

双极进化与算力重构：2026 AI 行业深度展望（海外篇）

---全球 AI 系列报告

报告摘要：

宏观格局：中美双极异构，资源约束下的殊途同归。美国依托资本与算力的无限供给，走“规模制胜”的技术理想主义道路。科技巨头（Microsoft/Meta/Google）资本开支不设上限，通过垂直整合（自研芯片+能源自主+闭源模型）构建不可撼动的硬科技壁垒，旨在摘取 AGI 圣杯。中国受制于芯片封锁与资源约束，走“效率突围”的工程激进主义道路。以字节/阿里为代表，通过架构创新（MoE/MLA）与极致工程化（DeepSeek），将推理成本压缩至美国的 1/10，率先在 C 端超级应用（如豆包）与 B 端降本增效中验证了商业闭环。

技术范式：Scaling Law 进化，推理算力迎来“J 型”爆发。Scaling Law 持续进化，传统的“堆参数”边际收益递减，算力价值从“预训练”向“推理端”转移。OpenAI o1/o3 的出现验证了“System 2 慢思考”逻辑，即通过增加推理时间来换取智能涌现。预计 2026 年推理算力需求将首次超越训练算力。随着编程、科研等复杂任务占比突破 50%，推理不再是廉价的“播放”，而是昂贵的“二次思考”，这将为推理芯片与边缘算力打开百倍级增量空间。

基建重构：能源成为第一瓶颈，北美供应链寻求“去英伟达化”。北美数据中心电力需求预计 2035 年增长 30 倍，电网扩容滞后导致严重的供需错配。科技巨头被迫转向“表后取电”，重启核电（三里岛）、押注 SMR（小型核堆）及自建燃气电站成为新常态。虽然英伟达在训练端仍维持统治，但推理需求的爆发为 ASIC（Google TPU, AWS Trainium）打开了缺口。出于成本控制与供应链安全，下游巨头将加速“去英伟达化”进程，也许在 2026 年一些和这些云厂商合作的代工厂将成为新的 Alpha 赢家。

应用落地：从 Chat 到 Agent，高价值场景的范式转移。智能体（Agent）崛起：AI 应用正从“生成内容”转向“解决问题”。能够自主规划、调用工具、闭环反馈的 AI Agent，将破解物流、法律、工业等行业“不可能三角”，实现从辅助人到替代人的价值跃迁。未来随着更多如 AI 智能眼镜、伴侣机器人等新硬件打破屏幕限制，创造全天候感知的原生需求，真正的“iPhone 时刻”可能就在不远的将来。

投资建议：

做多“算力基建的刚性”：传统的算力产业链依旧是核心，只要 Scaling Law 有效，巨头资本开支投入即为确定性 Beta，相关 AI 硬件和电力产业链值得关注。

对冲“硬件垄断的松动”：新的供应链往往意味着巨大的弹性，关注 ASIC 渗透率提升及先进封装产能外溢带来的 Alpha。

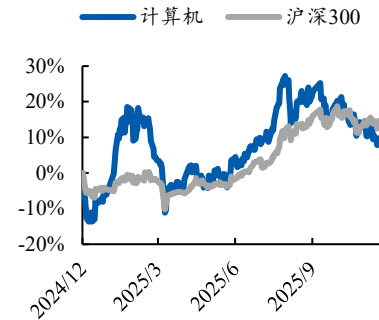
埋伏“应用场景的突破”：应用再一次到了爆发前夜，看好具备行业数据壁垒、能通过 Agent 重构工作流的垂直行业龙头和配套软件巨头。

风险提示：技术迭代不及预期、配套设施建设不及预期、行业竞争加剧。

优于大势

上次评级:优于大势

历史收益率曲线



涨跌幅 (%)	1M	3M	12M
绝对收益	-1%	-6%	12%
相对收益	-4%	-9%	-5%

行业数据

成分股数量 (只)	336
总市值 (亿)	45834
流通市值 (亿)	25851
市盈率 (倍)	83.49
市净率 (倍)	4.27
成分股总营收 (亿)	12390
成分股总净利润 (亿)	154
成分股资产负债率 (%)	44.08

相关报告

《量子计算政策频出，Google 量子算法推进产业进程》

--20251217

《AI 需求驱动存储革命，超级周期已然开启》

--20251213

《2 大利好同日发布，AI 应用蓄势待发》

--20251203

《未来已来，AI 应用进入发展新纪元》

--20251125

证券分析师：吴源恒

执业证书编号：S0550522100004

wuyh1@nesc.cn

目录

1.	中美 AI，步入“战略分流”	4
1.1.	AI 发展产业进一步进入深水区	4
1.2.	美国探寻上限，中国夯实底座	4
1.3.	展望 2026，美国继续深耕硬科技闭环；中国有望引爆应用层红利	7
2.	美国向左，深耕硬科技闭环：从“Scaling Law”到“算力主权”	8
2.1.	模型侧：从“概率生成”到“深度推理”	8
2.2.	应用侧：从“对话工具”到“自主执行的智能体”	13
2.3.	算力侧：从“规模暴力”到“效率、协同与场景”	18
2.3.1.	技术竞赛层面的极致内卷倒逼算力需求持续升级	18
2.3.2.	推理算力的爆发式增长打开了真实的需求空间	19
2.3.3.	战略共生进一步迫使算力投入的前置化与规模化	21
3.	投资建议：做多算力，对冲 ASIC，埋伏应用	23
3.1.	投资的持续性：从“进攻性军备”转向“生存性基建”	23
3.2.	“去英伟达化”的可能性：推理时代的供应链重构	26
3.3.	应用端爆发的潜力：System 2 带来的价值倍增	30
4.	风险提示	33

图表目录

图 1：北美互联网大厂资本开支（十亿美元）	5
图 2：中国互联网大厂资本开支（百万美元）	5
图 3：SuperCLUE 智能指数	6
图 4：SuperCLUE 大模型生成耗时指数（秒）-综合耗时	6
图 5：SuperCLUE 大模型综合价格指数（元/百万 tokens）	7
图 6：缩放定律（参数增加后精度损失连续减少）V.S. 涌现能力（1010-1011 参数后新能力的涌现）	10
图 7：传统 LLM 模型与 o1 模型的推理比较	10
图 8：谷歌模型训练模式	11
图 9：RLHF 与 RLVR 训练模式的区别	12
图 10：从快思考到慢思考	12
图 11：AI 数据飞轮	13
图 12：人类与大语言模型的交互 vs 自主 AI 智能体与大语言模型的交互	14
图 13：多模态播客生成的端到端 Agentic AI 工作流架构	15
图 14：多大型语言模型（LLM）联盟与推理 LLM 的协同响应	16
图 15：AI 应用在各大行业中的渗透率	17
图 16：AI 技术发展带来的新 Scaling Law 范式	18
图 17：推理模型目前占有所有使用量的一半	20
图 18：整个 AI 推理的 Token 使用量还在快速增长	20
图 19：编程已成为绝对的算力消耗主力	20
图 20 北美硬件商业投资闭环	22
图 21：美国人工智能数据中心的电力需求预计将激增	24
图 22：电力公司与数据中心高管看多人工智能对 2035 年电力需求的影响	24
图 23：北美各区已近建成数据中心及预期将建设数据中心的市場情况	25
图 24：北美硬件商业投资闭环	27
图 25：全球数据中心供应链	28
图 26：主流 GPU 厂商 AI 芯片研发之路	29
图 27：海外云计算厂商 AI 芯片研发之路	29

图 28: 火遍苏超的 AI 视频.....	30
图 29: 具备高 ROI 潜力和重构 workflow 能力的新赛道与垂直行业.....	31
图 30: 科技进化的三个阶段.....	32
表 1: 全球大模型迭代时间轴 (2023-2025)	9
表 2: 北美电力相关投资逻辑	26
表 3: 传统英伟达链相关公司不完全整理	28
表 4: AI 端侧相关公司不完全整理.....	32

1. 中美 AI，步入“战略分流”

1.1. AI 发展产业进一步进入深水区

历史车轮滚滚向前，2026 年全球 AI 产业将告别“盲目共识”，步入“战略分流”的深水区。站在 2025 年末的节点上，我们观察到全球 AI 产业正在物理与逻辑的双重层面，分裂为两套截然不同的进化物种。这一轮周期的核心叙事，已不再是全球同步的参数军备竞赛，而是中美双极在约束条件下的极致异构。

- 在大洋彼岸，美国科技巨头正在进行一场“向上的生长”，Google、Meta 与 OpenAI 通过自研芯片与系统级互联，试图打破英伟达的单极霸权，构建垂直封闭的算力帝国；
- 而在中国，科技大厂正在进行一场“向下的扎根”，字节跳动与阿里巴巴凭借对流量与场景的敏锐嗅觉，在算力受限的夹缝中，率先验证了 AI 在 C 端超级应用与 B 端降本增效中的商业闭环。

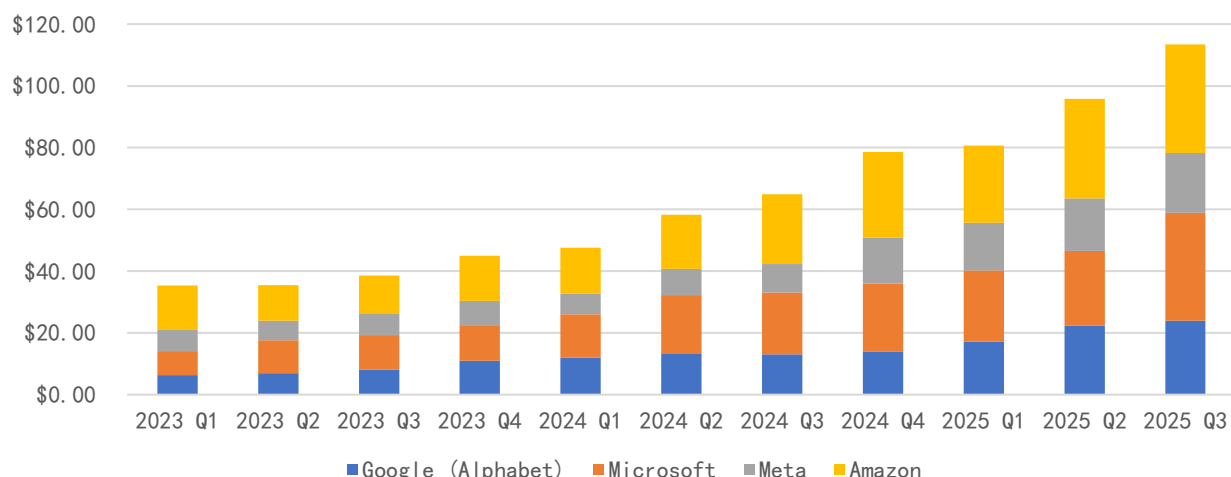
中美大模型的发展差异，本质上是资源约束与战略初心共同作用的结果。

- 美国凭借宽松的算力供给环境，走出了一条“规模制胜”的技术理想主义道路。以谷歌、OpenAI 为代表的企业将 AGI 视为需要极致理论突破的终极目标，核心模型研发周期常达一年半以上，执着于在多模态融合、科学推理等基础能力上追求极致。OpenAI 计划投入 1.4 万亿美元建设算力基础设施，依托超大规模算力集群实现模型后训练深度与推理效率的同步提升，坚信当算力规模达到阈值后智能会自然涌现。这种“保守”的技术深耕模式，使其在复杂逻辑推理、幻觉控制等核心维度长期保持领先，GPT-5 系列在科学计算、代码生成等领域的性能优势，正是建立在这种长期主义的技术积累之上。
- 中国团队则面临算力供给受限的现实挑战，选择了“效率突围”的工程激进主义路径。通过算法优化与架构创新，在有限资源下实现性能突破：DeepSeekV3.2 采用 MoE 架构，以 671B 参数规模完成训练仅用不到 280 万个 GPU 小时，不足 Llama3405B 训练时长的十分之一；华为盘古通过稀疏化技术与注意力机制改进，在不损失性能的前提下降低算力消耗，形成“技术优化+生态协同”的独特优势。这种“螺蛳壳里做道场”的智慧，推动中美顶级大模型的性能差距从 2023 年的 20% 缩至 2025 年的 0.3%，在智能体、多模态生成等细分领域实现局部反超。政策层面的精准赋能更强化了这一路径，中国“十四五”规划明确支持开源体系建设，杭州推出“AI 开源政策包”，单家企业每年最高可获 1000 万元算力补贴，推动魔搭社区汇聚全球头部开源模型，形成从技术研发到场景落地的良性循环。

1.2. 美国探寻上限，中国夯实底座

美国投资增速极快，2025 年 Q3 是美国 AI 算力“军备竞赛”的高潮期。Microsoft 和 Meta 处于“不计成本”的抢货阶段，而 Google 和 Amazon 则紧随其后，四大厂单季砸进去的 1134 亿美元，正在转化为海量的 GPU/TPU 订单和数据中心钢筋水泥。其中 Alphabet(Google)Q3 单季支出约 240 亿美元，主要用于支持 Gemini 和 Cloud 业务。CFO 明确表示 2026 年的资本开支将会有“显著增长”。Microsoft Q3 单季近 350 亿美元，而且 CFO Amy Hood 表示支出实际上是“受限于供应”的，如果芯片和数据中心交付能更快，花钱会更多。Meta 是几家中 Capex 增速最恐怖的，Q3 单季度近 200 亿美元（同比翻倍，+110%）。扎克伯格的逻辑是：宁可现在“浪费”算力，也不能在 AI 时代掉队。Amazon(AWS)将全年 Capex 指引暴力拉升至 1200 亿美元以上，相关投资在 2025 年全面爆发。这反映了 AWS 客户对 AI 模型训练和推理的巨大积压需求。

