增78.6%，主要投向AI和云服务；谷歌维持高位，聚焦服务器和数据中心；亚马逊连续五季环比增长，全年预计750亿美元。国内厂商中，阿里巴巴24Q3同比增239.63%，加码AI基础设施；腾讯同比增113.54%，布局GPU、CPU服务器及数据中心；百度虽承压，预计24年底或25年初迎来回升，重点投入AI模型和智能云服务。

全球服务器出货同比回升，B+C端应用逐步落地，思维链等新技术拉动推理算力需求增长。C端应用如ChatGPT访问量持续增长，国内多款AI产品MAU快速上升；B端AI赋能商业增长显著，AppLovin、Palantir等公司业绩大幅提升。AI推理侧需求因思维链（CoT）技术及模型参数量增加快速增长。CoT对千亿参数模型显著提升推理能力，同时推理次数和算力需求快速增加。中国AI芯片市场规模23年达1038.8亿元，预计25年增长至1780亿元，推理算力占比24年有望升至67.7%。国内外厂商如博通、Marvell、寒武纪等积极布局AI推理硬件，助推算力发展。

国产科技巨头在AI大模型与算力领域持续突破。字节跳动发布豆包Pro对标GPT-4o，API调用量大增，多场景渗透；召开冬季FORCE原动力大会，推出数据飞轮2.0，强化全模态数据管理；正式推出情感大模型，在豆包APP全量开放。小米升级MiLM2模型，参数灵活扩展，端云结合适配多场景；加速GPU集群建设提升算力。阿里倚天710芯片大规模落地，阿里云为双11提供百万核级算力。腾讯升级星脉网络2.0，优化网络协议与通信库，提升大模型训练效率。

多层面技术提升训练效率，测试性能领跑开源模型。2024年12月26日，DeepSeek上线并开源DeepSeek-V3模型，多项评测超同类开源模型，在重要领域与顶尖闭源模型相当，训练成本低。模型层采用MoE架构，经多阶段训练与能力提炼，在知识、代码、数学推理等测评中领先开源模型。架构层沿用V2架构，引入新技术，如无辅助损失负载均衡策略、MTP提升数据利用率。训练层通过DualPipe算法和FP8混合精度训练实现成本控制与效率提升。推理层先推出R1-Lite模型，后将R1推理能力迁移至V3提升其性能，推理算力包含GB300、博通、marvell等各类asic芯片。2025年1月发布的DeepSeek-R1模型在多测试中超越OpenAI的o1，在数学、编程及多种测试中表现出色。

风险提示：大模型研发进展不及预期、云厂商资本开支投入不及预期、国产算力迭代及供应不及预期。

重点公司盈利预测及投资评级

公司代码	公司名称	投资评级	昨收盘（元）	总市值（百万元）	EPS 2024E	EPS 2025E	PE 2024E	PE 2025E
688111	金山办公	优于大市	279.80	127,900	3.26	4.05	85.82	69.08
688041	海光信息	优于大市	133.03	307,600	0.78	1.00	170.55	133.03
000977	浪潮信息	优于大市	52.33	78,050	1.24	1.45	42.20	36.09

资料来源：Wind、国信证券经济研究所预测

行业研究·行业月报

计算机

优于大市·维持

证券分析师：熊莉
021-61761067
xiongli1@guosen.com.cn
S0980519030002

证券分析师：库宏垚
021-60875168
kuhongyao@guosen.com.cn
S0980520010001

证券分析师：艾宪
0755-22941051
aixian@guosen.com.cn
S0980524090001

联系人：云梦泽
021-60933155
yunmengze@guosen.com.cn

市场走势

资料来源：Wind、国信证券经济研究所整理

相关研究报告

《IDC专题报告:AIDC周期来临,各厂竞逐份额》——2025-01-22
《人工智能行业快评:-英伟达发布AI Blueprints,看好AI Agent发展》——2025-01-10
《国信证券人工智能专题：2024年美股SaaS回顾——整体估值修复，关注AI技术赋能》——2025-01-05
《IDC建设周期在即，关注国产算力机遇》——2025-01-02
《人工智能专题：小米AI布局》——2024-12-30

内容目录

图表目录

需求侧：海外资本支出呈现持续上扬的态势，而国内资本支出近期承压

美国对华芯片管制措施的持续加码，严重限制了中国获取高端半导体芯片的途径，这激发了对国产算力需求的急剧增加。以下是近期美国对华芯片管制措施：

图1：美国近期对华芯片管制措施

25 年 1 月 13 日	美国拜登政府周一宣布，将进一步限制人工智能芯片和技术的出口，将全球划分为三个等级，把先进的计算能力留在美国及其盟国，同时寻找更多方法阻止中俄伊朝等国的入径。
25 年 1 月 2 日	美国对华芯片管制新规正式生效，开始实施对半导体、人工智能和量子领域对华投资的限制。
24 年 10 月 28 日	美国政府正式宣布，限制美国企业和美国人在半导体、人工智能（AI）和量子领域向中国投资的新规将从 2025 年 1 月起生效。这一举措旨在防止美国的资本和专业知识被用于帮助中国开发先进关键技术。
24 年 6 月 21 日	美国财政部公布新规提案，进一步细化了对华芯片管制的具体措施，包括禁止美国投资着重先进半导体技术的中国公司等。

资料来源：中国科学院网信工作网，新浪财经，国信证券经济研究所整理

复盘云计算厂商历史，全球性资本开支增长仅有两次：第一次，2010 年美国制定“云优先”的发展战略，全球云计算蓬勃发展，各厂商资本开支均快速增长；第二次，23 年以来，全球人工智能快速发展，云厂商大力进行 AI 基础设施建设，驱动新一轮资本开支上升周期。

国外云产商资本开支持续增加

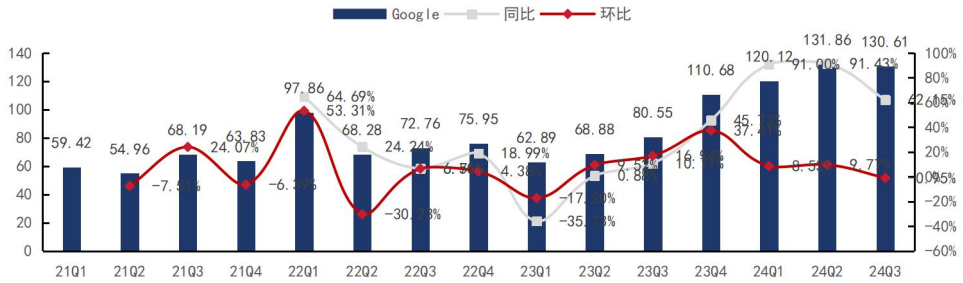
微软：24 年资本开支维持高同比增速。24Q3 资本支出（包括融资租赁）200 亿美元，同比+78.6%、环比+5.3%，其中用于购买财产、厂房和设备的现金支付为 149 亿美元，同比+50.7%。根据公司财报电话会议披露，资本开支总体投向 AI 和云，其中约一半用于长期资产，另一半用于采购服务器（包括 CPU 及 GPU）。公司预计，随着公司扩大人工智能服务规模，资本支出将继续增加。遇进一步提升市占率。

图2：微软 FY25Q1（=24Q3）资本开支 200 亿美金，同比+78.6%、环比+5.3%

资料来源：微软财报，国信证券经济研究所整理

谷歌：24 年资本开支维持高位。谷歌 FY24Q3 资本开支为 130.61 亿美元，同比+62.15%、环比-0.95%，环比基本持平。资本开支主要用于服务器、网络设备的购买以及数据中心的建设，融资租赁成本在财务上不显著；同时，预计 Q4 资本开支将与 Q3 持平。

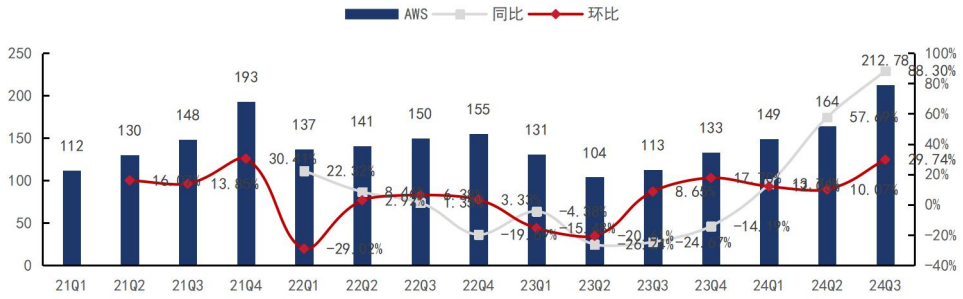
图3：谷歌 FY24Q3 资本开支 131 亿美金，同比+62.15%、环比-0.95%



资料来源：谷歌官网，国信证券经济研究所整理

亚马逊：资本开支连续五季度环比增长。24Q3 资本开支达 212.78 亿美元，同比+88.33%、环比+29.7%，连续五个季度环比增长，大部分支出主要投资于 AI 服务需求，同时包括支持北美和国际业务的技术基础设施。全年资本开支预计 750 亿，下个季度预计 231 亿美金。

图4：亚马逊 FY24Q3 资本开支 213 亿美金，同比+88.33%、环比+29.7%

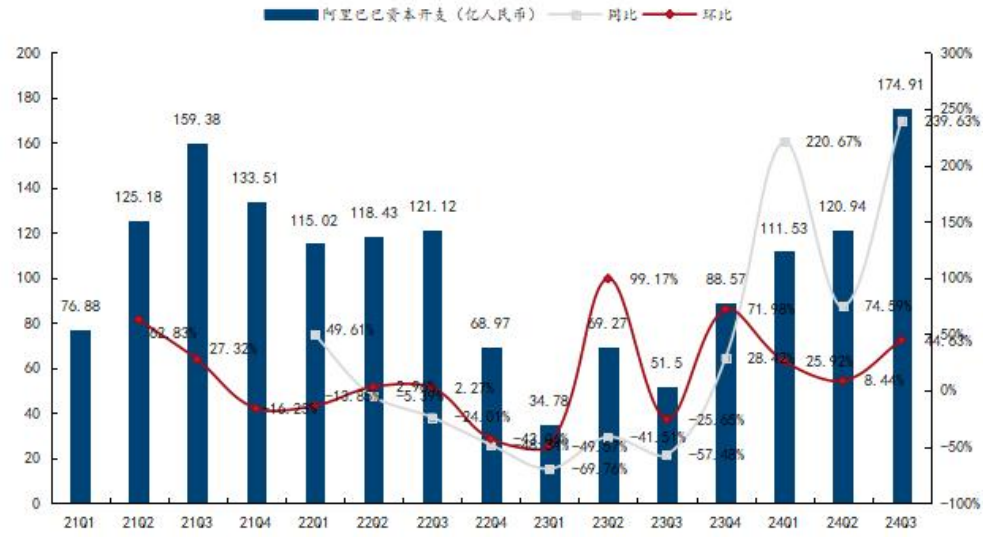


资料来源：亚马逊官网，国信证券经济研究所整理

国内云产商资本支出复苏有望

阿里巴巴：资本开支迅速复苏。2024 年 Q3 资本支出达到峰值 174.91 亿元，同比+239.63%，环比+44.63%，连续五个季度快速增长，主要投向 AI 基础设施，如数据中心、服务器等，以满足 AI 相关产品的需求。还投入电商 AI 算力，支持多款在研 AI 产品，提升用户体验和商家服务。预计未来阿里巴巴的资本支出将继续保持增长态势，主要投向 AI 领域，包括对 AI 初创公司的投资、数据中心建设以及相关服务器等基础设施的购置。

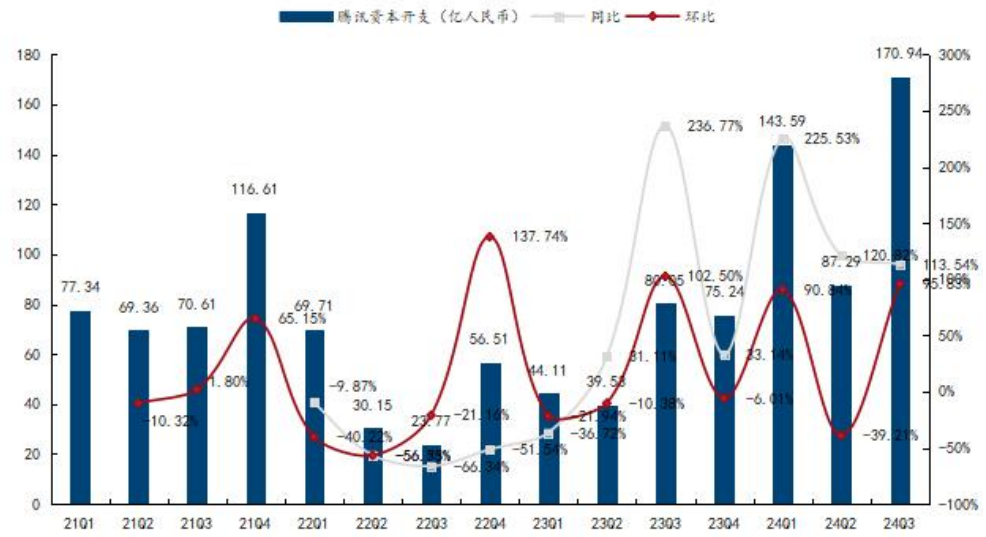
图5：阿里巴巴 FY24Q3 资本开支 174.91 亿人民币，同比+239.63%、环比+44.63%



资料来源：阿里巴巴官网，国信证券经济研究所整理

腾讯：资本开支于 2024 年同比显著提升。24Q3 达峰值 170.94 亿人民币，同比+113.54%，环比+95.83%，大部分支出主要投入 AI 领域，如对 GPU 和 CPU 服务器的投资增加。还用于数据中心建设，包括新建多个百万级服务器规模的数据中心，也投向与 AI 相关的企业，如 MiniMax、智谱 AI、百川智能等。预计未来腾讯资本支出仍将保持一定规模，持续投入 AI 领域及数据中心建设等，同时也会根据电商等业务的发展情况进行相应的资本布局。

图6：腾讯 FY24Q3 资本开支 170.94 亿人民币，同比+113.54%、环比+95.3%

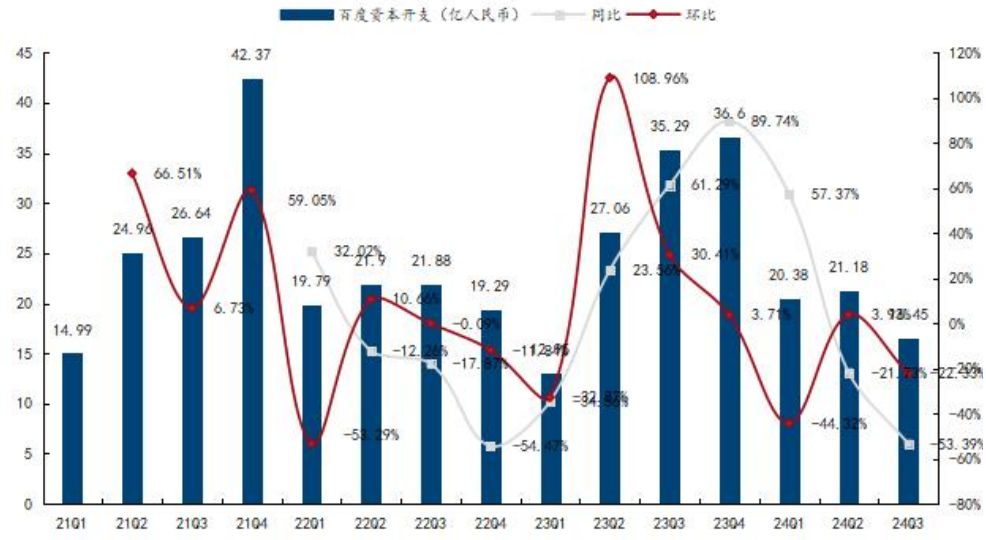


资料来源：腾讯官网，国信证券经济研究所整理

百度：资本开支近期承压。资本支出于 21Q4 达到峰值后，于 2022 年同比减少，并在 2023 年再次回升，资本开支上行周期间隔约 4-5 个季度，目前资本开支仍处于低位，根据历史经验，资本开支有望在 24Q4 或 25Q1 重回上行周期。24 年 Q3 资

本支出为 16.45 亿人民币，主要投入人工智能领域，用于购买支持大型人工智能语言模型训练的处理器和基础设施。同时，百度也在智能云服务方面有所投入，推动该业务增长。

图7：百度 FY24Q3 资本开支 16.45 亿人民币

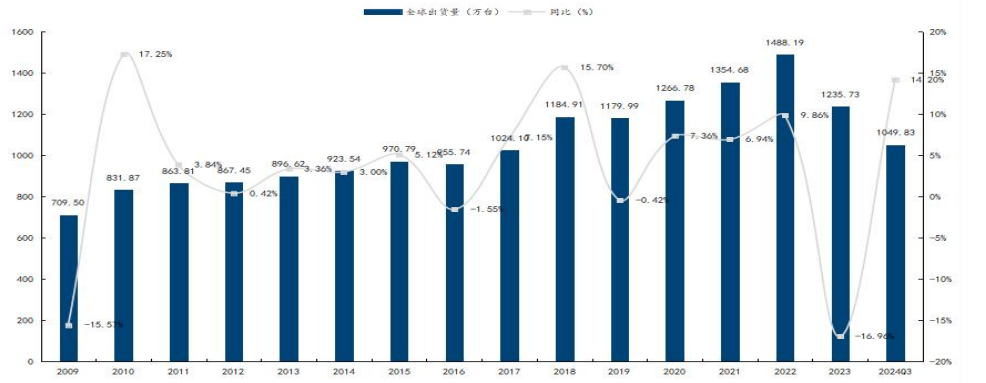


资料来源：百度官网，国信证券经济研究所整理

推理侧：全球服务器出货同比回升，B+C 端应用逐步落地拉动推理算力需求增长

全球服务器分年度出货情况。出货量于 2022 年达峰值，并在 2023 年同比下降。24Q1-Q3，服务器出货量同比再次回升，Q1-Q3 全球服务器出货量达 1049.83 万台，同比+14.2%。

图8：全球服务器分年度出货复盘



资料来源：IDC，国信证券经济研究所整理

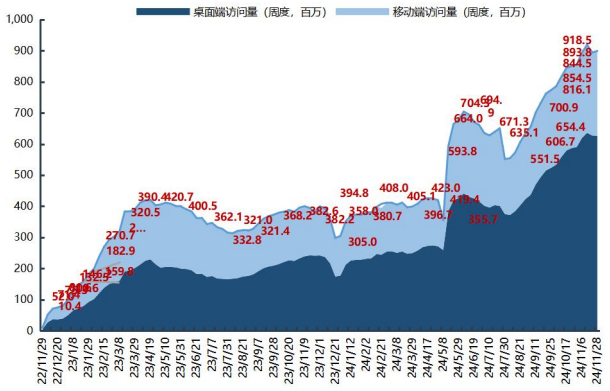
国内服务器出货情况：与全球出货走势类似，出货量于 2022 年达峰值，并在 23 年开始下滑，24Q1-Q3 服务器出货量同比回升，Q1-Q3 出货达 313.68 万台，同比+18.83%。

图9：国内服务器分年度出货复盘

资料来源：IDC，国信证券经济研究所整理

C 端：AI 应用数据持续增长。从全球来看，ChatGPT 周度访问量持续提升，根据 Similarweb 数据，最近一周（11 月 28 日-12 月 4 日）访问量合计为 8.99 亿次，环比+0.5%；根据 AI 产品榜数据，11 月 ChatGPT 的 MAU 数据为 287.25M，环比+11.27%。从国内来看，豆包、文小言、Kimi、智谱清言等 AI 应用快速发展，根据 AI 产品榜数据,11 月豆包、文小言、Kimi、智谱清言 MAU 分别为 59.98M、12.99M、12.82M、6.37M，分别环比+16.92%、3.33%、27.40%、22.18%。

图10：ChatGPT 周度访问量数据持续上升



资料来源：Similarweb，国信证券经济研究所整理

图11：11 月全球 TOP10 AI 产品有 7 款 MAU 环比增长

全球排名	AI产品榜	产品名	应用(APP)简短描述 aicpb.com	11月上榜应用 APP MAU	11月上榜应用 MAU变化
1		ChatGPT	The official app by OpenAI	287.25M	11.27%
2		豆包	AI 智能助手 | 抖音	59.98M	16.92%
3		Nova	聊天AI与AI写作机器人	49.63M	5.67%
4		ChatOn	Powered by ChatGPT & GPT-4o	28.84M	6.66%
5		Remini	人工智能修图	27.96M	-2.16%
6	c.ai	Character AI	Chat Ask Create	26.88M	5.74%
7		FaceApp	AI 人脸编辑器	26.48M	0.20%
8		Ask AI	Chat with Ask AI	26.35M	-7.16%
9		Talkie AI	Chat With Character | MiniMax	25.19M	22.14%
10		Chatbot AI	Chatbot AI & Smart Assistant	23.1M	3.85%

资料来源：AI 产品榜，国信证券经济研究所整理

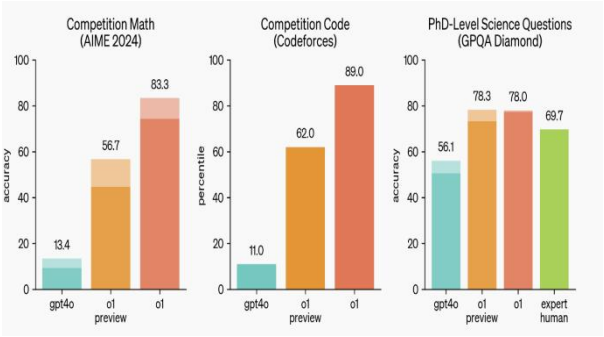
B 端：AI 赋能公司业绩增长。在广告领域，AppLovin 在 AppDiscovery 平台使用 AXON 2.0 AI 驱动技术拉动广告商支出增长，三季度收入同比+39%、净利润同比+300%，且四季度展望乐观；在数据分析领域，Palantir 推出 AIP 平台，集成多款大模型，用于数据分析，拉动其商业客户收入增长，三季度收入同比+30%（其中美国商业业务同比+54%），且上调全年收入指引为 28.05-28.09 亿美元（前值为 27.42-27.50 亿美元）。Salesforce、DocuSign、Asana 等公司受益于 AI 驱动，

三季度业绩表现出色。

AI 应用逐步落地，拉动推理算力需求增长。从单次推理来看，主要包括分词（Tokenize）、嵌入（Embedding）、位置编码（Positional Encoding）、Transformer 层、Softmax，推理主要计算量在 Transfomer 解码层，对于每个 token、每个模型参数，需要进行 2×1 Flops = 2 次浮点运算，则单词推理算力消耗为模型参数量×（提问 Tokens + 回答 Tokens）×2，随着模型参数量增长、模型向多模态发展，单次推理算力消耗持续增长。从推理次数来看，AI 应用逐步落地，模型推理次数提升，拉动推理算力需求快速增长。