图 1：北美互联网大厂资本开支（十亿美元）

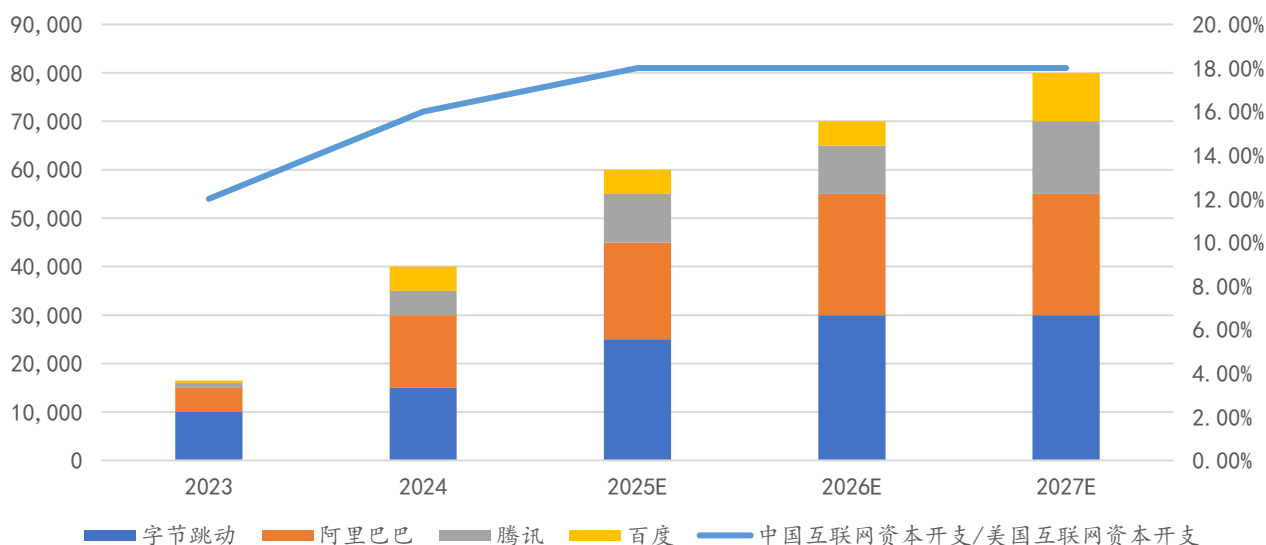


数据来源：各公司年报、东北证券

中国增速也很快，但和美国在总量上存在较大差距。中国互联网四巨头（字节跳动、阿里巴巴、腾讯、百度）的资本开支正在经历一轮显著的“补偿式增长”周期。2023 年是中国互联网大厂投入的低谷期，合计 Capex 仅不到 200 亿美元。但进入 2024 年，该数字迅速反弹至约 400 亿美元（同比翻倍），并预计在 2025 年进一步攀升至 600 亿美元量级，显示出国内大厂在 AI 基础设施建设上的紧迫感正在转化为实际的资金投入。

字节跳动是本轮 Capex 扩张的绝对主力，从 2024 年起占据了行业总投入的近半壁江山，我们预期公司 2027 年仍将维持约 250-300 亿美元的年投入规模。这主要归因于其 TikTok 全球算法推荐的巨大算力消耗，以及“豆包”等 AI 应用在 C 端流量变现上的激进策略。相比之下，腾讯与百度的投入曲线较为平滑，反映出更强的防御性与聚焦性。虽然中国在增速上试图追赶，但由于基数差异巨大，中美在 AI 算力底座上的绝对资金缺口仍在扩大。到 2027 年，美国大厂预计将维持 5000 亿美元左右的年投入规模，而中国仅为 1000 亿美元左右，两者体量相差 5 倍。

图 2：中国互联网大厂资本开支（百万美元）



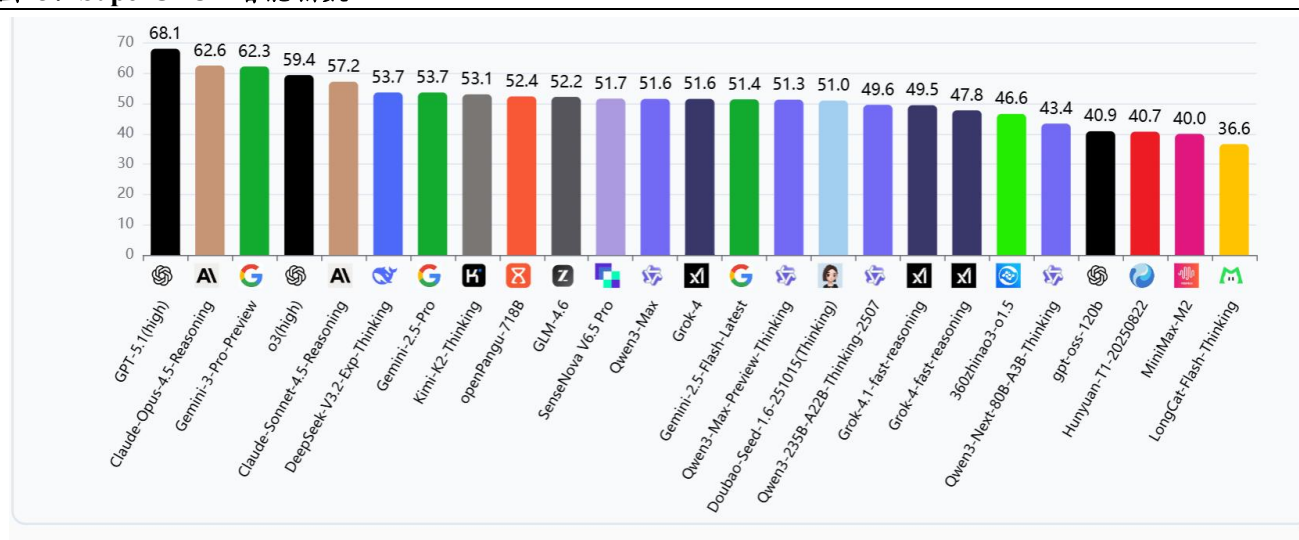
数据来源：各公司官网、东北证券

注：由于字节并未公开其具体资本开支，部分为预测数据为高盛预测数据

投入的差距也导致了中美大模型在跑分和生态上的区别，在大模型的发展中，美国在探寻智能的上限，而中国在夯实生态的底座。根据 SuperCLUE 在 2024 年底至 2025 年的测评数据（截至 2025 年 11 月最新动态），在代表 AI 最高水平的综合总榜中，美国依然掌握着‘核按钮’。OpenAI 的 o1 系列凭借 System 2 慢思考能力，在逻辑与科研高地（T0 层级）保持着对中国头部模型（DeepSeek-V3、Qwen-Max）的代际领先，这一‘智力鸿沟’短期内难以通过算力堆叠填平。

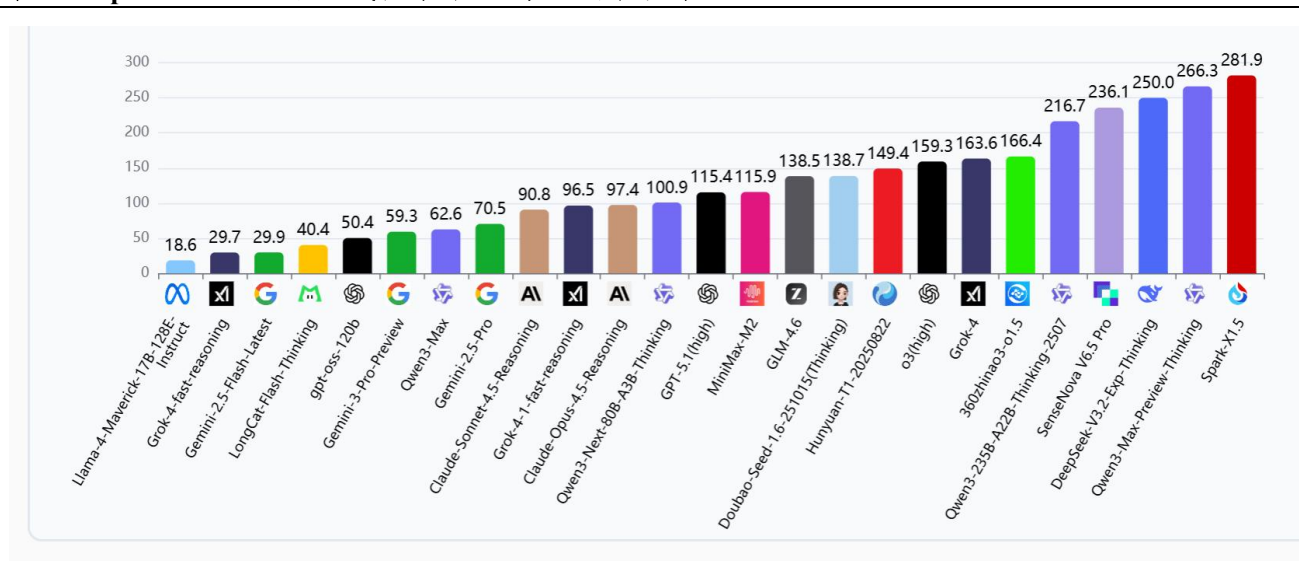
然而，在决定产业落地广度的开源以及性价比榜单中，中国已实现‘战略反攻’。阿里 Qwen 与 DeepSeek 系列模型凭借架构创新，在代码、数学及多语言任务上多次力压 Meta 的 Llama 系列，成为全球开源生态事实上的新标杆。这标志着中国 AI 产业已成功从‘模型复刻’阶段跨越至‘生态输出’阶段，为全球开发者提供了比美国方案更优的性价比选择。”

图 3：SuperCLUE 智能指数



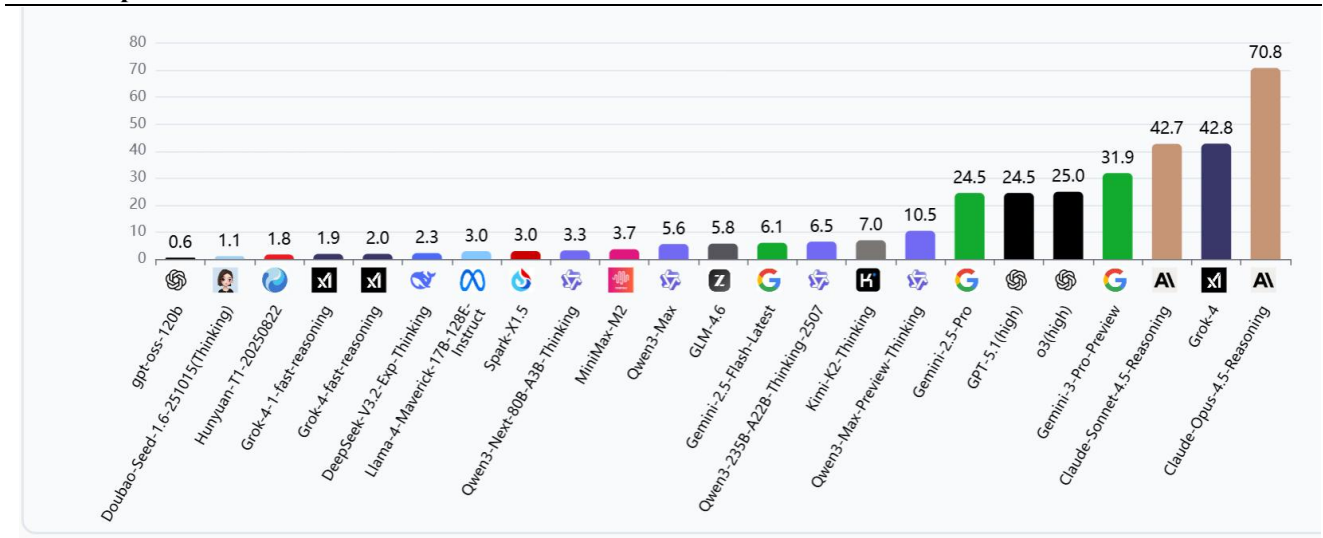
数据来源：Superclueai、东北证券

图 4：SuperCLUE 大模型生成耗时指数（秒）-综合耗时



数据来源：Superclueai、东北证券

图 5: SuperCLUE 大模型综合价格指数 (元/百万 tokens)



数据来源: Superclueai、东北证券

就像我们之前说的，中美大模型在 2025 年呈现出的“能力 vs. 生态”错位格局，并非偶然的技术选择，而是双方在算力供给、市场结构与数据禀赋三重约束条件下，经过两年残酷博弈后的必然演化结果。