OpenAI 发布 GPT-O1，通过思维链提升模型推理能力。24 年 9 月 12 日，OpenAI 发布 GPT-O1，同 GPT-4O 相比，GPT-O1 在数学、代码、科学问题（PhD 级别）评分显著提升。GPT-O1 在回复用户问题之前会生成一条较长的内部思维链，将复杂的问题拆分为更简单的步骤，且当前方法无效时，会进一步尝试其他方式，引入思维链将显著提升模型的推理能力。

图12：GPT-O1 在数学、代码、科学问题（PhD 级别）评分显著高于 GPT-4o



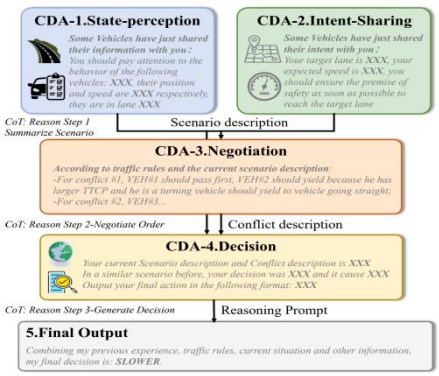
资料来源：OpenAI 官网，国信证券经济研究所整理

图13：GPT-O1 推理占比大幅提升

资料来源：JimFan（From X），国信证券经济研究所整理

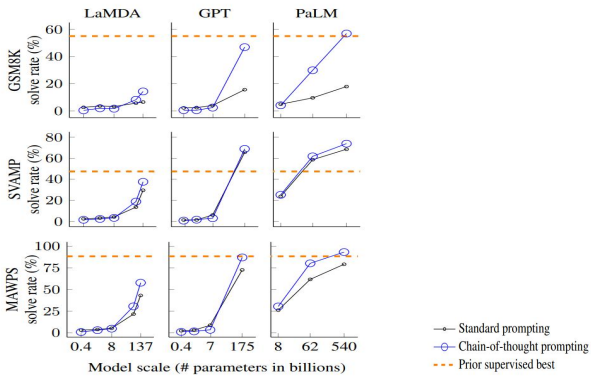
思维链（CoT）拉动推理算力增长。1）思维链（CoT）需要多步推理进而大幅提升推理算力的需求，同时推理时间的增长亦是推理算力消耗增长的反映；2）根据 Jason Wei 等人在 23 年发布的文章《Chain-of-Thought Prompting Elicits Reasoning in Large Language Models》，思维链仅对 1000 亿以上参数模型的推理有显著提升；此前，为节省推理算力消耗，大多数模型通过蒸馏等方式缩小模型参数量，而思维链反向限定模型参数量下限，进而拉动推理阶段算力需求增长。

图14：思维链多步推理提升推理阶段算力消耗



资料来源：Shiyu Fang 等著-《Towards Interactive and Learnable

图15：思维链（CoT）在 1000 亿参数模型上才能带来显著提升



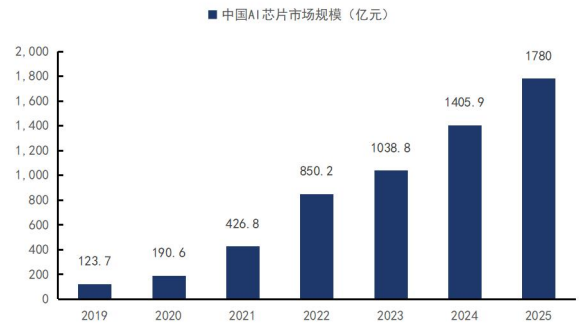
资料来源：Jason Wei 等著-《Chain-of-Thought Prompting Elicits

Cooperative Driving Automation: a Large Language Model-Driven Decision-Making Framework》-arXiv（2024）-P6，国信证券经济研究所整理

Reasoning in Large Language Models》-arXiv（2023）-P5，国信证券经济研究所整理

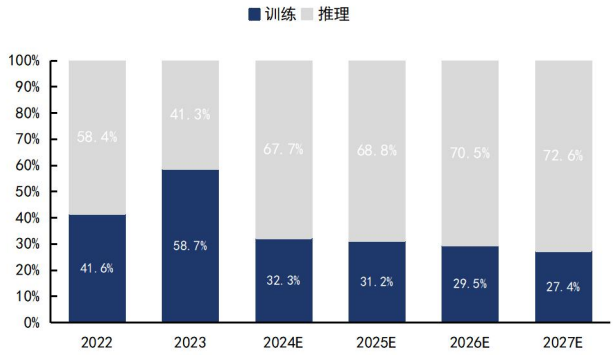
中国 AI 芯片市场规模快速增长，推理算力占比有望提升。根据亿欧智库数据，23 年中国 AI 芯片市场规模约 1038.8 亿元，预计 25 年增长至 1780 亿元，对标 23-25 年 CAGR 为 30.9%，中国 AI 芯片市场规模快速增长。随着 AI 应用逐步落地以及思维链等技术的运用，推理侧算力需求有望快速提升，根据《2023-2024 年中国人工智能计算力发展评估报告（IDC&浪潮）》发布数据，预计 24 年中国推理算力占比为 67.7%，同比+26.4 个 pct。

图16：中国 AI 芯片市场规模快速增长



资料来源：亿欧智库，国信证券经济研究所整理

图17：预计 24 年开始推理算力占比大幅提升



资料来源：《2023-2024 年中国人工智能计算力发展评估报告（IDC&浪潮）》-P18，国信证券经济研究所整理

- **博通：**从客户来看，公司为谷歌、Meta 等客户定制 AI Asic 芯片；从技术来看，公司可以提供高复杂度的定制加速卡（XPU）和 Asic，主要包括：1）计算：处理单元架构（客户提供）、设计流程和性能优化（博通）；2）内存：HBM PHY、集成与性能（博通）；3）网络 I/O：架构及执行（博通）；4）封装：2.5D、3D 和硅光子体系结构（博通）。
- **Marvell：**公司 19 年收购 Avera，随后宣布提供定制 Asic SOC 服务，目前客户主要有亚马逊等。
- **海光信息：**公司深耕 AI 芯片领域，采用 GPGPU 架构，其 DCU 芯片在推理领域表现出色。
- **寒武纪：**公司云、边、端三位协同，发布思元 370 加速卡，推理领域表现出色。
- **云天励飞：**公司 Deep Edge 系列推理卡已经适配了包括云天天书、通义千问、百川智能、以及 Llama2/3 等在内的近十个主流大模型。

国产科技巨头在 AI 大模型与算力领域持续突破

字节跳动：召开冬季原动力大会，豆包 API 调用量急剧增长

字节火山引擎冬季 FORCE 原动力大会在上海召开，大会展示了多款 AI 相关技术

成果的前沿更新。其中包括 1）涵盖视觉理解、图文以及 3D 模型生成等多功能，多场景实现的“豆包”大模型。预计于 2025 年 1 月上线的“豆包 pro”将全面对标 GPT-4o。2）AI 图片视频创作平台“即梦 AI”。3）涵盖高效记忆，具有 AI 搜索推荐功能的企业级模型“火山方舟大模型”，火山引擎在汽车智能化快速拓展，豆包已服务国内超 80%的汽车品牌。4）企业级应用及开发平台“扣子”及企业专属 AI 应用创新平台 HiAgent。政策端，央企市值管理意见出炉，聚焦四大看点。

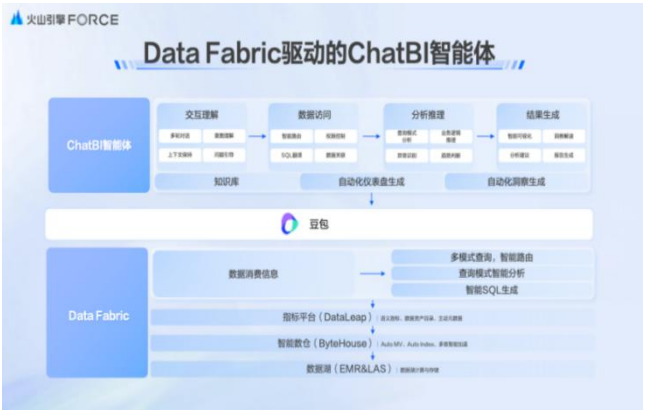
算力方面：火山引擎冬季 FORCE 原动力大会发布数据飞轮 2.0 版本，通过 AI 创新重新定义企业数据智能，用“多模态数据湖”更好应对大模型时代的数据管理。本次 2.0 版本的升级包括：智能数据洞察 DataWind ChatBI 智能体、增长分析 DataFinder 智能分析助手、A/B 测试 DataTester 智能实验助手、客户数据平台 VeCDP 智能营销助手、增长营销平台 GMP 创意助手、大数据研发治理套件 DataLeap 运维助手和 E-MapReduce 全模态数据处理引擎等，全系列火山引擎数智平台产品 AI 能力的发布。

两大数据飞轮 2.0 核心解决方案首次公开亮相。其一为 “DataFabric 驱动下的 ChatBI 智能体解决方案”，亮点在于赋予业务自定义的数据智能体能力，有效降低业务调用与理解数据的难度；其二是 “多模态数据湖解决方案”，该方案专注于处理全模态数据，扩容企业潜在数字资产规模。

图18：火山引擎数据飞轮 2.0 模式图

资料来源：火山引擎，国信证券经济研究所整理

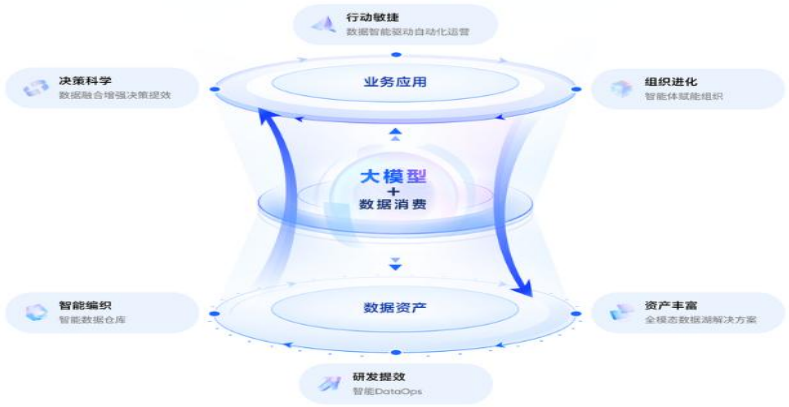
图19：火山引擎 Data Fabric 驱动下的 ChatBI 智能体解决方案



资料来源：火山引擎，国信证券经济研究所整理

数据飞轮 2.0 将全面融合大模型，通过一体化数智研发与一站式数据智能运营，使工作流程化繁为简。同时，在多元计算引擎的处理下，企业可以快速处理结构与非结构数据，激发企业更多潜在数据资产。截止目前，数据飞轮 2.0 基本实现数据生产、管理与应用各环节全方位 AI 能力深度融合，推动企业数据消费便捷化、资产建设低门槛化，加速企业数据价值实现进程。