- **算力路径：暴力美学 vs. 极致工程** 美国依托 H100/B200 顶级集群，信奉 Scaling Law 的资源堆叠模式探索 AGI 上限；中国受制于芯片封锁，倒逼出 MoE、MLA 等架构层面的极致工程创新。这种“进化压力”反而使中国模型构建了 5-10 倍的推理成本优势，为商业化铺平道路。
- **商业策略：封闭收租 vs. 开源突围** 美国（OpenAI）凭借先发优势走“iOS 式”封闭系统，赚取高额 SaaS 利润；中国巨头（阿里/字节）选择“安卓式”开源策略，通过将底层模型“大宗商品化”来消解美国的技术溢价，迫使竞争维度下沉至中国擅长的应用生态与云基建。
- **数据场景：抽象逻辑 vs. 具象落地** 美国掌握代码与学术期刊等“抽象数据”，确立了逻辑推理与编程的统治地位；中国坐拥复杂的移动互联与产业链“具象数据”，使其在文本创作、情感陪伴及垂类 Agent 等落地场景中展现出更强的适应性。

1.3. 展望 2026，美国继续深耕硬科技闭环；中国有望引爆应用层红利

因此在 2026 年，美国科技巨头的核心焦虑不再是“买不到英伟达”，而是“过于依赖英伟达”。我们看到了一场前所未有的“算力垂直整合运动”。最典型的信号来自 Google。过去，TPU 只是 Google 内部的秘密武器；但现在，我们发现 Google 正在转型为“算力军火商”，开始向 Meta、Apple 甚至 xAI 出售 TPU 硬件与服务。这不仅仅是商业模式的改变，更是为了分摊高昂的研发成本，试图构建一个平行于英伟达之外的算力生态。与此同时，OpenAI 也不再满足于软件层的统治。他们与博通（Broadcom）合作的自研 ASIC 芯片已经流片，预计 2026 年正式量产。

美国的 AI 竞争正在从软件层的“百模大战”，下沉到硬件层的“硅片战争”。在这个战场上，决定胜负的“胜负手”已经从芯片设计转移到了先进封装。2026 年台积电 CoWoS 产能虽然翻倍，但仅英伟达一家就要吃掉近 60%。因此，2026 年美股科技的 Alpha，将属于那些掌握“非英伟达体系”算力供给能力的公司，比如博通，以及那些掌握产能瓶颈的代工巨头。

在大洋彼岸疯狂堆砌硬件的同时，中国科技大厂正在极其艰难的算力约束下，走出了一条“应用为王”的独特路径。如果说美国是在做“加法”（堆算力），那么中国巨

头正在做“除法”（降成本）和“乘法”（裂变流量）。看“乘法”：字节跳动的“豆包”是一个现象级样本。在没有任何先发优势的情况下，豆包凭借“推荐算法+精准买量”的暴力美学，仅用半年时间就将日活推到了 1.7 亿级别，是同类竞品的 20 倍。这证明了中国互联网最擅长的“超级 App”打法，在 AI 时代依然奏效。更值得期待的是，2026 年春节，字节计划发布对标 Sora 的视频生成模型。这将是传媒与内容行业的“iPhone 时刻”，它将直接打穿短剧与网文的生产成本。看“除法”：阿里巴巴的通义千问。面对算力封锁，阿里选择了极致的工程化突围。他们计划明年将 80% 的算力迁移至自研 PPU，目标是支撑 5 万亿 token 的调用量。这意味着，中国 AI 产业正在率先完成从“拼参数”到“拼 ROI”的进化。

综上所述，2025 年的中美 AI 格局，是“美国科研红利”与“中国工程红利”的双轨并行。美国正在向上捅破天花板，试图摘取 AGI 的圣杯；而中国正在向下扎根，试图通过极致的性价比与开源生态，将 AI 变成滋养万业的黑土地。而在此背景下，2026 年双方的宏观图景已经非常清晰：美国向左，深耕硬科技闭环；中国向右，引爆应用层红利。

2. 美国向左，深耕硬科技闭环：从“Scaling Law”到“算力主权”

2026 年美国巨头正在通过“垂直一体化”来重构 AI 产业。面对 Scaling Law 的边际成本上升，他们不再单纯依赖英伟达，而是通过自研芯片、能源自主和 System 2 模型，构建一个不可撼动的硬科技壁垒。

2.1. 模型侧：从“概率生成”到“深度推理”

过去三年，全球大模型行业完成了从“技术探索”到“产业落地”的关键跨越，其迭代轨迹并非零散的功能升级，而是围绕算力分配、生态格局与交互形态的三大核心方向系统性演进，深刻重塑了行业竞争逻辑与商业价值判断。

- **2023 年作为“探索时代”，**行业核心发展在于对 Scaling Law 的探索，普遍相信“参数规模与算力投入决定智能上限”，OpenAI 凭借 GPT-4 系列的性能优势独领风骚，引发全行业的恐慌性追赶，“做大模型、建通用底座”成为唯一共识，暴力堆叠 GPU 与扩大参数规模成为技术竞争的核心范式。
- **2024 年进入“实用时代”，**用户需求从“能对话”升级为“能看懂图、能处理长文本、反应更快”，多模态融合与交互效率成为核心诉求，行业发展呈现明显架构分叉；DeepSeek 通过创新的 MLA (Multi-Head Latent Attention) 架构与 DeepSeek MoE 技术，打破了“算力即唯一壁垒”的认知，在保持高性能的同时将推理显存占用压缩至极致，证明算法效率的核心价值，为 2025 年的行业价格战埋下关键伏笔。
- **2025 年则迈入“推理与效率时代”，**System2 推理、价格崩塌与端侧落地成为关键词，模型发展形成鲜明双轨制：向上突破的 OpenAI o1 与 GPT-5，通过强化学习 (RL) 与自我博弈技术内化思维链，消耗更多推理侧算力实现“慢思考”能力，在数学竞赛中正确率达 83.3%，编程能力较前代提升 8 倍，持续试探 AGI 上限；向下普及的 DeepSeek 与 Llama 系列，则通过动态序列调整、批处理优化、低精度推理等技术极致压缩成本，其中 DeepSeek 成为行业“性价比度量衡”，迫使所有厂商直面“价格贵 10 倍是否能力强 10%”的拷问，推动各大旗舰模型降低推理成本，让 AI 从昂贵奢侈品转变为水电煤般的基础设施。

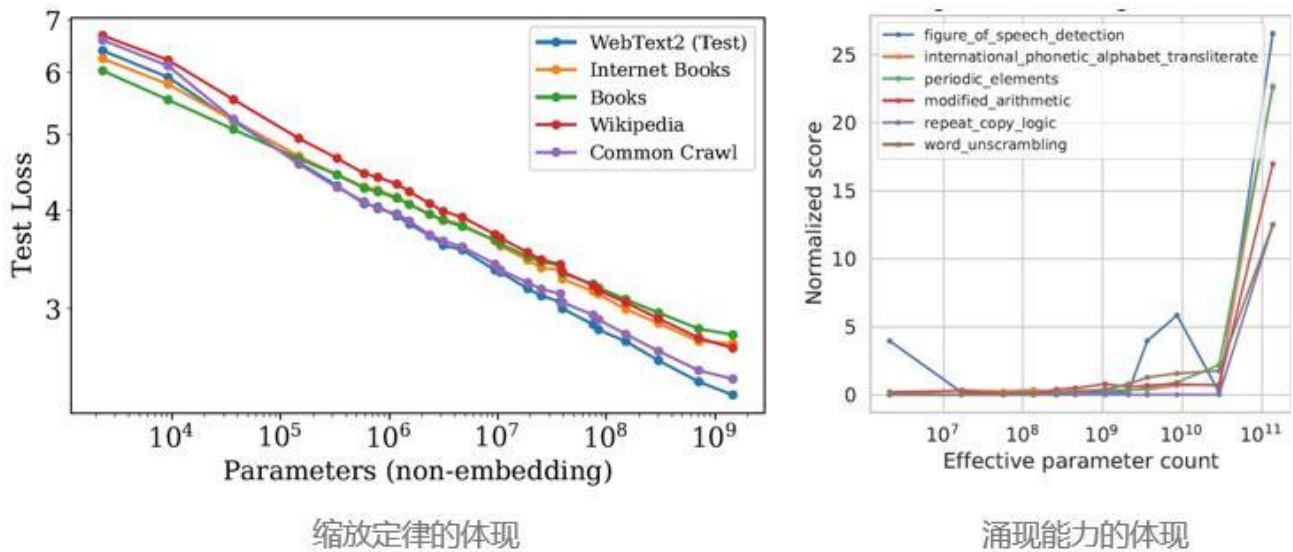
表 1：全球大模型迭代时间轴（2023-2025）

时间轴	模型名称	开发厂商	参数量	核心突破与行业指示
2023. 03	GPT-4	OpenAI	~1.8T (MoE)	定义了“智能涌现”的标准，证明了 ScalingLaw（缩放定律）在万亿参数级依然有效。
2023. 07	Llama2	Meta	7B-70B (Dense)	打破了 OpenAI 的垄断，让企业和个人能低成本私有化部署，开启了“开源 vs 闭源”之争。
2024. 03	Claude3opus	Anthropic	未公开	首次有模型在代码编写和复杂逻辑上体感超越 GPT-4，强调了“人设”和“安全性”的重要性。
2024. 05	GPT-4o	OpenAI	未公开	将 AI 从“文本工具”推向“实时助理”，极低的延迟让语音交互真正可用。
2024. 09	Qwen2.5	Alibaba	72B (官方)	开源性价比碾压。全球开发者发现：用 Qwen72B 写代码比用 Llama405B 更好、更快、更便宜
2025. 01	DeepSeek-V3/R1	DeepSeek	671B MoE (极致性价比)	R1 模型以极低的推理成本复现了 OpenAI o1 的思考能力。API 价格仅为对手的 1/20，直接引爆了 2025 年的“算力通缩”。
2025. 02	Gemini2.0 (Flash/Pro)	Google	未公开	不再只是聊天，而是能实时看屏幕、听声音并操作电脑。Flash 版本速度极快，成为语音助手的标准后台。
2025. 03	Qwen-Max (v3)	Alibaba	405B+ (自进化架构)	依旧是最强开源模型。进一步强化 Coder 属性，不仅能写代码，还能看懂 UI 设计图直接生成前端。
2025. 08	GPT-5	OpenAI	System2 (慢思考) 普及化	将 o1 的推理能力完美融入 GPT 通用模型。它不再急于回答，而是会“想”几秒钟再输出，不仅能写代码，还能自己 De-bug。
2025. 11	Gemini3.0	Google	超大 MoE+ 物理世界理解	针对 GPT-5 发布的强力反击。Gemini3 在复杂的科学推理和长视频理解上再次刷新 SOTA，且深度集成了 Deep Research（深度研究）功能。
2025. 12	GPT-5.2/DeepSeek-V4	OpenAI / DS	端侧推理 vs 云端巨兽	DeepSeekV4 致力于让 GPT-5 级别的能力跑在单卡甚至高端手机上；而 OpenAI 继续堆砌云端算力，探索超级智能。

数据来源：东北证券不完全统计

在 GPT-4 为代表的早期阶段，大模型行业的竞争本质是“算力堆砌的规模竞赛”。彼时，行业共识聚焦于“参数越大、训练算力越足，模型能力越强，也会出现较强的涌现能力”，头部企业纷纷加码训练端投入——OpenAI 训练 GPT-4 据称消耗了超 1 万枚 A100 GPU，Meta Llama2 的训练算力成本高达数亿美元，参数规模从千亿级向万亿级冲刺成为最直观的技术标杆。这一阶段，芯片投资的核心诉求是“大显存”，以支撑超大规模模型的训练过程，英伟达 A100/H100 凭借高显存容量与算力密度，成为行业绝对的“训练神器”。

图 6：缩放定律（参数增加后精度损失连续减少）V.S. 涌现能力（1010-1011 参数后新能力的涌现）



数据来源：OpenAI、东北证券

而随着 OpenAI o1 模型的发布，行业迎来算力价值的根本性转移，推理时算力成为新的竞争核心。o1 通过引入“多步推理规划”机制，证明模型在回答复杂问题时，“思考时间”（即推理过程中消耗的算力）与智力表现呈强正相关，其在数学推理、逻辑分析等任务中的准确率较前代提升 35%，核心原因并非参数规模扩大，而是推理阶段的算力投入增加了 2-3 倍，模型能够通过多轮迭代验证、工具调用模拟等方式优化结论。这一突破彻底打破了“唯参数论”，推动行业重新审视算力分配逻辑，即是未来的 AI 系统计算开销将更多地集中在推理服务上，而非单纯的预训练计算，o1 模型证明，在推理过程中给予更多时间生成“思考标记”与提高各项任务的准确率直接相关。