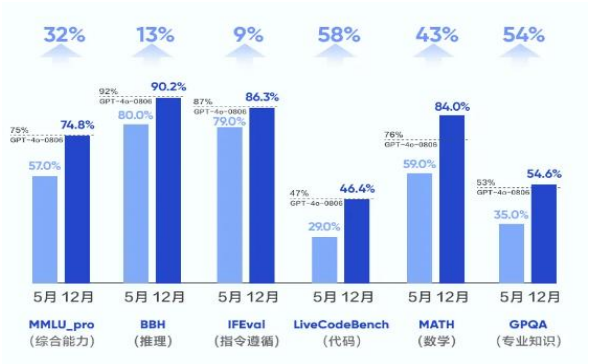
图20：火山引擎多模态数据湖解决方案



资料来源：火山引擎，国信证券经济研究所整理

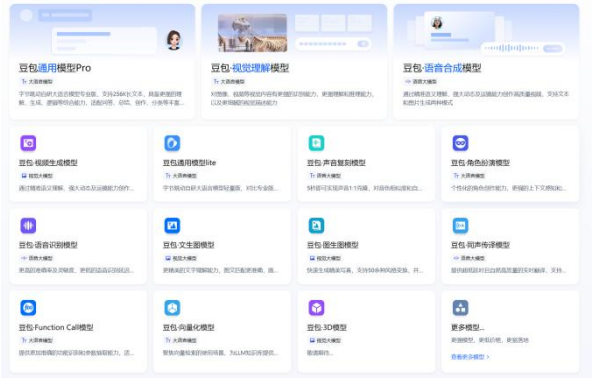
字节跳动模型能力持续提升，产品矩阵愈加丰富。1）模型能力持续提升：豆包通用模型 pro 完成新版本迭代，综合任务处理能力较 5 月份提升 32%，在推理上提升 13%，在指令遵循上提升 9%，在代码上提升 58%，在数学上提升 43%，在专业知识领域能力提升 54%。豆包通用模型 Pro 已全面对齐 GPT-4o 的能力，但使用价格远低于后者，推理输入价格为 0.0008 元/千 tokens。**2）产品矩阵愈加丰富：**除豆包通用模型 Pro 外，字节新发布视觉理解模型，模型能够综合理解用户给出的文本和图像信息，并给出准确的回答，在金融、医疗、教育、旅游等诸多行业有广阔的应用前景，且模型输入价格仅为 0.003 元/千 tokens，比行业价格便宜 85%；此外，豆包音乐模型、文生图模型持续迭代升级。

图21：字节跳动模型能力持续提升



资料来源：字节跳动官网，国信证券经济研究所整理

图22：字节跳动模型产品矩阵愈加丰富

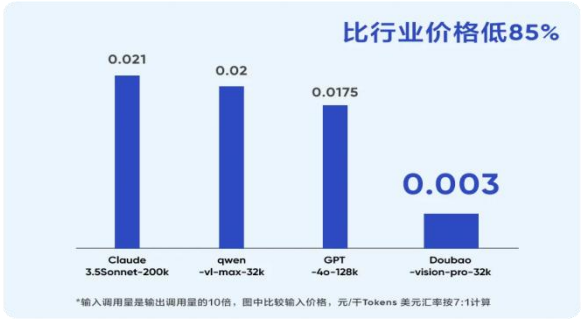


资料来源：字节跳动官网，国信证券经济研究所整理

图23：字节跳动视觉理解模型能力增强，场景拓宽

图24：字节跳动视觉理解模型价格远低于同行

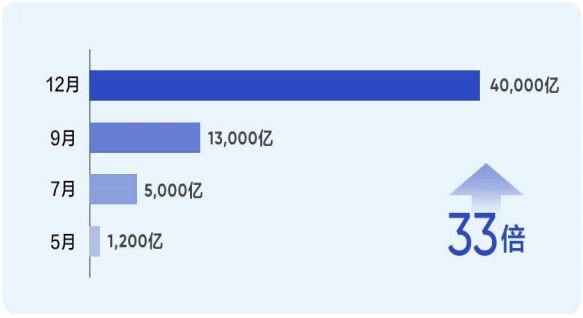
资料来源：火山引擎，国信证券经济研究所整理



资料来源：火山引擎，国信证券经济研究所整理

豆包大模型 API 调用量迅速提升，多场景快速渗透。1）API 调用量快速提升：截至 12 月中旬，豆包通用模型的日均 tokens 使用量已超过 4 万亿，较五月首次发布时日均 1200 亿增长了 33 倍；目前，豆包大模型已经与八成主流汽车品牌合作，并接入到多家手机、PC 等智能终端，覆盖终端设备约 3 亿台，来自智能终端的豆包大模型调用量在半年时间内增长 100 倍。**2）多场景快速渗透**：最近 3 个月，豆包大模型在信息处理场景的调用量增长了 39 倍，帮助企业更好的分析和处理内部及外部的数据；客服与销售场景增长 16 倍，帮助企业更好的服务客户，扩大销售；硬件终端场景增长 13 倍，AI 工具场景增长 9 倍，学习教育等场景也有大幅增长。

图25：豆包大模型 API 调用量迅速提升



资料来源：火山引擎，国信证券经济研究所整理

图26：豆包大模型在多场景迅速渗透

资料来源：火山引擎，国信证券经济研究所整理

豆包 API 调用收入快速增长，拉动推理算力需求提升。我们的计算基于以下假设：1）目前豆包 Pro 大模型能力已全方位对齐 ChatGPT-4o，我们假设豆包通用大模

型的参数量已达到万亿以上，与 ChatGPT 参数量齐平，目前万亿模型普遍采用 MoE 架构，我们假设每次调用 API 时使用的模型参数为 1100 亿；2）数据中心算力调用情况在一天内可出现多次波峰及波谷，我们假设算力设备投建时，所需的算力容量为平均需求的 3 倍；3）假设数据中心中，单卡平均利用率为 45%；4）假设豆包的所有模型 API 调用平均价格为 1.5 元/百万 Tokens；5）假设未来 API 调用将有 40%为付费调用，剩余 60%为免费用户的调用；6）目前国外领先云服务商如微软、Meta 等普遍预期计算设备折旧年限为 6 年，我们假设字节跳动数据中心折旧年限同为 6 年。基于上述假设，我们预期在日均 100 万亿 Tokens 调用量时，公司年收入可达 219 亿元。

图27：豆包 API 调用算力需求及收入（基于日均 100 万亿 API 调用

假设日均API调用量为100万亿Tokens	
通用模型	数据
豆包通用模型参数量（亿）	1100
预期日均Tokens调用量（亿）	1000000
每日推理总时长（秒）	86400
平均每秒调用Tokens（亿）	11.57
假设峰值算力需求/平均需求（倍）	3
推理所需理论算力=2*模型参数量*每秒调用Tokens/100（EFLOPS）	763.89
H800单卡算力（TFLOPS）	1979
假设单卡利用率（%）	45%
实际需要卡数（万张）	**85.78**
API调用平均单价（元/百万Tokens）	1.5
API调用付费率（%）	40%
日均API调用收入（亿元）	0.6
年收入（亿元）	**219**
单张H800芯片价格（万元）	15
显卡折旧年限（年）	6
年均算力投入（亿元）	**214.44**

资料来源：火山引擎、英伟达，国信证券经济研究所整理

图28：豆包 API 调用算力需求及收入（基于日均 1000 万亿 API 调用

假设日均API调用量为1000万亿Tokens	
通用模型	数据
豆包通用模型参数量（亿）	1100
预期日均Tokens调用量（亿）	10000000
每日推理总时长（秒）	86400
平均每秒调用Tokens（亿）	115.74
假设峰值算力需求/平均需求（倍）	3
推理所需理论算力=2*模型参数量*每秒调用Tokens/100（EFLOPS）	7638.89
H800单卡算力（TFLOPS）	1979
假设单卡利用率（%）	45%
实际需要卡数（万张）	**857.77**
API调用平均单价（元/百万Tokens）	1.5
API调用付费率（%）	40%
日均API调用收入（亿元）	6
年收入（亿元）	**2190**
单张H800芯片价格（万元）	15
显卡折旧年限（年）	6
年均算力投入（亿元）	**2144.43**

资料来源：火山引擎、英伟达，国信证券经济研究所整理

1 月 20 号，豆包实时语音大模型正式推出，并在豆包 APP 全量开放，将豆包 APP 升级至 7.2.0 版本即可体验。据字节的模型优点评测，豆包实时语音大模型在情绪理解和情感表达方面与 GPT-4O 相比优势明显。尤其是“一听就是 AI 与否”评测中，超过 30%的反馈表示 GPT-4o“过于 AI”，而豆包实时语音大模型相应比例仅为 2%以内。尤其情商层面，模型在情感理解、情感承接以及情感表达等方面也取得显著进展，能较为准确地捕捉、回应人类情感信息。

小米：MiLM2 升级发布，性能与技术全面提升，着手搭建万卡算力集群

2024 年 11 月，**小米大模型已经实现了从一代到二代（MiLM2）的升级迭代**。主要升级包括：

丰富了模型的参数矩阵，参数规模同时向下和向上扩充，实现了云边端结合，参数尺寸最小为 0.3B，最大为 30B；

在 10 大能力维度上，相比于第一代模型平均提升超过 45%，其中指令跟随、翻译、闲聊等关键能力处于业界前列；

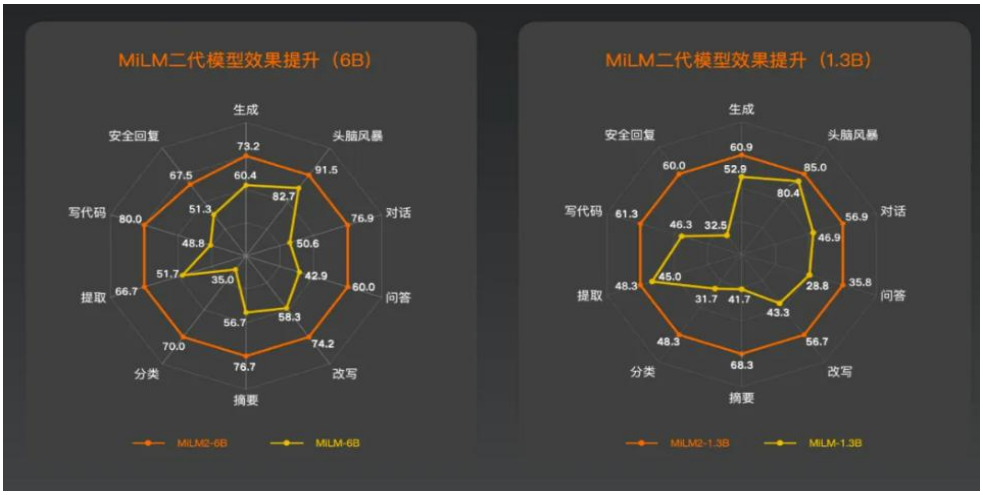
在端侧部署上支持 3 种推理加速方案，包括大小模型投机、BiTA、Medusa，并且量化损失降低 78%；

支持的最长窗口为 200k（第一代为 4k），在长文本评测中，效果处于业界前列。

小米大模型团队采用自主构建的通用能力评测集 Mi-LLMBM2.0，对最新一代的

MiLM2 模型进行了全方位评估。该评测集涵盖了广泛的应用场景，包括生成、脑暴、对话、问答、改写、摘要、分类、提取、代码处理以及安全回复等 10 个大类，共计 170 个细分测试项。以 MiLM2-1.3B 模型和 MiLM2-6B 模型为例，对比去年发布的一代模型，在十大能力上的效果均有大幅提升，平均提升幅度超过 45%。

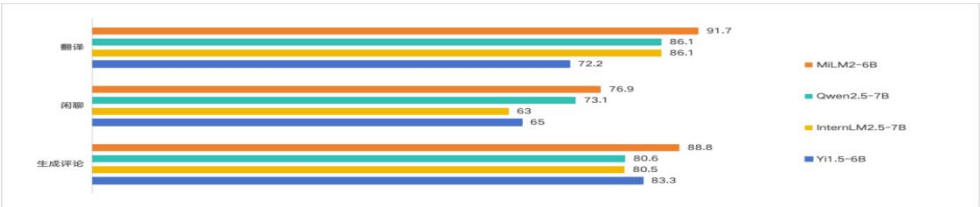
图29：MiLM 二代模型效果提升图



资料来源：新浪财经，国信证券经济研究所整理

小米的「人车家全生态」战略，旨在构建一个涵盖人、车、家等多元化生活场景的超级智能生态系统。在这个系统内，实时交互成为常态，每时每刻都需要精确对接用户千差万别的个性化需求，这对于大模型的生成、闲聊、翻译等能力提出了更高的要求。在这些关键能力上，MiLM2-6B 模型的评测成绩十分优异，对比业内同参数规模模型也有较优的效果。