图 7：传统 LLM 模型与 o1 模型的推理比较



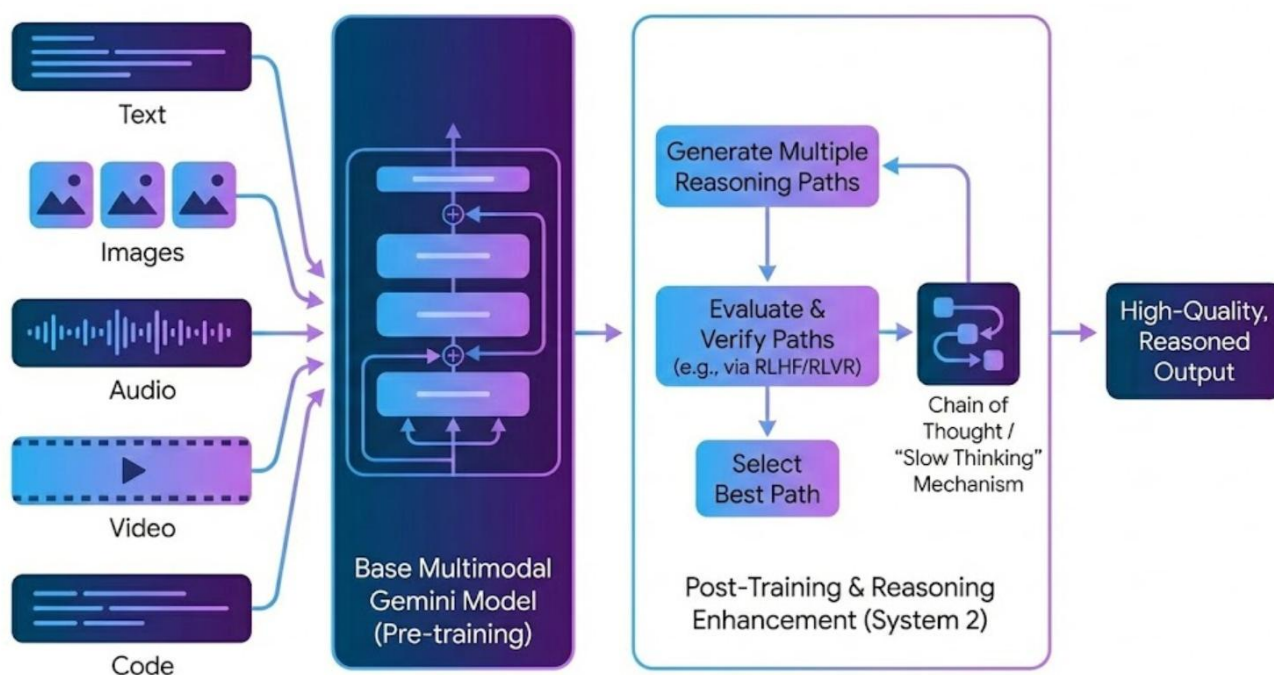
数据来源：OpenAI、东北证券

这一转变标志着大模型行业正从“参数至上”的粗放增长期，迈入“智能密度”与“推理效率”并重的精细化运营阶段。随着模型规模逼近万亿参数大关，这种“暴力美学”的投入产出比开始急剧下降。单纯堆砌算力和参数不再能线性地解决复杂的逻辑推理与泛化难题，反而带来了难以承受的能源消耗与训练成本。行业开始意识到，盲

目追求参数规模的物理极限，正逐渐触碰到高质量训练数据的“天花板”——人类互联网上现存的高质量文本已被“吃干榨净”，继续通过增加低质数据来换取参数增长，无异于饮鸩止渴。为了突破 Scaling Law 的瓶颈，训练的重心从单纯的“预训练 (Pre-training)”向更高权重的“后训练” (Post-training) 与“推理训练”(Inference)转移。

以谷歌为例，为了应对 Scaling Law 的边际效应递减，其技术路径明显从单纯的“堆算力、喂数据”转向了“强化学习驱动的推理能力”与“合成数据自我进化”。它不是分别训练不同模态的模型再进行拼接，而是从一开始就将文本、图像、音频、视频和代码等多种数据类型输入到一个统一的“多模态 Transformer 架构”中进行联合训练。这种设计使得 Gemini 能够天生理解和推理不同模态之间的关系，从而在问答、图像描述、代码生成和复杂推理任务上表现出色。

图 8：谷歌模型训练模式

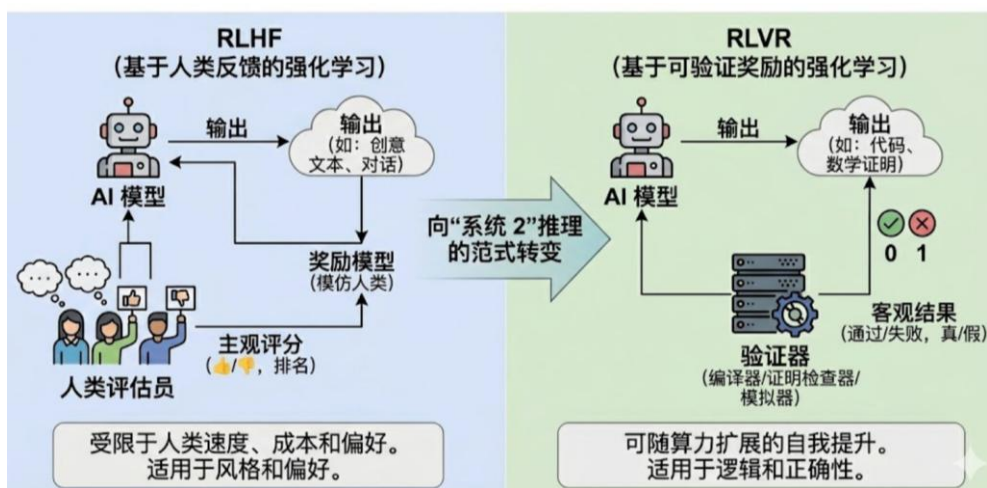


数据来源：Google DeepMind、东北证券

谷歌解决这一瓶颈的核心策略可以概括为以下三个维度：

- **从 RLHF 进阶到 RLVR：用“可验证”的奖励信号训练逻辑：**传统的 RLHF（基于人类反馈的强化学习）依赖人类打分，但在高难度的数学或代码问题上，人类很难给出精准反馈。谷歌通过 AlphaProof 和 AlphaGeometry 等项目，探索出了一条 RLVR（Reinforcement Learning with Verifiable Rewards，基于可验证奖励的强化学习）的新路径。简单说 RLHF 是教模型“像人一样说话”（讨好人类），而 RLVR 是教模型“像科学家一样思考”（追求真理）。
- **核心逻辑：**在数学证明或代码生成等领域，答案的对错是客观的（可以被编译器或形式化证明器验证）。谷歌利用 DeepMind 的 AlphaZero 算法思路，让模型在这些“真理明确”的封闭环境中进行大规模的自我对弈和搜索。
- **解决瓶颈：**这种方式不再依赖昂贵且稀缺的人类专家数据，而是让模型通过“试错-反馈-强化”的循环，自主探索出人类未曾教过的解题路径。这直接突破了高质量人类数据耗尽的限制。

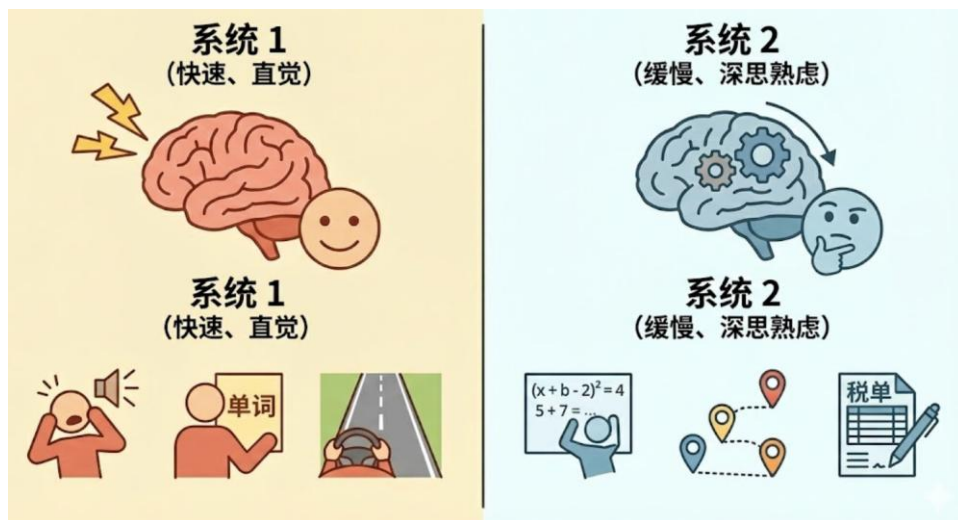
图 9：RLHF 与 RLVR 训练模式的区别



数据来源：东北证券制作

- **引入“慢思考” (System 2) 机制：Test-time Compute (推理时计算) 的工程化。** 这类似于 OpenAI o1 的思维链，谷歌在 Gemini 1.5 及后续版本中集成了 System 2 这种慢思考模式。
 - **训练重心的转移：**模型不再被训练成“脱口而出”的反射机器，而是被训练去生成中间推理步骤。训练的目标函数从“预测下一个 token”变成了“预测正确的推理路径”。
 - **计算换智能：**通过在推理阶段消耗更多算力（让模型在输出最终答案前进行多轮内部搜索、自我批判和验证），模型能够处理极高复杂度的任务。这种模式将算力需求从“训练端”部分转移到了“推理端”，绕开了单纯增加参数带来的边际收益递减。

图 10：从快思考到慢思考

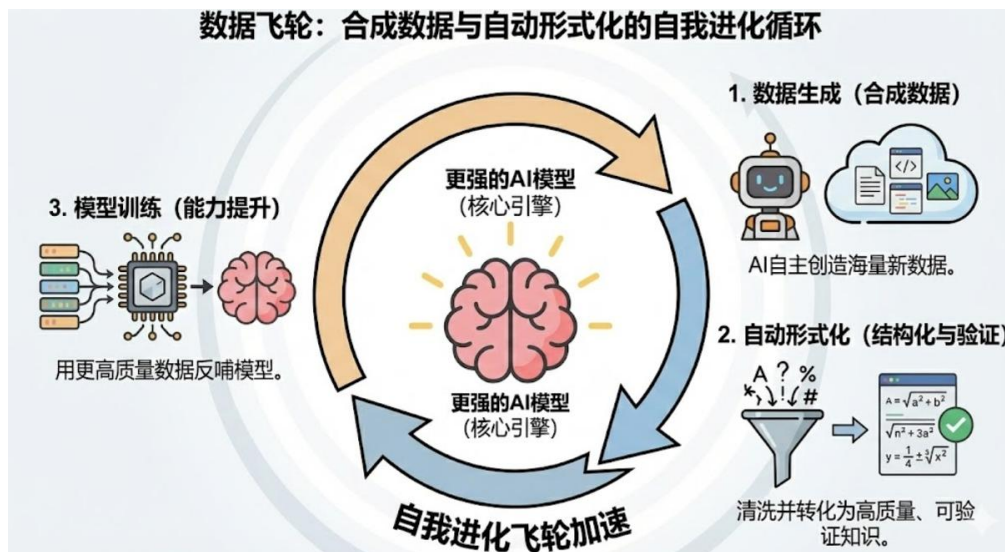


数据来源：东北证券制作

- **数据飞轮：合成数据与自动形式化：**为了解决高质量数据匮乏的问题，谷歌利用现有的大模型来“生产”训练数据，形成自我进化的飞轮。
 - **自动形式化：**在 AlphaProof 中，谷歌使用一个微调过的 Gemini 模型，将海量的自然语言数学题“翻译”成形式化语言。
 - **价值闭环：**这些被形式化的数据成为了完美的训练材料。模型自己生成问题、自己尝试解决、自己验证对错，并将成功的案例反向输入到训练集中。这种

“合成数据+自我训练”的闭环，使得模型能够在没有外部新数据的情况下，依然通过挖掘现有知识的深度来提升能力。

图 11：AI 数据飞轮



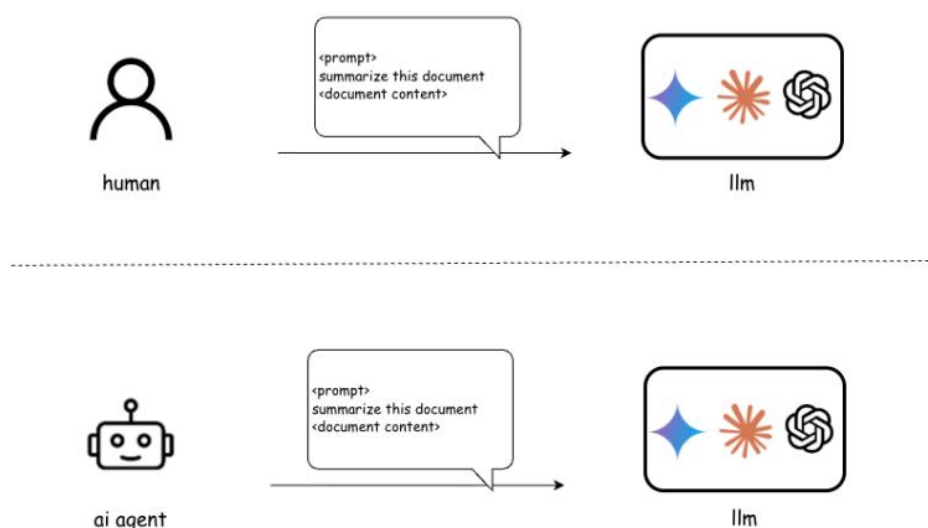
数据来源：东北证券制作

总结而言，大模型的算力角逐已发生结构性跃迁。谷歌对 Scaling Law 瓶颈的突围，揭示了行业竞争的核心已从拼“记忆广度”（看谁见过更多数据）的资源战，升级为拼“思维深度”（看谁能利用强化学习与搜索算法进行逻辑推演）的技术战。未来的行业壁垒，不再单纯取决于谁能囤积更多的算力卡来训练庞大的静态模型，而在于谁能掌握合成数据的“炼金术”以打破数据枯竭的僵局，以及谁能通过算法优化赋予模型类似人类的“慢思考”能力。随着算力需求从训练端的“一次性大规模投入”向推理端的“持续性高强度思考”分流，这不仅是对 Scaling Law 的修正，更标志着人工智能正在重新定义“智能”的本质——从复述知识转向解决问题。