图30：生成、闲聊、翻译能力对比图



资料来源：新浪财经，国信证券经济研究所整理

在坚持轻量化部署的大原则下，小米自研大模型团队构建并不断扩充了自研大模型的模型矩阵，将大模型的参数规模灵活扩展至 0.3B、0.7B、1.3B、2.4B、4B、6B、13B、30B 等多个量级，充分考虑多元化的业务场景及资源限制。

0.3B~6B：终端（on-device）场景，应用时通常是一项非常具体的、低成本的任务，微调后可以达到百亿参数内开源模型效果。

6B、13B：支持多任务微调，微调后可以达到几百亿开源模型的效果。

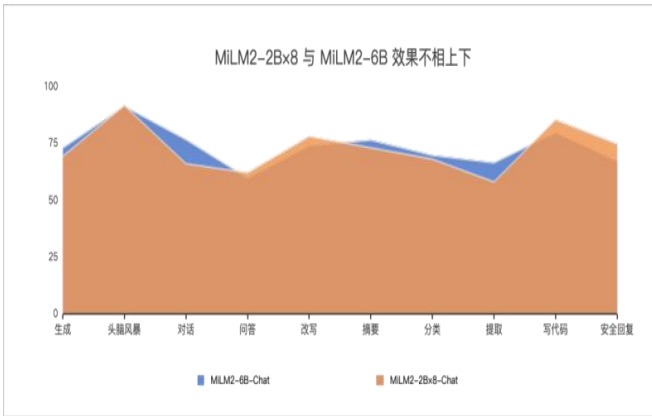
30B：云端场景，具备坚实的 zero-shot/上下文学习或一些泛化能力，模型推理能力较好，能够完成复杂的多任务。

小米自研大模型矩阵不仅包含多样的参数量级，同时也纳入了各种不同的模型结构。在二代模型系列中，大模型团队加入了两个 MoE 结构的模型：MiLM2-0.7B×8 和 MiLM2-2B×8。以 MiLM2-2B×8 为例，根据评测结果，该模型在整体性能上与 MiLM2-6B 不相上下，而解码速度实现了 50%的提升，提升了其运行效率。

图31：小米第二段自研大模型 MiLM2 模型矩阵

资料来源：新浪财经，国信证券经济研究所整理

图32：MiLM2-2B×8 与 MiLM2-6B 效果对比



资料来源：新浪财经，国信证券经济研究所整理

据界面新闻，**小米正在着手搭建 GPU 万卡集群**，大力投入 AI 大模型。小米大模型团队在成立时已拥有 6500 张 GPU 资源，据小米 2023 周年演讲，小米大模型技术主力突破方向为轻量化、本地部署。

阿里：服务器芯片倚天 710 已成功落地规模；双 11 提供了 100 万核的算力

在 24 年 11 月 19 日的 Arm 年度技术大会上，**阿里巴巴宣布其自研的基于 Arm 架构的服务器芯片倚天 710 已成功落地规模，当前已实现数百万核的应用，客户数也已超过 1000。**这一重大进展不仅展现了阿里在芯片研发领域的雄心，也彰显了其在数据库、大数据以及视频等领域的商用实力，标志着中国在高性能计算领域的进一步崛起。

图33：倚天 710

资料来源：中电网公众号，国信证券经济研究所整理

技术特点方面，倚天 710 使用了多项先进的技术创新，包括多核处理器和高效的缓存管理，这使得其在处理复杂任务时的稳定性和速度都具备明显优势。芯片支持的众多现代编程模型与生态系统，也将进一步丰富用户的选择，降低开发门槛。使用倚天 710 进行数据库、大数据处理或者视频服务时，用户能够享受到更大规模的计算能力与小于传统芯片的延迟，优化全链路的用户体验。

与一般服务器芯片相比，阿里的倚天 710 在硬件级别的资源调度与管理表现出色，使得企业在面对不断膨胀的数据量时，能够更加从容应对。此外，芯片具备支持 AI 计算的能力，非常适合深度学习、自然语言处理等需求苛刻的应用场景。

今年，阿里云首度在公共云上为天猫双 11 提供了超过 100 万核 CPU 的算力资源支撑，弹性规模刷新纪录。

图34：阿里云工作示意图

资料来源：阿里云招聘公众号，国信证券经济研究所整理

阿里云通过一系列技术创新，打破传统模式，无需预购算力，按需即取，首次将

云计算的即时弹性提升至百万核级的极限水平，并首次大规模应用了云基础设施处理器 CIPU。

阿里云自研云基础设施处理器 CIPU，向上接入飞天云操作系统，向下统一调度并优化计算、存储、网络等资源，将全球数百万台服务器连成一台超级计算机，按需调配双 11 所需的算力资源。

基于公共云底座，阿里云为双 11 全新打造了统一的算力需求与资源匹配调度引擎，可高效调度张北、南通、北京、上海、新加坡、德国等全球多地的公共云算力，而且用完即释放。

经估算，此举可让双 11 的弹性成本整体节省 25-50%左右。再加上算力需的时域错峰，算力闲置的成本也降低了 8%。

腾讯：星脉 2.0 全面升级，专为 10 万级 GPU 的网络通信而生

24 年 7 月 1 日，腾讯宣布了星脉网络的全新升级。作为一套软硬协同的高性能网络体系，星脉网络 2.0 的**核心目标是通过高性能、自研的网络设备、自研通信协议 TiTa、集合通信库 TCCL 和全栈网络运营系统**，打造一个高效、稳定的计算环境，以支持万亿级参数规模的 AI 大模型训练。

图35：星脉网络

资料来源：53AI 知识库，国信证券经济研究所整理

自研网络设备，交换机容量、光模块速率都提升一倍。在硬件方面，腾讯星脉网络 2.0 进行了显著升级。交换机的容量从原来的 25.6T 提升至 51.2T，大大增加了数据传输的容量。光模块的速率从 200G 升级到 400G，显著降低了网络延迟，提升了数据传输速度。同时，CNIC 网卡作为公有云业内首款为 AI 训练设计的网卡，整卡带宽达 400Gbps，具备 3.2T 整机通信带宽。这些硬件升级不仅提升了通信效率，还减少了网络拥塞，显著提高了整体网络性能。

自研通信协议 TiTa，采用主动拥塞控制算法，在拥塞发生前就进行调控。为了提升通信效率，腾讯自研的 TiTa 协议采用了主动拥塞控制算法。这种算法通过端侧网卡主动感知并调整数据包发送速率，从而在拥塞发生前就进行调控，避免网络性能大幅下降。相比传统被动拥塞控制算法，TiTa 协议能够更有效地避免网络拥

堵，减少数据包丢失，提高网络吞吐量，优化数据包发送速率，降低通信过程中的延迟和网络拥堵。

集合通信库 TCCL，实现了 GPU 间数据的高效传输。集合通信库 TCCL 通过 NVLINK+NET 异构并行通信技术，实现了 GPU 间数据的高效传输。每个 GPU 网卡构建了独立的网络通道，实现数据并行传输，间接提升了传输链路的带宽。此外，Auto-TuneNetworkExpert 自适应算法能够根据不同的机型、网络规模、模型算法和数据包大小等因素，动态调整网络参数，确保在各种场景下实现最优性能。这些优化使得 TCCL 通信库不仅提升了数据传输带宽和速度，还通过自适应算法根据不同场景优化网络参数，提高了资源利用率，减少了资源浪费。

灵境仿真平台，将 GPU 故障定位时间从传统的天级缩短到分钟级。腾讯的灵境仿真平台作为网络运营系统的一部分，能够收集训练过程中的日志和 GPU 相关信息，通过仿真模拟还原训练任务，定位训练中的卡顿和性能抖动问题，这一功能使得问题定位时间从传统的天级缩短到分钟级。此外，全栈网络运营系统在星脉网络 2.0 中得到了全面升级，并提供 360 度无死角的立体监控，能够更快发现和定位网络问题，并快速修复故障，确保训练任务的连续性。这种全方位的监控和快速修复能力，显著提升了训练的稳定性和高可用性。

星脉网络 2.0 的四个关键组件相互协同配合，共同共同提升大模型训练过程中的网络性能。可以用赛车来类比：如果将调度 GPU 集群训练大模型比作赛车，目标是通过赛场软硬件系统的升级来发挥最大性能。硬件（交换机、光模块）相当于赛道，升级后的带宽提升至 3.2T，如同拓宽和改善了赛道，增加宽度和容量。TiTa 协议如同赛事指挥中心，智能化调控“车速”，避免拥堵。TCCL 通信库如同专业车队管理系统，通过 NVLINK+NET 异构并行通信和自适应算法优化赛车性能。运营系统则如同专业的抢修队，全方位监控和修复故障，确保比赛顺利进行。

图36：星脉网络 2.0

资料来源：53AI 知识库，国信证券经济研究所整理

DeepSeek：多层面技术提升训练效率，测试性能领跑开源模型

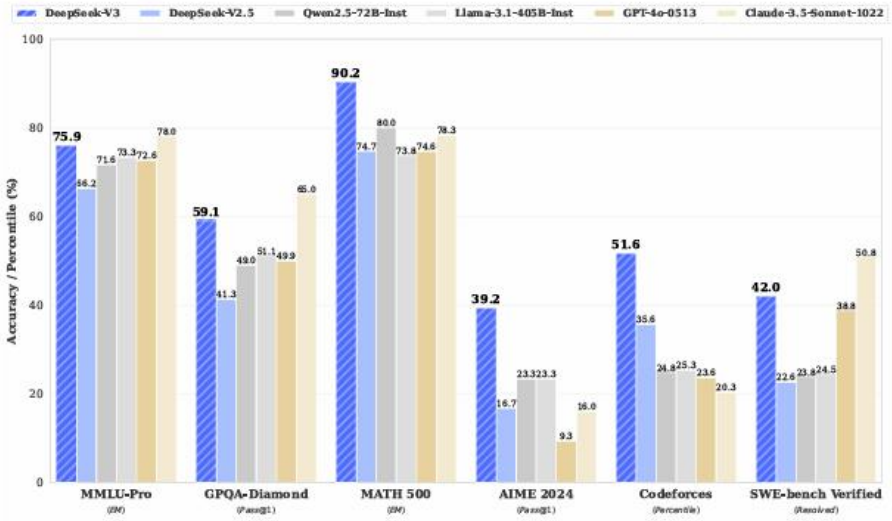
2024 年 12 月 26 日，DeepSeek 宣布上线并同步开源 DeepSeek-V3 模型，多项评测成绩超 Qwen2.5-72B 和 Llama-3.1-405B 等开源模型，并在数学、代码等重要领域测评表现与 GPT-4o 和 Claude-3.5-Sonnet 等顶尖闭源模型相当，训练成本仅为 557.6 万美元。

模型层：在多项测评中能力领先其他开源模型

模型采用 MoE 架构，运用 R1 提炼推理能力。据 DeepSeek-V3 Technical Report，DeepSeek-V3（以下简称模型）采用 MoE（专家混合模型）架构，总参数量达 6710 亿，每个 Token 激活 37 亿参数。在预训练阶段，模型使用 14.8T 的高质量数据集进行训练，并在后续对模型进行了两个阶段的上下文长度扩展，第一阶段将上下文长度扩展至 32K，第二阶段进一步扩展至 128K。在后训练部分，DeepSeek 使用了监督微调（SFT）和强化学习（RL）来提升模型能力，并从 DeepSeek R1 系列模型中提炼推理能力，以提高模型在实际运用中的表现。