2.2. 应用侧：从“对话工具”到“自主执行的智能体”

大模型的交互形态演进，本质是“模型与人类、模型与世界”连接方式的升级，过去三年完成了从“信息交互”到“任务执行”的关键跨越。早期阶段，大模型以“聊天机器人”为核心形态，交互局限于“提问-回答”的双向信息传递——用户输入文本需求，模型输出文字反馈，核心价值是“信息检索与整理”，代表模型如 GPT-3.5、早期 Claude1，仅能作为辅助工具解决简单的信息类问题。

图 12：人类与大语言模型的交互 vs 自主 AI 智能体与大语言模型的交互



数据来源：A Practical Guide for Designing, Developing, and Deploying Production-Grade Agentic AI Workflows、东北证券

随着技术迭代，交互形态逐步突破“纯文本对话”的边界，同时 AI 也在替代人类完成与 LLM 的自主交互，实现任务执行的自动化。Gemini1.5 支持上传整本书籍、超长视频文件，实现“内容解析-关键信息提取-个性化总结”的全流程处理；Claude3.5 推出的 Artifacts 功能，允许用户在对话中直接生成代码、表格、简易 App，并实时预览效果，实现了“需求提出-成果生成”的闭环交互。

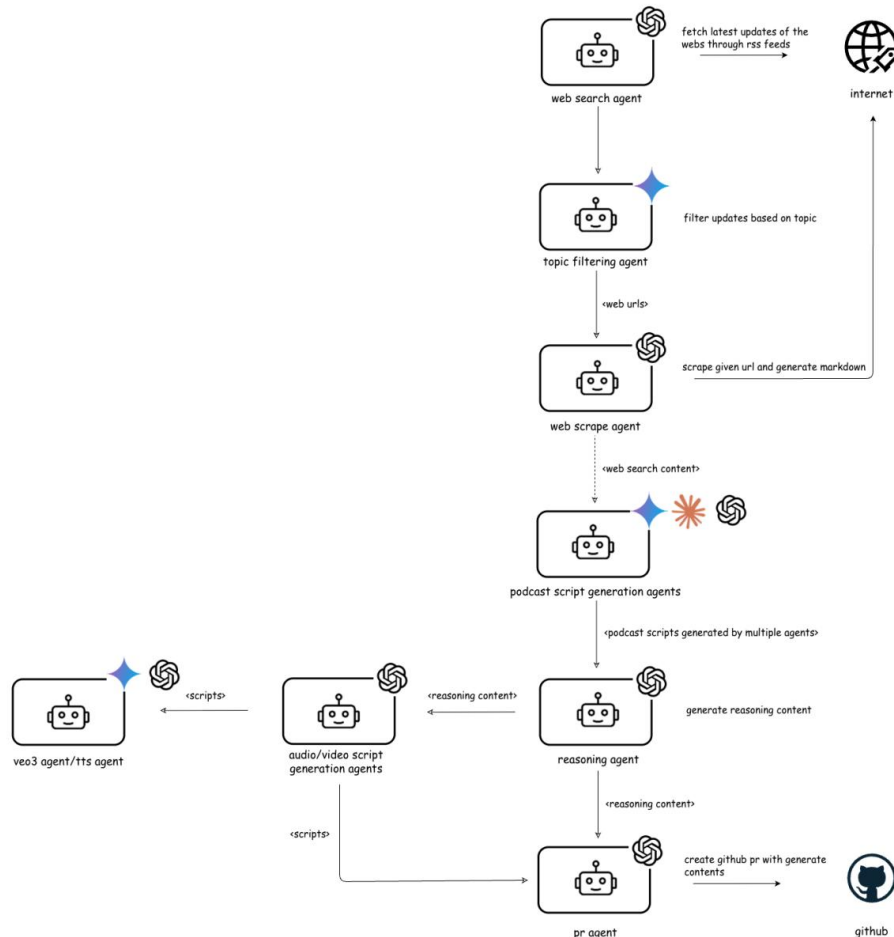
当前，行业正加速向“智能体（Agent）”方向演进，模型的核心定位从“回答问题”转变为“执行任务”。这一阶段的模型具备三大核心能力：自主规划（基于复杂需求拆解为可执行的步骤）、工具调用（自主调用搜索引擎、办公软件、代码编译器等外部工具）、闭环反馈（根据执行结果优化调整方案）。这一演进趋势不仅将重塑人机协作模式——未来人类将更多聚焦于“目标设定”与“结果审核”，而模型承担具体执行工作，还将推动工具生态的爆发式增长，围绕 Agent 的工具平台、插件市场、闭环反馈系统等配套生态将快速成熟，最终构建起“模型+工具+场景”的全链路智能生态。

比如下图就是一个“从实时新闻到多模态播客”的实际场景，完整展示了生产级 Agentic AI 工作流的组件分工与协同逻辑，核心流程与角色如下：

1. **输入触发**：用户提供“主题 + 新闻源 URL”，启动工作流；
2. **数据获取层**：Web Search Agent（调用 RSS 和 MCP 搜索端点收集最新新闻）→ Topic Filtering Agent（筛选与主题匹配的新闻 URL）→ Web Scrape Agent（爬取网页内容并转换为结构化 Markdown，为后续生成提供统一数据格式）；
3. **内容生成层**：Podcast Script Generation Agents（多模型联盟，如 OpenAI、Gemini、Anthropic 分别生成差异化脚本）→ Reasoning Agent（整合多模型脚本，解决冲突、剔除推测性内容，生成统一合规的最终脚本）；
4. **多模态输出层**：Audio/Video Script Generation Agents（将最终脚本转换为 TTS 语音提示、Veo-3 视频提示）→ Veo Agent（生成 JSON 格式的 Veo-3 视频指令）+ TTS 生成器（输出 MP3 音频）；
5. **发布层**：PR Agent（整合脚本、音视频文件，通过 MCP 集成的 GitHub 服务器自动创建分支、提交内容并发起 Pull Request）。

这个场景就是 AI 展现出的通过将复杂任务拆解为单一职责的智能体，搭配工具、LLM 与外部系统（如 GitHub），实现从数据输入到多模态输出的全流程自主化，同时体现“可观测、可审计、可复用”的生产级特性。

图 13：多模态播客生成的端到端 Agentic AI workflows 架构



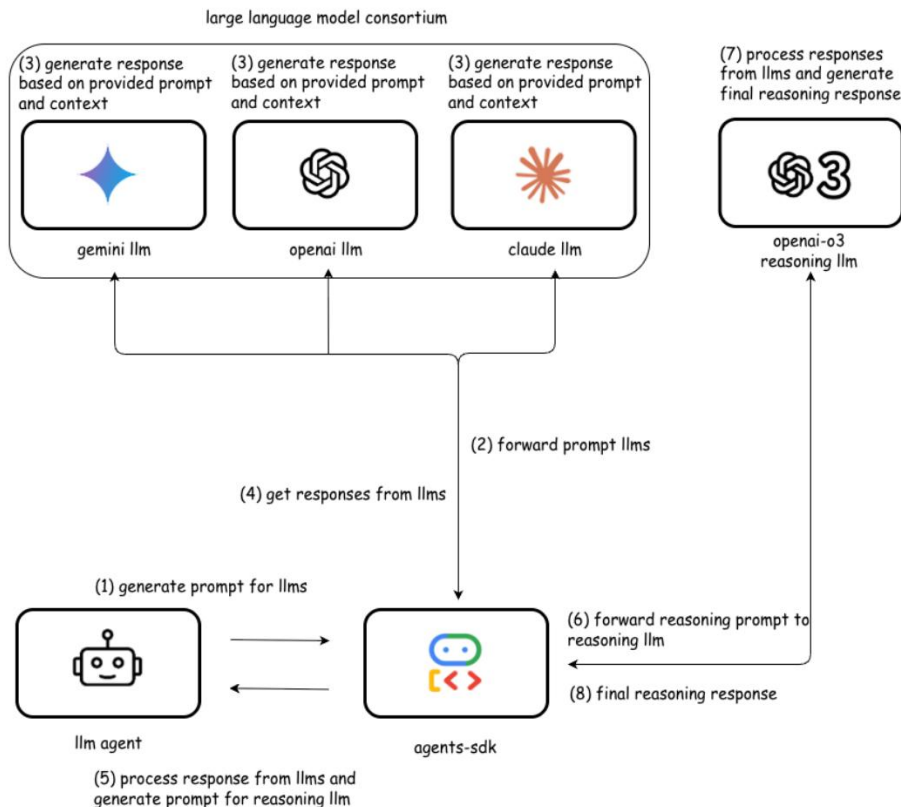
数据来源：A Practical Guide for Designing, Developing, and Deploying Production-Grade Agentic AI Workflows、东北证券

在生产级 AI 的演进过程中，“动态切换不同大模型以适配多元需求、优化输出质量”已成为核心趋势。本质上正是 AI Agent 实现多模型切换的核心架构：智能体通过统一的 SDK 生成标准化提示词，将其同步分发至 Gemini、OpenAI、Claude 等异构大模型组成的“模型联盟”，各模型基于自身能力特性独立生成响应，再由专用推理 LLM 对多元输出进行冲突解决、事实对齐与逻辑整合，最终形成兼具准确性、全面性与合规性的结果。这种切换并非简单的模型堆砌，而是 Agent 基于任务需求的精准调度，背后依托于统一接口抽象、模型元数据管理与动态路由策略的技术支撑，确保不同模型的能力无缝衔接。

AI Agent 之所以倾向于多模型切换，核心是为了破解单一模型的能力局限，虽然有的时候单一模型效果会更好。不同大模型在训练数据、架构设计上的差异，使其形成了鲜明的能力分工：部分模型擅长深度逻辑推理与专业领域内容生成，部分模型在多模态处理或长文本解析上效率更优，还有些轻量模型能以更低成本实现快速响应。在实际 workflows 中，Agent 会根据任务环节的核心需求自动完成模型切换。这种动态切换模式还实现了“质量与成本的双重优化”：对于简单问答、基础摘要等轻量需求，Agent 切换至低成本的开源模型或轻量闭源模型，控制推理成本；对于金融

风控、医疗诊断等高精度需求，则自动切换至性能顶尖的旗舰模型，保障结果可靠性。同时，通过模型切换形成的“多模型交叉验证”机制，让 Agent 的输出在事实一致性、逻辑严谨性上显著提升，正如文档中提到的，这种模式使多模态播客脚本的事实准确率提升 40%，幻觉得分优化至行业领先水平。未来，随着自适应选择网络、动态加权融合等技术的成熟，AIAgent 的模型切换将更趋智能，不仅能根据任务类型匹配模型，还能结合实时负载、响应延迟等动态调整策略，真正实现“按需调度、最优输出”，推动生产级 AI 向更高效、可靠、灵活的方向演进。

图 14：多大型语言模型（LLM）联盟与推理 LLM 的协同响应



数据来源：A Practical Guide for Designing, Developing, and Deploying Production-Grade Agentic AI Workflows、东北证券

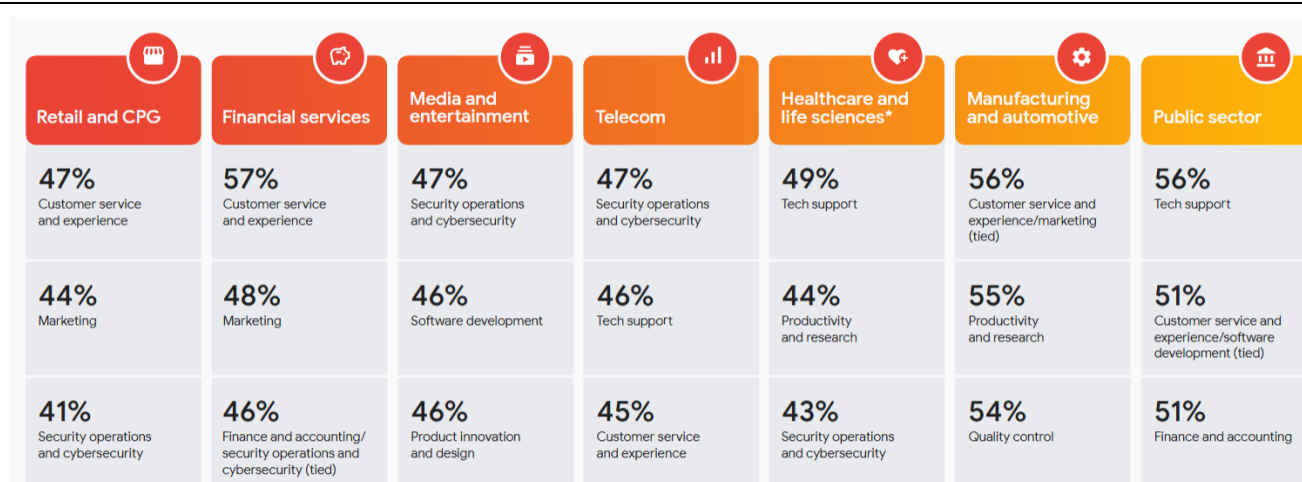
模型推理效率的提升也带来了下游应用范围以及效率的增长。根据《Google 谷歌：2025 年人工智能投资回报率分析报告》，AI 智能体的部署已在零售与消费品、金融服务、媒体与娱乐、电信、医疗健康、制造业以及公共部门这七大主要行业中全面铺开。各行业的 Top 3 应用场景反映了企业正根据自身的核心业务痛点和价值驱动力进行差异化投资。

- **零售与快消行业将客户服务、营销、安全运营列为核心场景，核心在于其直面海量消费者的业务属性：**高频重复的咨询需求（如订单查询、售后问题）恰好适配 AI 智能体 7x24 小时自主响应、快速调取知识库的能力，而营销场景中对用户数据的精准分析与个性化内容生成，以及安全运营中对交易风险、数据泄露的实时监测，均能通过 AI 智能体整合多源数据、自动化执行校验流程的优势高效落地；
- **金融服务行业同样将客户服务置于首位，叠加营销与财务 / 安全运营的双重需求：**AI 智能体既能替代人工处理大量标准化的开户咨询、账单查询等任务，又能在营销中精准匹配客户风险偏好与金融产品，在财务核算、安全

风控中严格遵循合规规则,通过多工具调用(如对接交易系统、风控数据库)实现无偏差执行。

- **媒体娱乐行业聚焦安全运营、软件开发与产品创新,核心是其业务流程中存在大量需要跨环节协同的非结构化任务(如内容版权保护、视频制作工具开发、创意方案生成):** AI 智能体可通过编排多专业角色(数据爬取、风险识别、代码生成),将分散的工作流整合为自动化时间表;电信行业的安全运营、技术支持与客户服务,则贴合其“网络运维复杂、用户基数庞大”的特点, AI 智能体能够实时监测网络安全漏洞、快速响应设备故障咨询、自主处理套餐变更等高频需求,大幅降低人工运维压力。
- **医疗健康行业以技术支持、生产力研究、安全运营为核心,关键在于医疗领域对专业知识的高依赖与数据的海量性:** AI 智能体可快速检索医学知识库辅助医护人员解决技术难题,整合临床数据支撑科研分析,同时严格遵循医疗数据隐私规范完成安全校验;
- **制造与汽车行业将客户服务 / 营销、生产力研究、质量控制列为重点,源于其生产流程标准化程度高、供应链协同复杂的特性:** AI 智能体能够自动化处理售后咨询、优化生产排程、通过机器视觉整合实现产品质量的实时检测;
- **公共部门则聚焦技术支持、客户服务 / 软件开发、财务核算,核心是公共服务对“高效响应、透明合规”的要求:** AI 智能体可自主处理政务咨询、简化办事流程、规范财务报销审核,在提升服务效率的同时确保流程可追溯

图 15: AI 应用在各大行业中的渗透率



数据来源: Google Cloud、东北证券

基于这一逻辑推演,未来大量的 AI 需求将爆发性地向以下三大“深水区”行业渗透。这些行业的共性在于,其核心业务流程均由大量“重复度高(可被训练)、规则明确(可被验证)、且需多环节协同(需长程推理)”的任务链构成。而 AI Agents 的出现,标志着生产力工具从“单点辅助”向“全流程托管”的跨越——它们不仅能理解孤立的任务,更能通过自主规划工作流、跨系统整合工具与数据、以及标准化的执行与纠错,精准破解了这些行业长期存在的“人力认知成本高昂、跨部门响应效率低下、复杂流程协同困难”的“不可能三角”。

这意味着, AI 智能体将不再仅仅是依附于旧有流程的“提效外挂”,而是逐步演化为重塑行业底层逻辑的“数字基础设施”。未来,谁能率先完成从“人驱动流程”到“Agent 驱动业务”的范式跃迁,谁就将彻底打破传统服务业的人力规模瓶颈,掌握定义下一代生产关系与价值分配的主动权。

2.3. 算力侧：从“规模暴力”到“效率、协同与场景”

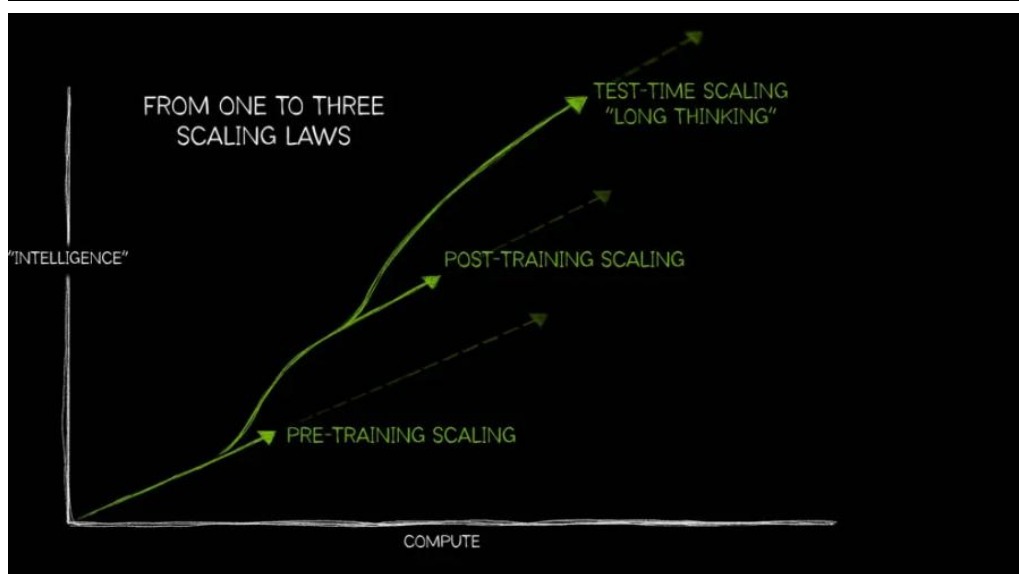
在 AI 需求爆发的潜在需求下，北美算力建设如火如荼。站在 AI 应用大规模爆发的黎明前夜，北美科技巨头们正以一种近乎疯狂的决绝，在此刻押注未来。面对即将到来的智能体时代与海量推理需求，硅谷不再等待市场验证，而是选择率先开启一场史无前例的算力基建狂潮。从沙漠中拔地而起的吉瓦级（GW）数据中心，到争分夺秒重启的核电站，北美算力建设正以“战时速度”飞奔，试图通过超饱和的算力供给，提前锁定未来数字经济的入场券与定价权。

尽管市场当前对于算力建设可能存在的“泡沫化”风险仍怀有隐忧，但我们研判认为，2026 年算力基础设施的建设非但不会放缓，反而将进入加速上行周期。支撑这一趋势的核心逻辑在于，当前 AI 算力投资正受到三大核心引擎的强力驱动：技术竞赛层面的极致内卷倒逼算力需求持续升级，推理算力的爆发式增长打开了真实的需求空间，以及战略共生进一步迫使驱动算力投入的前置化与规模化。这三股力量的共振将有效对冲市场情绪波动，推动算力建设向更高量级跃升。

2.3.1. 技术竞赛层面的极致内卷倒逼算力需求持续升级

市场曾热议 Scaling Law 面临“失效”，但 2025 年的技术突破与实证研究还是证明了目前算力的堆叠还没到极限，就像前文提到的失效的是传统“单纯堆参数”路径，而非定律本身。谷歌 DeepMind 通过 DiLoCo 分布式优化技术验证，即使总通信量减少 99%，Scaling Law 仍能通过优化算法实现超越数据并行训练的性能提升，证明模型性能与计算资源的正相关关系仍是客观规律。前 OpenAI 多模态后训练负责人毕树超进一步指出，Scaling Law 本质反映数据结构的幂律特性，需通过指数级增长的算力突破稀有知识模式的学习瓶颈，这一客观规律构成了厂商算力投入的底层逻辑。由于训练数据和算力增长受限，我们不能再盲目追求模型规模，而应转向提升模型的“能力密度”，即在有限资源下使模型更高效、更智能，这才是法则的核心指引。当然我们也需承认，Scaling Law 更多是一种规律的总结，而非客观定律，因此即便这一规律失效，只要大模型训练尚未触及算力扩张的天花板，技术层面的算力堆叠就仍会持续推进。

图 16：AI 技术发展带来的新 Scaling Law 范式



数据来源：英伟达、东北证券

因此我们认为，在技术战争白热化的明天，2026 年头部厂商在训练端的投入必然是不减反增，呈现出“基础设施超大规模化”与“技术路径精细化”并行的双重特征。这本质上是一场以垄断为终局的算力军备竞赛：

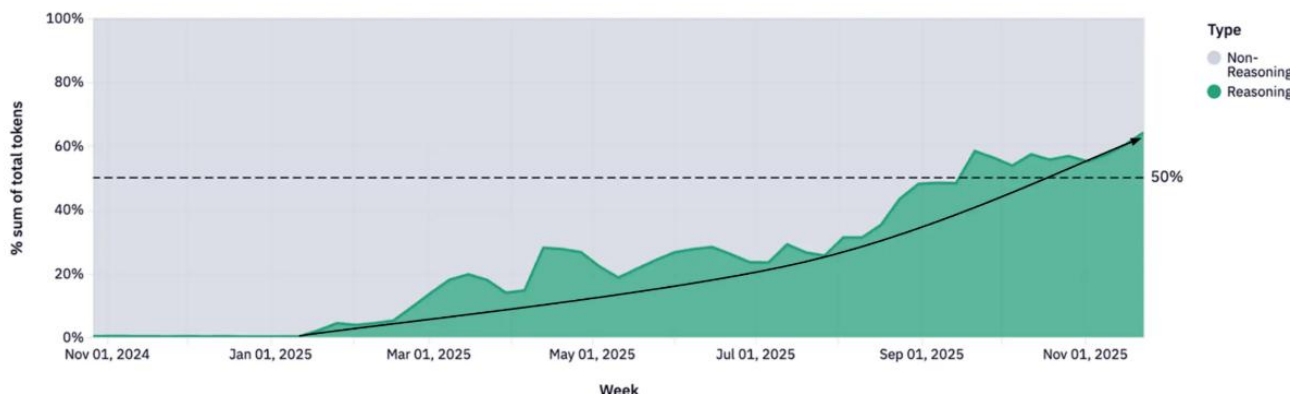
- **基建层面的“绝对规模突破”**：OpenAI 与英伟达达成的 10 吉瓦 (GW) 级数据中心合作，标志着算力建设进入了“国家电网级”时代。这项高达 1000 亿美元的单一项目投资，不仅部署了数百万块 GPU，更构建了一个算力规模超当前主流超算中心 10 倍以上的“超级大脑”，其战略意图直指下一代通用人工智能的训练霸权。
- **商业层面的“战略性成本容忍”**：尽管千亿参数模型的单次训练成本已突破 6000 万美元，但头部厂商正通过“算力领先 → 模型性能碾压 → 用户生态垄断”的飞轮效应，将高昂的 CAPEX 转化为不可逾越的市场护城河。OpenAI 凭借 GPT-4o 系列在 C 端占据 7 亿周活用户的案例证明，超额的算力投入是获取市场定价权的唯一路径。
- **技术层面的“从堆砌到协同”**：算力竞争已从粗放的“参数堆砌”进化为“算力+算法+数据”的系统工程。以 DeepSeek-V3 为例，其通过算法优化实现了媲美顶尖模型的性能，但这并未否定算力的重要性，反而证明了“高效算力堆砌”正在取代“盲目算力堆砌”。定制化的高密度算力集群，依然是支撑算法创新的物理底座。
- **竞争层面的“生存入场券”**：对于 OpenAI、Google、Meta 等巨头而言，拥有最强的基础模型不再仅仅是为了增长，更是为了“生存”。在 Scaling Law 被数学证明失效之前，这是一场无人敢率先退出的“囚徒困境”——停止投入即意味着被生态淘汰。

2.3.2. 推理算力的爆发式增长打开了真实的需求空间

推理需求也正将成为 AI 算力需求的核心拉动力，这一趋势已被行业数据与技术演进逻辑充分验证。就像前文提到的，随着代理式 AI、多模态交互、行业定制化解决方案的规模化渗透，推理需求正从“辅助训练”的配角转变为驱动算力增长的主角。与训练算力“一次性高投入、头部厂商集中布局”的特征不同，推理算力具备“高频复用、场景分散、长尾延伸”的属性：从消费端的 AI 助手实时交互、内容生成，到产业端的工业质检、医疗诊断、自动驾驶决策，每一个 AI 应用的落地都需要持续的推理算力支撑，且单场景的规模化推广（如千万级用户的 AI 办公工具）会直接催生指数级的算力增量。更关键的是，推理扩展定律的显现（复杂任务性能提升与推理算力消耗正相关）、边缘计算与端云协同的普及，进一步拓宽了推理算力的应用边界，使其覆盖从数据中心到终端设备的全场景，成为贯穿 AI 产业链、连接技术与市场的核心枢纽，最终主导未来 AI 需求的增长曲线。