在实际测评中，DeepSeek 表现领先开源模型，并进一步缩小与闭源模型的差距。1）知识层面：在教育类基准测试如 MMLU、MMLU-Pro 上，模型超越了目前所有开源模型，其表现与领先的闭源模型如 GPT-4o 和 Claude-Sonnet-3.5 相当。在常识性测试中，模型在 SimpleQA 和 Chinese SimpleQA 中的评分领先，在英语事实性知识（SimpleQA）方面稍逊于 GPT-4o 和 Claude-Sonnet-3.5；2）代码、数学与推理层面：在数学相关测试中，模型在所有非思维链推理模型中表现出色，在 MATH-500 等特定测试中超越了 o1-preview。模型在编程竞赛基准测试中表现领先，在工程类任务中模型表现略逊于 Claude-Sonnet-3.5。

图37：DeepSeek-V3 在各项测试中表现领先



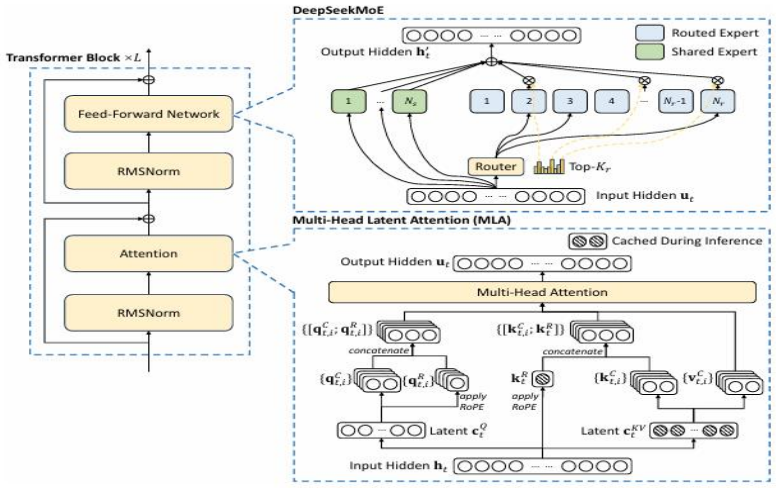
资料来源：DeepSeek 著-《DeepSeek-V3 Technical Report》-Github（2025）-P1，国信证券经济研究所整理

架构层：基本沿用 V2 架构，引入 MTP 等全新技术

沿用 V2 基本架构，引入无辅助损失的负载均衡策略。模型在架构层面沿用 V2 模

型中的多头潜在注意力（MLA）以及 DeepSeekMoE 架构，以实现经济训练。在此基础上，模型引入了无辅助损失的负载均衡策略，以减少因负载均衡所带来的性能下降。MLA 架构用于减少注意力键值（KV）缓存时的空间占用，通过对注意力键和值进行低秩压缩来实现，帮助模型在维持性能的同时减少计算资源的消耗。

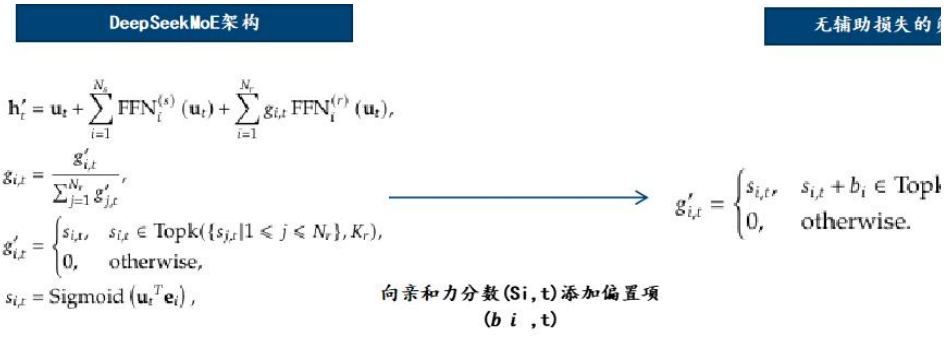
图38：DeepSeek-V3 模型架构



资料来源：DeepSeek 著-《DeepSeek-V3 Technical Report》-Github（2025）-P7，国信证券经济研究所整理

DeepSeekMoE 较传统 MoE 有多方面改进。与传统 MoE 架构相比，DeepSeekMoE 使用了更细粒度的专家，并将部分专家设置为共享专家，能够更精确地针对特定的问题提供解决方案。同时，传统 MoE 架构采用辅助损失来鼓励负载平衡，以免不平衡的专家载荷导致计算效率降低，但这可能在某些情况下影响模型性能。DeepSeekMoE 引入了无辅助损失的负载平衡策略，在每个专家模型的任务匹配程度评分中添加一个偏置项，用于调整每个专家在决定哪些专家应该处理哪些任务时的负载，同时使用补充序列级辅助损失，以此来优化整个系统的性能和效率。

图39：DeepSeekMoE 引入无辅助损失的负载平衡策略



资料来源：DeepSeek 著-《DeepSeek-V3 Technical Report》-Github（2025）-P8、P9，国信证券经济研究所整理

模型通过 MTP 提升数据利用效率。DeepSeek 在模型训练时设定了多 Token 预测（MTP）目标，将预测范围扩展到每个位置的多个未来 Token。MTP 增加了模型训练过程中的信号密度，提高模型对数据的整体利用效率，同时增强模型生成文本的连贯性。MTP 策略主要旨在提高主模型的性能，在推理过程中可直接丢弃 MTP 模块，主模型独立正常运行。

训练层：通过工程优化，进一步实现成本控制

为了促进模型的高效训练，DeepSeek 实施了工程优化。首先，模型使用了 DualPipe 算法，以实现高效的管道并行。与现有方法相比，DualPipe 具有更少的管道气泡（等待数据处理或通信延迟形成的停滞区域），在模型训练的前向和后向传播过程实现了重叠计算和通信，从而提高了整体的训练效率。其次，DeepSeek 引入了 FP8 混合精度训练，优化了训练期间的内存占用。

DualPipe 技术优化通信成本。在大规模分布式训练系统中，每个计算节点需要频繁地与其他节点交换信息，导致部分时间在等待数据的传输，计算资源不能持续进行数据处理，资源利用率低下。以 DeepSeek-V3 为例，在模型训练时跨节点的专家并行性带来的通信开销导致计算与通信的比率约为 1:1。为了解决这一问题，DeepSeek 在单独的前向和后向块内部重叠计算和通信，通过采用双向管道调度，同时从管道的两端供给数据，使大部分通信可以完全重叠，从而实现通信成本的降低。

图40：DualPipe 技术极大优化通信效果

DualPipe调度实例

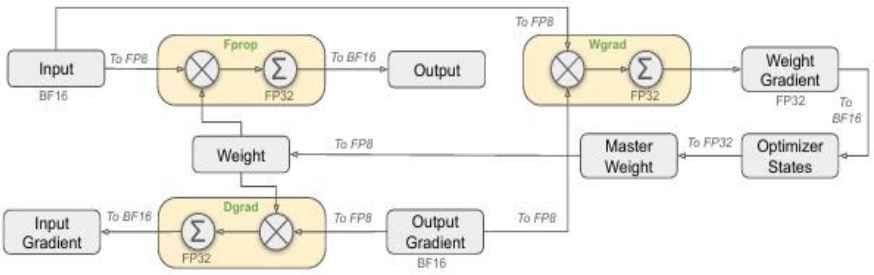
不同的管道并行方法对比

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

资料来源：DeepSeek 著-《DeepSeek-V3 Technical Report》-Github（2025）-P13，国信证券经济研究所整理

使用 FP8 数据格式提升训练效率。DeepSeek 在训练模型时提出了使用 FP8 数据格式的细粒度混合精度框架，大部分计算密集型操作在 FP8 精度下进行，少数关键操作保持在原始的数据格式中。这种设计理论上使计算速度比传统的 BF16 方法提高了一倍。同时，使用 FP8 精度存储可以减少内存需求，使得训练过程更加高效。然而，低精度计算可能会引入更多的数值不稳定性和精度损失，为了解决这一问题，DeepSeek 在 GEMM（通用矩阵乘法）操作的内部维度引入了每组缩放因子，根据较小的元素组调整缩放比例，使量化过程能更好地适应离群值，从而保障了低精度训练结果的准确性。

图41：混合精度框架使用 FP8 格式



资料来源：DeepSeek 著-《DeepSeek-V3 Technical Report》-Github（2025）-P15，国信证券经济研究所整理

推理层：将 R1 推理能力迁移至模型中

推出类 o1 推理模型，为 V3 模型提供基础。2024 年 11 月 20 日，DeepSeek 发布 DeepSeek-R1-Lite，R1 系列模型使用强化学习训练，推理过程包含大量反思和验证，思维链长度可达数万字，在数学、代码以及各种复杂逻辑推理任务上，取得了与 o1-preview 相似水准的推理效果。目前该系列模型仍处于迭代开发阶段，正式版 DeepSeek-R1 模型技术仍未开源。

图42：R1 预览版取得与 o1-preview 媲美的性能

	DeepSeek-R1-Lite-Preview	OpenAI o1-preview	GPT-4o	Claude 3.5 Sonnet	Qwen-2.5-72B-Instruct	DeepSeek V2.5
AIME (pass@1) 美国数学竞赛	**52.5**	44.6	9.3	16.0	23.3	16.7
MATH-500 (greedy) 美国数学竞赛	**91.6**	85.5	76.6	78.3	83.1	74.7
GPQA Diamond (pass@1) 理工科博士生测试	58.5	**73.3**	53.6	65.0	49.0	41.3
Codeforces (Rating) 世界级编程竞赛	**1450**	1428	759	717	732	882
LiveCodeBench (2024.8-2024.11) 世界级编程竞赛	51.6	**53.6**	33.4	36.3	31.1	29.2
Zebra Logic 自然语言解谜	56.6	**71.4**	28.2	33.4	26.6	22.1

*所有测评在最大推理长度 32K 下得到，测试结果通过测试集重复测试多次求平均得到，避免温度带来的随机影响。

资料来源：DeepSeek 官微，国信证券经济研究所整理

以 R1 推理能力为底座，将能力迁移至 V3 中。在后训练部分中，对于推理相关的数据集，DeepSeek 利用内部的 DeepSeek-R1 模型生成数据。DeepSeek 首先开发一个专门针对特定领域（如编程、数学或一般推理）的专家模型，使用结合了监督式精调（SFT）和强化学习（RL）的训练管道，使用这个专家模型作为最终模型的数据生成器。这种方法确保了最终训练数据保留了 DeepSeek-R1 的优势，使用此训练数据训练的 V3 模型能够极大的提升自身的推理能力。对于非推理数据，DeepSeek 使用 DeepSeek-V2.5 生成响应，并招募人类注释员来验证数据的准确性和正确性。通过从 DeepSeek-R1 系列模型中提取推理能力，V3 模型实现了在数学、编程等领域性能上的提升。

推理算力包含 GB300、博通、marvell 等各类 asic 芯片。

图43：通过迁移 R1 推理能力提升 V3 模型性能

后训练前

	Benchmark (Metric)	# Shots	DeepSeek-V2 Base	Qwen2.5 72B Base	LLaMA-3.1 405B Base	DeepSeek-V3 Base
	Architecture	-	MoE	Dense	Dense	MoE
	# Activated Params	-	21B	72B	405B	37B
	# Total Params	-	236B	72B	405B	671B
English	Pile-test (BPB)	-	0.606	0.638	**0.542**	0.548
	BBH (EM)	3-shot	78.8	79.8	82.9	**87.5**
	MMLU (EM)	5-shot	78.4	85.0	84.4	**87.1**
	MMLU-Redux (EM)	5-shot	75.6	83.2	81.3	**86.2**
	MMLU-Pro (EM)	5-shot	51.4	58.3	52.8	**64.4**
	DROP (F1)	3-shot	80.4	80.6	86.0	**89.0**
	ARC-Easy (EM)	25-shot	97.6	98.4	98.4	**98.9**
	ARC-Challenge (EM)	25-shot	92.2	94.5	**95.3**	**95.3**
	HellaSwag (EM)	10-shot	87.1	84.8	**89.2**	**88.9**
	PIQA (EM)	0-shot	83.9	82.6	**85.9**	84.7
	WinoGrande (EM)	5-shot	**86.3**	82.3	85.2	84.9
	RACE-Middle (EM)	5-shot	73.1	68.1	**74.2**	67.1
	RACE-High (EM)	5-shot	52.6	50.3	**56.8**	51.3
	TriviaQA (EM)	5-shot	80.0	71.9	**82.7**	**82.9**
	NaturalQuestions (EM)	5-shot	38.6	33.2	**41.5**	40.0
	AGIEval (EM)	0-shot	57.5	75.8	60.6	**79.6**
Code	HumanEval (Pass@1)	0-shot	43.3	53.0	54.9	**65.2**
	MBPP (Pass@1)	3-shot	65.0	72.6	68.4	**75.4**
	LiveCodeBench-Base (Pass@1)	3-shot	11.6	12.9	15.5	**19.4**
	CRUXEval-I (EM)	2-shot	52.5	59.1	58.5	**67.3**
	CRUXEval-O (EM)	2-shot	49.8	59.9	59.9	**69.8**
Math	GSM8K (EM)	8-shot	81.6	88.3	83.5	**89.3**
	MATH (EM)	4-shot	43.4	54.4	49.0	**61.6**
	MGSM (EM)	8-shot	63.6	76.2	69.9	**79.8**
	CMath (EM)	3-shot	78.7	84.5	77.3	**90.7**
Chinese	CLUEWSC (EM)	5-shot	82.0	82.5	**83.0**	**82.7**
	C-Eval (EM)	5-shot	81.4	89.2	72.5	**90.1**
	CMMLU (EM)	5-shot	84.0	**89.5**	73.7	88.8
	CMRC (EM)	1-shot	**77.4**	75.8	76.0	76.3
	C3 (EM)	0-shot	77.4	76.7	**79.7**	78.6
	CCPM (EM)	0-shot	**93.0**	88.5	78.6	92.0
Multilingual	MMMLU-non-English (EM)	5-shot	64.0	74.8	73.8	**79.4**

后训练后

	Benchmark (Metric)	DeepSeek V2-0506	DeepSeek V2.5-0905	Qwen2.5 72B-Inst.	LLaMA-3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3
	Architecture	MoE	MoE	Dense	Dense	-	-	MoE
	# Activated Params	21B	21B	72B	405B	-	-	37B
	# Total Params	236B	236B	72B	405B	-	-	671B
English	MMLU (EM)	78.2	80.6	85.3	**88.6**	**88.3**	87.2	**88.5**
	MMLU-Redux (EM)	77.9	80.3	85.6	86.2	**88.9**	88.0	**89.1**
	MMLU-Pro (EM)	58.5	66.2	71.6	73.3	**78.0**	72.6	75.9
	DROP (3-shot F1)	83.0	87.8	76.7	88.7	88.3	83.7	**91.6**
	IF-Eval (Prompt Strict)	57.7	80.6	84.1	86.0	**86.5**	84.3	86.1
	GPQA-Diamond (Pass@1)	35.3	41.3	49.0	51.1	**65.0**	49.9	59.1
	SimpleQA (Correct)	9.0	10.2	9.1	17.1	28.4	**38.2**	24.9
	FRAMES (Acc.)	66.9	65.4	69.8	70.0	72.5	**80.5**	73.3
	LongBench v2 (Acc.)	31.6	35.4	39.4	36.1	41.0	48.1	**48.7**
Code	HumanEval-Mul (Pass@1)	69.3	77.4	77.3	77.2	81.7	80.5	**82.6**
	LiveCodeBench (Pass@1-COT)	18.8	29.2	31.1	28.4	36.3	33.4	**40.5**
	LiveCodeBench (Pass@1)	20.3	28.4	28.7	30.1	32.8	34.2	**37.6**
	Codeforces (Percentile)	17.5	35.6	24.8	25.3	20.3	23.6	**51.6**
	SWE Verified (Resolved)	-	22.6	23.8	24.5	**50.8**	38.8	42.0
	Aider-Edit (Acc.)	60.3	71.6	65.4	63.9	**84.2**	72.9	79.7
	Aider-Polyglot (Acc.)	-	18.2	7.6	5.8	45.3	16.0	**49.6**
Math	AIME 2024 (Pass@1)	4.6	16.7	23.3	23.3	16.0	9.3	**39.2**
	MATH-500 (EM)	56.3	74.7	80.0	73.8	78.3	74.6	**90.2**
	CNMO 2024 (Pass@1)	2.8	10.8	15.9	6.8	13.1	10.8	**43.2**
Chinese	CLUEWSC (EM)	89.9	90.4	**91.4**	84.7	85.4	87.9	90.9
	C-Eval (EM)	78.6	79.5	86.1	61.5	76.7	76.0	**86.5**
	C-SimpleQA (Correct)	48.5	54.1	48.4	50.4	51.3	59.3	**64.8**

资料来源：DeepSeek 著-《DeepSeek-V3 Technical Report》-Github（2025）-P25、P31，国信证券经济研究所整理

2025 年 1 月，DeepSeek 再次发力，发布推理模型 DeepSeek-R1。该模型在 Codeforces、MATH-500 等多个第三方测试中超越 OpenAI 的最新模型 o1，在数学、编程和常识等任务上取得优秀成绩，API 调用成本仅为后者的 1/10（每百万 Token1 元￥）。DeepSeek-R1 具备强大的推理和分析能力，采用与 DeepSeek-V3 相同的专家混合（MoE）架构，在提升性能的同时，保持了较低的成本优势。

2025 年 1 月 20 日晚，拥有 660B 参数的超大规模模型 DeepSeek R1 正式发布。

这款模型在数学任务上表现出色，如在 AIME 2024 上获得 79.8% 的 pass@1 得分，略超 OpenAI-o1；在 MATH-500 上得分高达 97.3%，与 OpenAI-o1 相当。

编程任务方面，如 Codeforces 上获得 2029 Elo 评级，超越 96.3%的人类参与者。在 MMLU、MMLU-Pro 和 GPQA Diamond 等知识基准测试中，DeepSeek R1 得分分别为 90.8%、84.0% 和 71.5%，虽略低于 OpenAI-o1，但优于其他闭源模型。

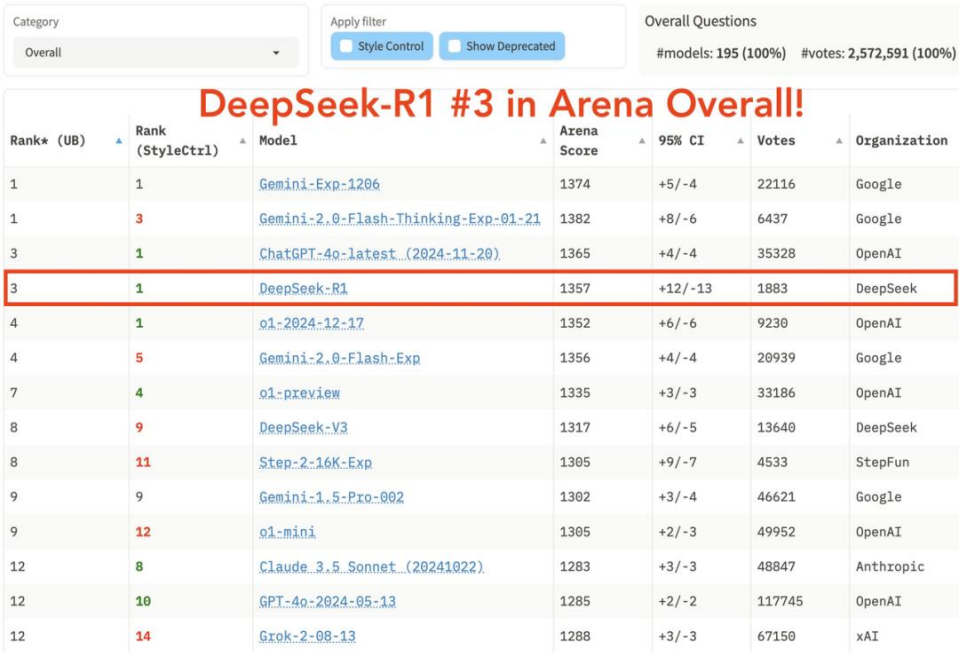
在最新公布的大模型竞技场 LM Arena 的综合榜单中，DeepSeek R1 排名第三，与 o1 并列。

在「Hard Prompts」（高难度提示词）、「Coding」（代码能力）和「Math」（数学能力）等领域，DeepSeek R1 位列第一。

在「Style Control」（风格控制）方面，DeepSeek R1 与 o1 并列第一。

在「Hard Prompt with Style Control」（高难度提示词与风格控制结合）的测试中，DeepSeek R1 也与 o1 并列第一。

图44：DeepSeek-R1 测试排行



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization
1	1	Gemini-Exp-1206	1374	+5/-4	22116	Google
1	3	Gemini-2.0-Flash-Thinking-Exp-01-21	1382	+8/-6	6437	Google
3	1	ChatGPT-4o-latest (2024-11-20)	1365	+4/-4	35328	OpenAI
3	1	DeepSeek-R1	1357	+12/-13	1883	DeepSeek
4	1	o1-2024-12-17	1352	+6/-6	9230	OpenAI
4	5	Gemini-2.0-Flash-Exp	1356	+4/-4	20939	Google
7	4	o1-preview	1335	+3/-3	33186	OpenAI
8	9	DeepSeek-V3	1317	+6/-5	13640	DeepSeek
8	11	Step-2-16K-Exp	1305	+9/-7	4533	StepFun
9	9	Gemini-1.5-Pro-002	1302	+3/-4	46621	Google
9	12	o1-mini	1305	+2/-3	49952	OpenAI
12	8	Claude 3.5 Sonnet (20241022)	1283	+3/-3	48847	Anthropic
12	10	GPT-4o-2024-05-13	1285	+2/-2	117745	OpenAI
12	14	Grok-2-08-13	1288	+3/-3	67150	xAI

资料来源：APPSO 公众号，国信证券经济研究所整理

AI Agent 应用：海内外 AI 应用百家争鸣，各领域百花齐放

AI+企业服务

ServiceNow：核心产品是 Now Platform，这是一个基于云的企业工作流平台，目前该平台支持以下几大核心领域：IT 服务管理（ITSM）、IT 运维管理（ITOM）、客户服务管理（CSM）：、人力资源服务管理（HRSM）、安全运营、应用开发平台。

AI 逻辑：ServiceNow 的 Now Platform 的 AI 功能提高工作流程软件效率，够帮助开发者、管理员、用户、客服、公司员工等多方实现提效，如问答交互、协调运营、AI 主动学习与即时决策、定制化运营模式，从而提高产品付费率和 ARPU。