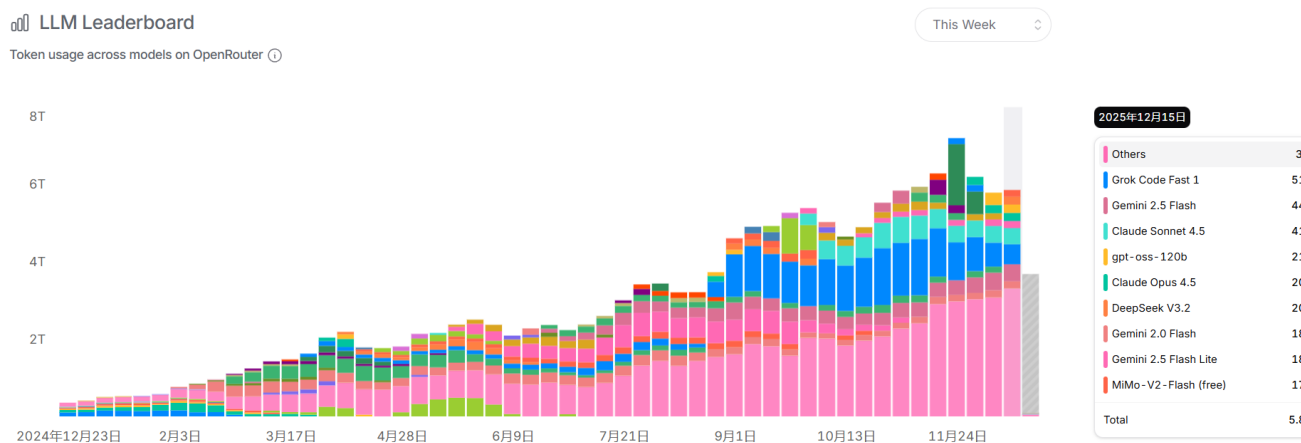
OpenRouter 的最新实盘数据显示，全球 AI 推理需求在 2025 年并未放缓，反而呈现出加速上扬的“J 型曲线”特征。从总量上看，推理算力的消耗量已实现近 10 倍的爆发式增长——周度 Token 消耗量从 2023 年底的不足 1 万亿 (1T)，一路飙升至 2025 年 11 月的 7.5 万亿 (7.5T) 以上，且在该年底仍保持着陡峭的增长斜率，未见任何见顶迹象。这种增长背后的驱动力发生了本质性的结构反转。数据显示，“推理模型 (Reasoning Models)”已正式接管市场。截至 2025 年 11 月，具备慢思考能力的推理模型流量占比首次突破 50%，超越了传统的非推理模型。这意味着，算力的消耗不再仅仅源于用户量的增长，更源于模型为了“思考”更复杂的问题而消耗了倍数级的推理时算力。

图 17：推理模型目前占有所有使用量的一半



数据来源：OpenRouter、东北证券

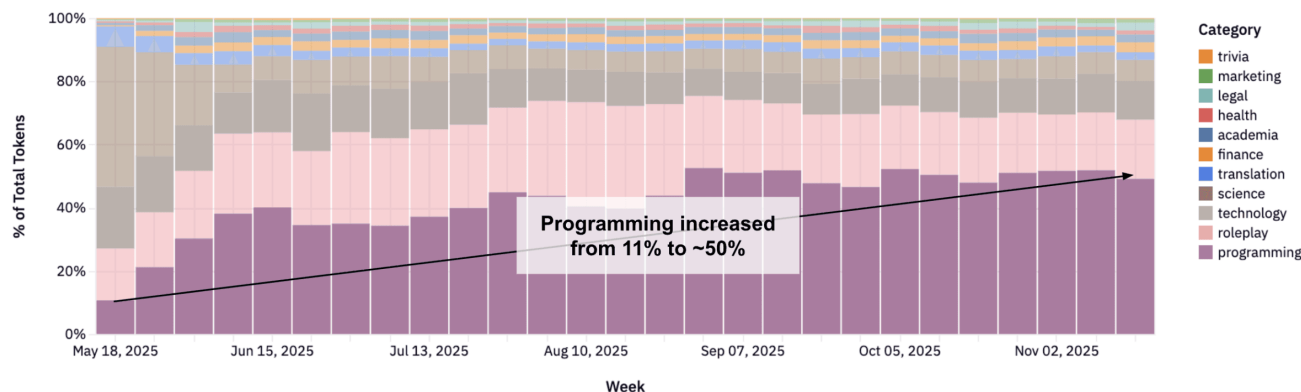
图 18：整个 AI 推理的 Token 使用量还在快速增长



数据来源：OpenRouter、东北证券

在具体的应用场景中，编程（Programming）已成为绝对的算力消耗主力。数据清晰地展示了“编程”类推理需求的统治级表现：其占比从 2025 年 5 月的仅 11% 左右，在短短半年内极速扩张至 50%，彻底主导了推理算力的流向。除编程外，角色扮演(Roleplay)依然维持着第二大场景的地位，而科技(Technology)、科学(Science)与金融(Finance)等专业领域的长尾需求虽然占比相对较小，但也随着总盘子的扩大而在绝对值上持续增长。这表明，AI 推理正在从泛娱乐闲聊，全面转向以代码生成和复杂逻辑解决为核心的生产力工具。

图 19：编程已成为绝对的算力消耗主力



数据来源：OpenRouter、东北证券

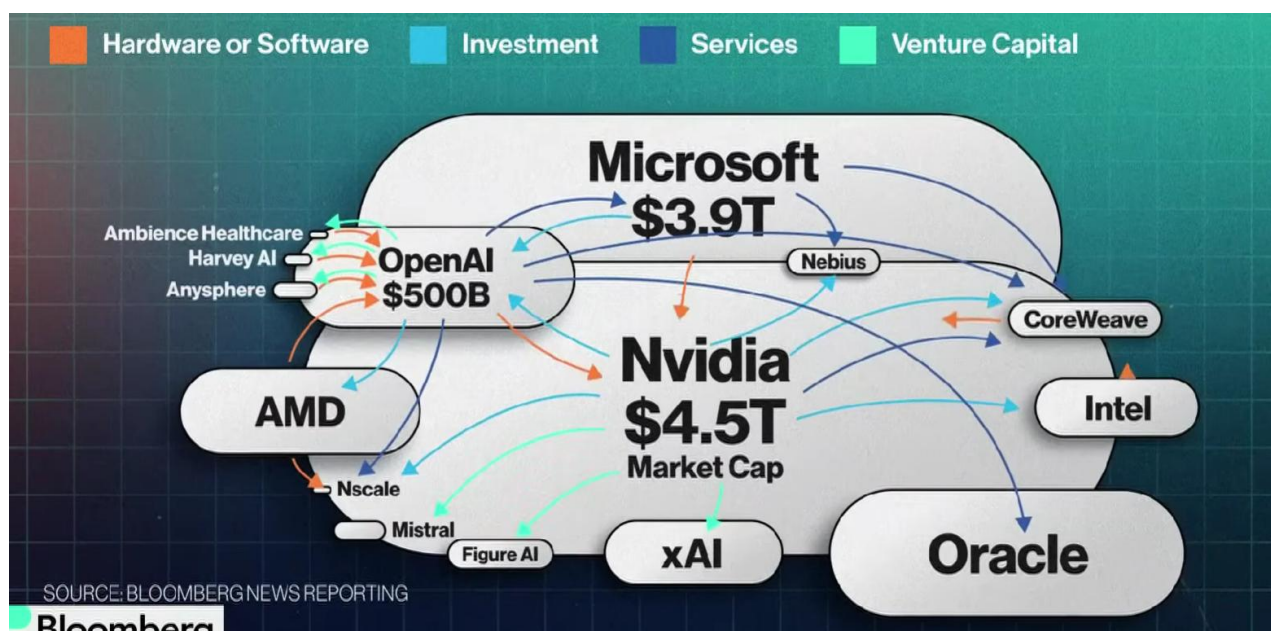
在这些数据的支持下，我们有理由认为 2026 年，AI 产业将迎来历史性的“算力分水岭”：推理算力需求或将首次全面超越训练算力，成为驱动半导体增长的第一引擎。这一结构性翻转并非简单的线性增长，而是由“推理扩展定律”与“应用场景深化”双重定义的计算范式革命：

- **技术范式的“维度升维”**：推理即训练的延伸 OpenAI o1/o3 系列模型的问世，验证了“Test-time Compute（测试时计算）”的新摩尔定律：智能水平与推理时间呈现正相关。通过链式推理（CoT）与思维树搜索，模型在复杂任务上的准确率提升了 40% 以上，但代价是算力消耗的指数级通胀——o3 高强度版本的单任务推理成本高达 20 美元，DeepSeek R1 在复杂数学任务上的计算量更是传统大模型的 150 倍。这意味着，推理不再是廉价的“播放”，而是昂贵的“二次思考”。
- **业务场景的“密度跃迁”**：从 Chat 到 Agent 随着代理式 AI 与多模态技术的落地，推理算力正从“低密度的文本生成”向“高密度的跨模态决策”演进。在医疗领域，整合病史与影像的多模态诊断将效率提升 3 倍，其背后是单次推理算力需求激增 5 倍的代价；在自动驾驶端，为了将跨模态决策延迟压缩至 0.1 秒以内，边缘节点的算力密度被迫提升一个数量级。应用场景的价值越高，对算力密度的渴求越贪婪。
- **商业逻辑的“通胀效应”**：更为关键的是，慢思考机制的引入，彻底改变了推理的单位经济模型。未来的每一次交互，不再是简单的 Token 预测，而是一场复杂的“思维链推演与自我博弈”。随着 ChatGPT、Copilot 等国民级应用的日活（DAU）持续激增，这种由“深度思考”带来的算力乘数效应，将使处理海量并发请求的推理成本迅速超越模型训练成本，为推理芯片市场打开指数级的增量空间。

2.3.3. 战略共生进一步迫使算力投入的前置化与规模化

在算力军备竞赛进入深水区之际，AI 大模型厂商与芯片巨头之间的关系已超越了简单的买卖供需，演变为一种“深度绑定的战略共生体”。为了消除未来技术迭代与商业回报的巨大不确定性，双方正共同构建一套严密的“双向锁定机制”。一方面，以 OpenAI、Google、Microsoft 为代表的模型厂商，通过坚定押注 Scaling Law 与 System 2 慢思考路径，将未来五年的竞争壁垒直接固化为对“万卡集群”的持续性依赖——这实际上是主动锁定了自身的巨额资本开支路径，用今天的投入换取未来的技术定义权；另一方面，以 NVIDIA 为首的芯片霸主，则敏锐地利用了这种“生存焦虑”，通过将产品迭代周期从两年压缩至一年（Tick-Tock 加速）以及构建全栈封闭生态（如 CUDA+NVLink），将下游的算力饥渴转化为长期的订单确定性。这种“算法定义硬件，硬件固化生态”的螺旋式上升，不仅极大抬高了行业的入场门票，更在客观上联手封锁了未来算力格局的演进方向，确立了赢家通吃的游戏规则。

图 20 北美硬件商业投资闭环



数据来源：Bloomberg、东北证券

2026 年，这种深度绑定的商业诉求，正从供给侧强力驱动算力投入的前置化与规模化：

- **长期协议构筑“产能护城河”**：英伟达通过“硬件交付+资本注入”的双轮驱动策略，与 OpenAI 等核心客户签署了 10GW 级别的系统部署协议。这种高达千亿美元级的阶段性投资，实质上是以“产能预付+技术定制”的模式，提前锁定了未来 5 年的 GPU 采购清单，为芯片厂商穿越半导体周期提供了极高的业绩确定性。
- **缩短周期制造“持续性焦虑”**：为了支撑万亿市值并向资本市场证明需求的永续性，芯片厂商主动将产品迭代周期从传统的两年压缩至一年（Tick-Tock 加速：H100 → Blackwell → Rubin）。这种“技术性折旧加速”策略，迫使大模型厂商为了维持“Scaling Law”的领先优势，必须陷入“永不停歇”的硬件采购循环，从而在客观上形成了一个完美的商业闭环。
- **市值与算力的“双螺旋共振”**：芯片厂商的市值逻辑已完全锚定于 AI 算力的刚性需求——2025 年英伟达数据中心业务营收占比突破 70%，且头部模型厂商贡献了全球高端 GPU 65% 的销量。芯片厂商通过在幕后推波助澜“算力军备竞赛”，成功将下游的生存焦虑转化为了自身的市值溢价与市场份额的双重护城河。

纵观北美算力市场的演进路径，我们看到的是一场由“技术规律（Scaling Law）”、“市场需求（推理需求增长）”与“市场意志（深度绑定）”共同编织的宏大叙事。这三股力量还在 2026 年交汇成一股强大的协同效应：技术法则定义了算力需求的指数级增长，推理应用则提供了各式各样的需求场景，而产业链策略则铺设了确定性的供给轨道。尽管竞争的焦点正从粗放的“规模暴力”转向精细化的“效率、协同与场景”，但这并未改变算力作为“数字石油”的核心地位。相反，在三重逻辑的层层加码下，算力基建正成为定义未来十年科技版图的最关键变量，驱动整个行业加速冲向万亿市值的星辰大海。

3. 投资建议：做多算力，对冲 ASIC，埋伏应用

从行业发展的角度，我们建议关注以下三点，投资的持续性，去英伟达化的可能性已经应用端爆发的潜力。从投资的角度，我们认为 2026 年海外 AI 线的思路在于——做多“算力基建的刚性”，对冲“硬件垄断的松动”，并提前埋伏“高价值智能体”的商业化质变。