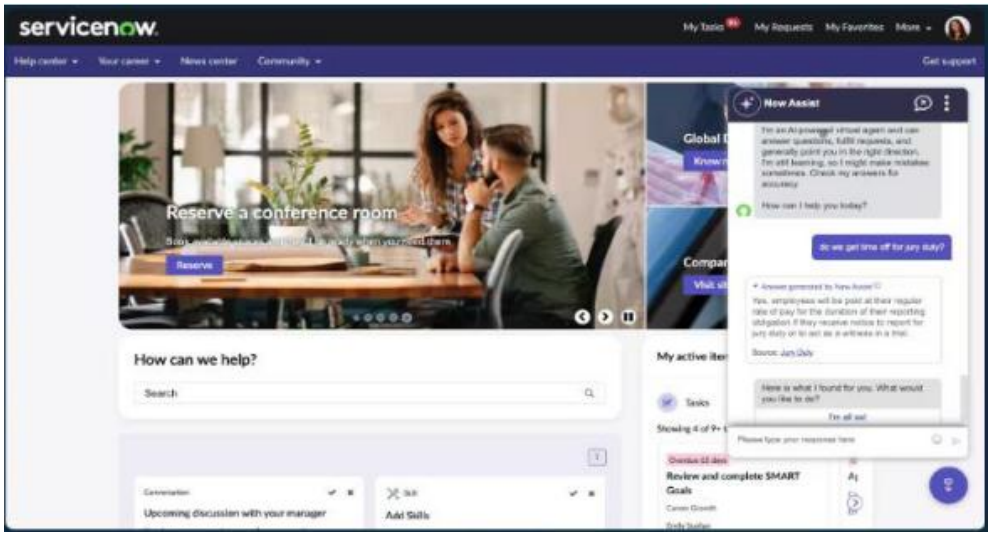
1）Now Assist：于 2023 年 6 月推出，整合了此前发布的生成式 AI 功能：包括内容总结生成、对话交流、代码生成、虚拟客服、AI 搜索和生成式 AI 控制器等，

同时面向企业发布了大模型 Now LLM。

2）Agentic AI：AI 代理将在后台自主工作，处理任务、管理流程并与员工协作。

3）Xandu：为 NowPlatform 添加了超过 350 个开箱即用的新一代 AI 功能的最新版本。

图45：NOW 平台 HR 系统界面



资料来源：ServiceNow 公司官网，国信证券经济研究所整理

赛意信息：赛意信息聚焦于工业互联网、智能制造等领域的技术与商业模式应用，提供面向 23 个行业和 11 条业务线的产品和解决方案，为企业提供高端软件咨询、实施、集成服务。

AI 产品：善谋 GPT 平台。善谋 GPT（赛意 AI 中台）深度融合赛意信息在财税、人力、营销、供应链、研发与生产制造等领域的知识和最佳实践，致力于构建企业智能化创新引擎，通过上下文记忆、知识/库表索引、Prompt 工程、Agent 执行、通用工具集等扩充大模型的存储记忆、适配应用、调度执行和领域专业能力，形成体系化的企业服务大模型。为企业提供多模型对接、向量管理、私有模型预训练与应用等能力，帮助企业快速落地 AI，实现流程智能化管理、交互、引导与流转。

核心能力包括：模型集成、模型训练、AI 应用开发、Agent 构建

图46：赛意信息善谋 GPT 平台架构

资料来源：赛意信息公司官网，国信证券经济研究所整理

AI+数据与安全

Palantir：专注于大数据分析的定制化软件公司，其技术被广泛应用于国防、安全和金融领域，美国是其主要市场。

AI 逻辑：AI 背景下数据分析、处理、训练需求提升，Palantir AIP 的技术优势能**帮助企业完成数字化转型，**使用大模型的实时数据决策**改善运营并获取竞争优势，**同时 AIP 降低大模型使用门槛，为**企业所有员工赋能，**灵活部署的特点也**扩大了用户覆盖面，提高了商业化用户的渗透率。**

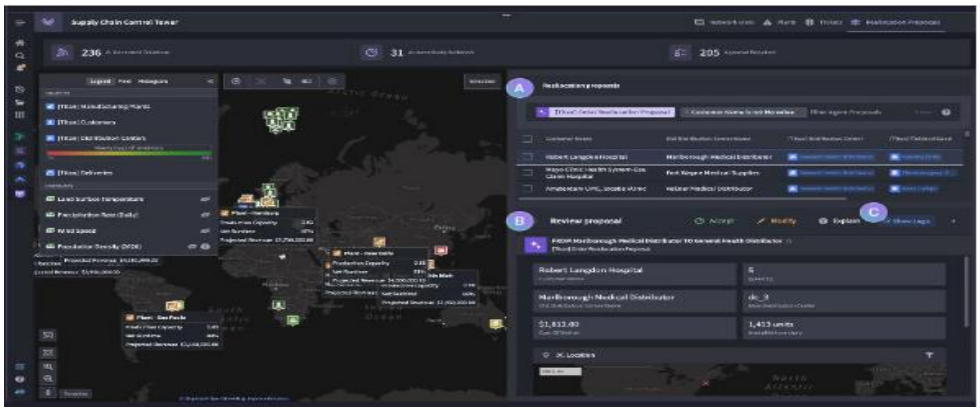
AIP 技术特点：2023 年 4 月，Palantir 正式推出其开发的 AI 产品 AIP，其核心在于 AI 应用而非开发大模型 LLM，面向的企业客户群体包括制造业、医疗、政府、能源等各行各业的企业，主要为传统企业提供深度数据分析的能力。AIP 的技术特点包括：

① 数据的整合、管理及安全：AIP 拥有着强大的数据管理能力，能够控制大模型学习接口和运行范围（细化数据颗粒度等），采取多层次的安全防护措施（数据加密、访问控制等），对构建安全有效的 AI 至关重要，因为 AI 模型的表现往往取决于训练数据的质量和多样性。

② 实时分析决策系统与用户友好界面：AIP 能够即时响应数据的变化，对于金融、网络安全等行业十分重要，能够给客户即时反馈，从而做出更精准的判断。同时 AIP 提供的用户友好界面使缺乏编程经验的人也能操作，降低产品使用门槛，为企业所有员工赋能。

③ 灵活性与可扩展性：AIP 支持定制化开发，可以根据项目的规模（无论是小规模的试点还是大规模的企业机构运作）进行部署，有着灵活性和可扩展性。

图47：Palantir AIP 界面



资料来源：Palantir 公司官网，国信证券经济研究所整理

渊亭科技：Utenet 渊亭科技专注认知智能全栈技术研发与产品化落地，聚焦金融、政务、国防、工业互联网四大行业，为客户提供认知中台、AI 中台、数据中台三大中台产品及 AI+行业解决方案，打通“数据-AI-认知”的闭环服务。

金融：渊亭科技为银行、证券、保险、互联网金融等各类型金融机构，提供针对反洗钱、反欺诈、智能营销、异常交易监测、智能投研等业务场景的智能化解决方案，提升金融风控能力与运营效率。

公安政务：凭借渊亭科技人工智能、知识图谱、数据挖掘、认知计算、AI 中台等全栈技术能力，为各级政府、公共事业单位及城市管理者提供优质的智能应用，提高办公、监管、服务、决策的智能化水平

国防：渊亭科技致力于国防科技建设，凭借多年军工用户服务经验和人工智能全栈技术能力，构建与国防应用需求相适应，满足任务需求的智能化信息系统与解决方案，助力实践强军护国、军地共建以及军民融合的时代目标

图48：渊亭科技核心产品

资料来源：渊亭科技公司官网，国信证券经济研究所整理

AI+广告电商

Shopify：一个基于 SaaS 模式的电商服务平台，商业模式包含订阅解决方案和商家解决方案，其中商家解决方案业务收入占比 70%+。

AI 产品功能：1）**Shopify magic**:包含文本生成、图片生成、应用评论摘要等一系列 AI 功能，所有的 AI 功能都集成在 Magic 中；2）Sidekick:AI 助理，为卖家用户们提供日常的任务处理和商业分析服务。

收费方式：任何套餐都可以使用 Shopify Magic，Sidekick 则需要申请体验资格

AI 实施效果：1）24Q1 超过 50%的商家支持互动是由 AI 协助的，同时 AI 提供了 8 种语言的 24/7 实时支持，这都节省了人员的工作量；2）23 年末和 24 年 1 月，AI 显著提升了营销效率，保持 ROI 标准的情况下主要营销渠道中的商家广告量从第四季度到第一季度增长了近 130%；3）AI 工具大幅提升了销售团队的工作效率，与不使用 AI 的团队相比，赢得了多 31%的销售机会以及 35%的单个销售机会价值；4）AI 回复建议推出后，商家平均约有 50%的回复都使用了建议回复，可以快速提高转化率。

图49：Shopify Magic 主要功能展示

资料来源：Shopify 公司官网，国信证券经济研究所整理

国内相关领域公司：值得买。公司已深入布局消费内容、营销服务与消费数据三大业务板块：凭借消费内容为全网用户提供高效、精准、专业的消费决策支持；通过营销服务帮助电商、品牌商获取用户，扩大品牌影响力；同时沉淀来自消费内容和营销服务业务的数据资源，形成涵盖人、货、场、媒等多种维度的底层数据体系，在支持自身业务发展的同时，对外输出各类数据产品和服务，赋能合作伙伴乃至整个消费行业。

2023 年，公司将 AIGC 确定为重点战略项目，基于 AI 对现有业务商业生态的重塑，通过集团各业务板块的高效协同，实现整体 GMV 超 280 亿。

值得买 AI 应用包括：

小值：值得买科技基于值得买消费大模型基础所研发的 Agent 产品 。可通过对话深度理解用户需求，为存在不同决策难点的消费者提供个性化的建议，提升消费决策的质量和效率。

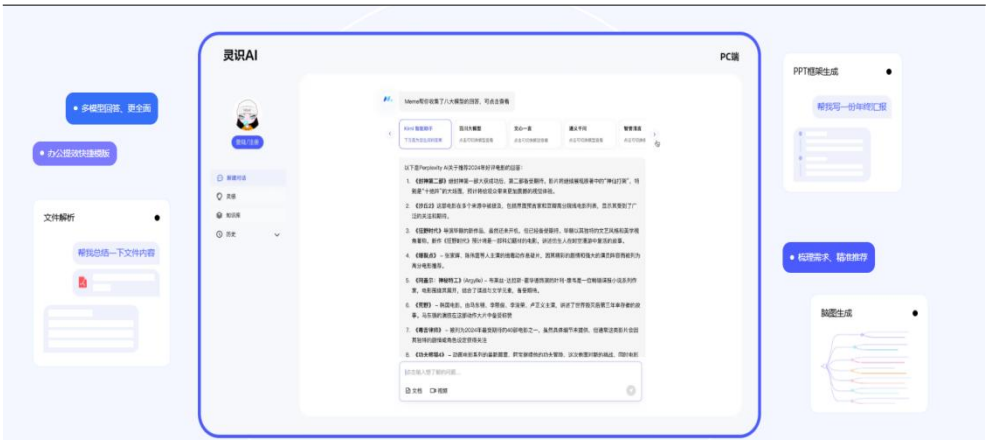
灵识：利用 AI 能力帮助用户快速理解和整理内容的生产力工具。

AIGC 产品：创作者工具、灵眸、神灯素材助手

灵台：凭借海量的商品库和消费库内容，支持通过多种形式，对外提供海量商品数据、用户评价、购买指南等优质预料，及口碑总结、商品对比、商品推荐、全

网比价等服务。

图50：值得买灵识 AI



资料来源：值得买公司官网，国信证券经济研究所整理

投资建议:建议关注国产算力

随着全球大模型持续迭代，模型能力持续提升，AI 应用的蓬勃发展，拉动推理侧算力需求提升，持续看好国产算力发展；此外，字节跳动大模型 API 调用量快速提升，潜在算力需求巨大，相关产业链有望受益，建议关注海光信息、浪潮信息等。

风险提示

大模型研发进展不及预期。

云厂商资本开支投入不及预期。

国产算力迭代及供应不及预期。

免责声明

分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即报告发布日后的 6 到 12 个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A 股市场以沪深 300 指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数（HSI.HI）作为基准；美国市场以标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数 10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数 10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数 10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数 10%以上

重要声明

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所

深圳
深圳市福田区福华一路 125 号国信金融大厦 36 层
邮编：518046　总机：0755-82130833

上海
上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层
邮编：200135

北京
北京西城区金融大街兴盛街 6 号国信证券 9 层
邮编：100032