3.1. 投资的持续性：从“进攻性军备”转向“生存性基建”

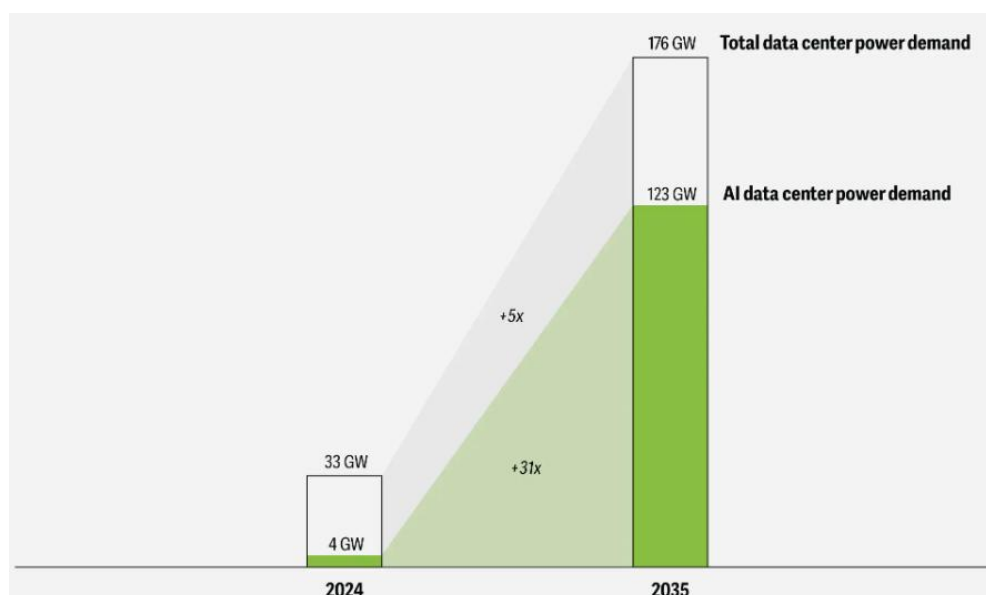
市场普遍担忧巨额资本开支后的回报缺口，但从产业竞争格局来看，算力投入具有极强的刚性与不可逆性。从目前各大北美互联网企业的高管访谈来看，大部分高管持有的观点是：投资过度的后果只是浪费一些钱（随后可以通过折旧消化），但投资不足的后果是公司倒闭（被技术时代抛弃）。

- **囚徒困境下的“资本底座”：**就像我们说的，对于云巨头而言，算力投入已不再是单纯的商业选择，而是关乎生存的入场券。只要 Scaling Law 依然有效，任何削减投入的行为都等同于主动退出“牌桌”。因此，即便短期应用回报滞后，来自科技巨头的 CAPEX 仍将维持高位，这种“为了不被淘汰而付费”的逻辑保障了上游产业链业绩的长期下限。
- **主权级算力的接力：**除了企业竞争，主权 AI 在成为新的增量资金来源。中东、欧洲及亚太国家正在将算力视为与电力同等重要的国家战略资源，这种**政府级采购**将有望在后期接力企业 CAPEX，拉长算力繁荣的周期。

所以在算力投资是一种技术正确与政治正确的情况下，我们并不担心全球整体的投资信心和趋势。因此，我们认为 2026 年的研究视点不应该纠结于“会不会投”或“投多少”，因为这是确定性最高的贝塔；真正的预期差则在于物理世界的供给瓶颈（如电力）以及需求结构变迁下的基建形态演变（训练 vs 推理）。

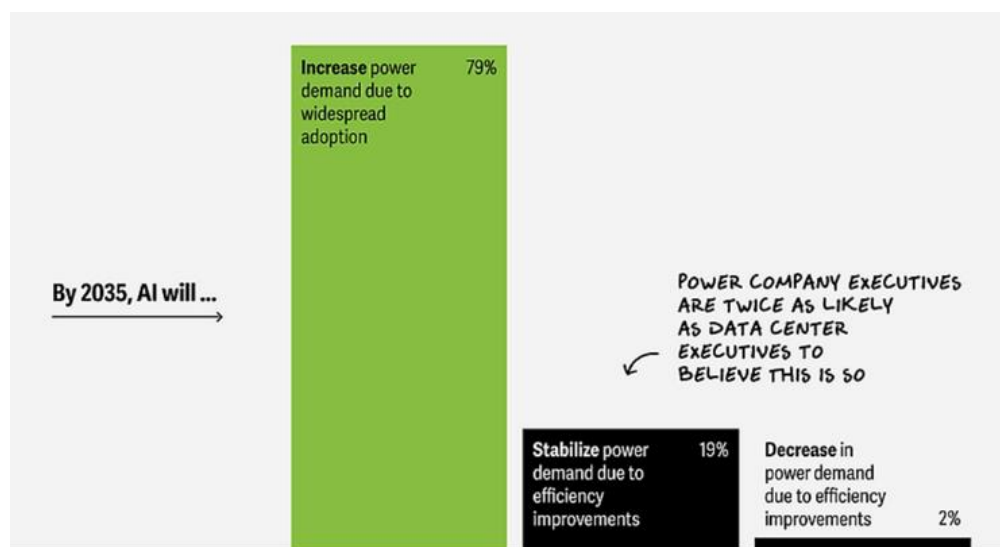
因此在北美电力供应已取代芯片产能，成为制约人工智能产业扩张的第一大瓶颈。核心矛盾在于 AI 算力的几何级增长与电网扩容的线性增长之间存在不可调和的时间错配。面对美国电网规划滞后引发的结构性供需错配，北美科技巨头正被迫重构能源策略，从单纯的电力采购方转型为能源基建方。“算力-能源”的双重垄断正在形成。根据德勤 2025 年的最新预测，美国 AI 数据中心的电力需求将经历一次“S 型”陡峭爬坡——从 2024 年的 4GW 激增至 2035 年的 123GW，增幅超 30 倍。这种增长并非全行业普涨，而是由 AI 驱动的结构替代。预计到 2035 年，AI 相关负载在数据中心总能耗中的占比将从当前的 12% 飙升至 70%。未来的数据中心将彻底异化为吞噬电力的“算力巨兽”。

图 21：美国人工智能数据中心的电力需求预计将激增



数据来源：德勤、东北证券

图 22：电力公司与数据中心高管看多人工智能对 2035 年电力需求的影响



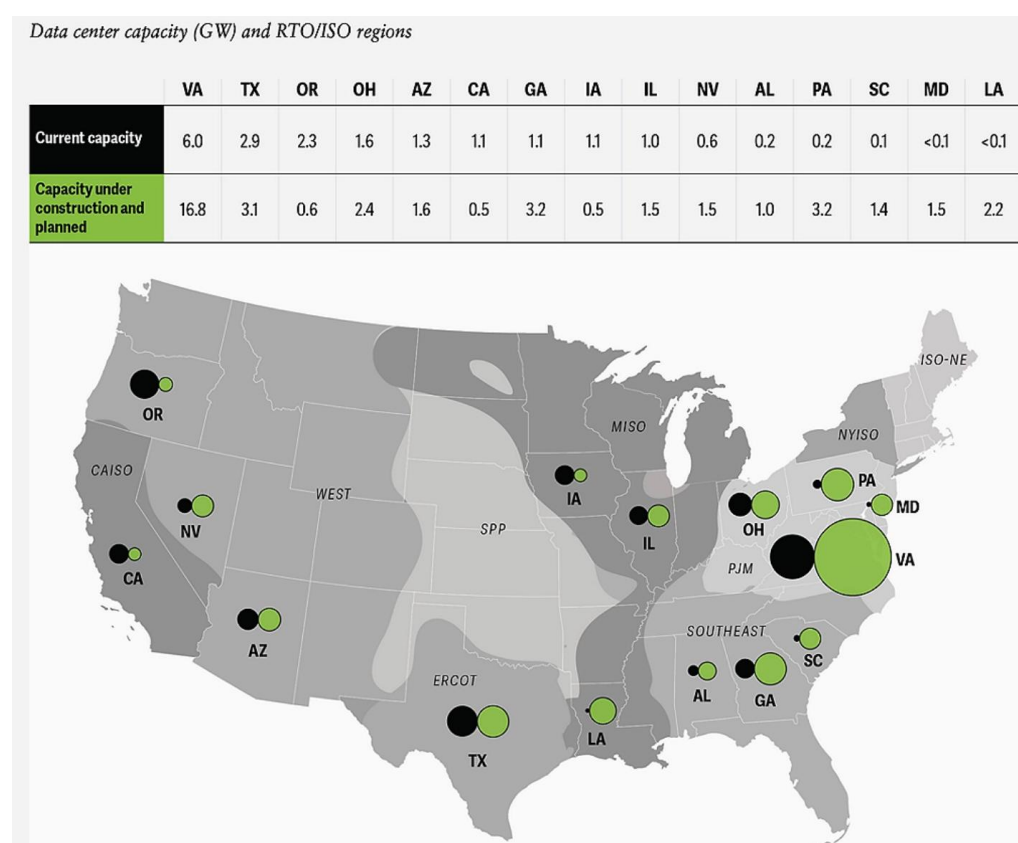
数据来源：德勤、东北证券

推理需求与电网瓶颈共振，驱动北美 AIDC 走向“去中心化”，现有电网的挑战已经开始从“点状负荷”到“面状冲击”。这种全天候运行的高密度电力集群正在给区域电网带来前所未有的物理挑战，部分核心区域已频繁出现谐波失真、负荷削减警告甚至局部停电事故，电网的承载能力已逼近物理极限。

在 AI 需求的增长下核心区的拥堵还在增加，弗吉尼亚州 (VA) 作为传统的“数据中心巷”，虽然规划容量仍高达 16.8 GW，但其电网接入等待时间已长达七年，拥堵效应显著。同时新兴区的需求也非常旺盛：为了规避核心区的电力瓶颈，算力正在向德克萨斯州 (TX, 规划 3.1 GW)、佐治亚州 (GA, 规划 3.2 GW)、俄亥俄州 (OH, 规划 2.4 GW) 以及宾夕法尼亚州 (PA, 规划 3.2 GW) 等能源充沛或靠近用户的地区扩散。这些地区的规划容量 (绿色部分) 往往数倍于现有容量 (黑色部分)，印证了算力基建正在经历一场全国范围内的再平衡。这是因为未来推理算力对“物

理距离”的敏感性肯定与训练算力可以集中在偏远沙漠不同，推理算力必须“逐人而居”。随着 AI 模型从云端大规模部署到更靠近用户的边缘侧，为了满足自动驾驶、实时语音交互等场景的毫秒级时延要求，基础设施开发不再局限于少数几个枢纽州，而是向全美更多州扩展。这种“应用侧的去中心化需求”叠加“供给侧的电力分布不均”，正共同重塑美国 AIDC 的地理格局，使其从“单极集中”迈向“多极分布式”的新时代。

图 23：北美各区已近建成数据中心及预期将建设数据中心的市場情况



数据来源：德勤、东北证券

同时，供给侧除了发电的供给错位以外，还存在一定的时间错配。建设一个 AI 数据中心需要 18-24 个月，但建设配套的高压输电网络需要 5-7 年（受限于许可审批、土地征收），因此这就成为了美国 AI 算力所面对的最大困难。

面对公共电网的脆弱性，北美科技巨头的能源策略已发生根本性转折：从被动的“绿电采购（PPA）”转向主动的“资产持有与建设”。这种“造电厂”的行为不仅是履行 ESG 承诺，更是为了确保业务连续性的生存之战。

- **策略一：重启核电与“表后取电”**：为了绕过漫长的电网排队（7 年），大厂开始直接将数据中心建在电厂旁边，通过物理专线直接取电，不再经过公共电网。
 - 案例：微软促成了著名的 **三里岛核电站（Three Mile Island）** 重启计划，签署了为期 20 年的购电协议，独占其 835MW 的电力。
 - 案例：亚马逊与 Talen Energy 合作，直接购买了紧邻宾夕法尼亚州核电站的数据中心园区，以获得稳定的基荷电力。
- **策略二：押注 SMR（小型模块化核反应堆）**：由于大型核电站建设周期太长，大厂开始投资下一代小型核堆。

- **案例：**谷歌与 Kairos Power 签署协议，计划在 2030 年代初部署多个 SMR 反应堆，总容量达 500MW。
- **案例：**亚马逊 领投了 X-energy，计划在 2039 年前上线超过 5GW 的 SMR 新产能。
- **策略三：自建燃气调峰电站**
 - 虽然大厂承诺“碳中和”，但在缺电的现实面前，天然气作为最可靠的过渡能源被重新启用。
 - **案例：**雪佛龙与 GE Vernova 合作，计划利用天然气轮机为数据中心提供“表后”供电，确保 24/7 不间断运行，不再看公用事业公司的脸色。

我们认为，基于北美科技巨头这种“绕过大电网、自建能源岛”的战略大转向，对于国内投资者而言，存在一定的产业链和映射机会。虽然中国企业直接参与美国核电或燃气电站建设的门槛极高（受限于地缘政治和资质），但在**配套设备、短缺物资以及替代能源方案**上，中国制造业的优势非常明显，建议关注**四大核心受益产业链**：

表 2：北美电力相关投资逻辑

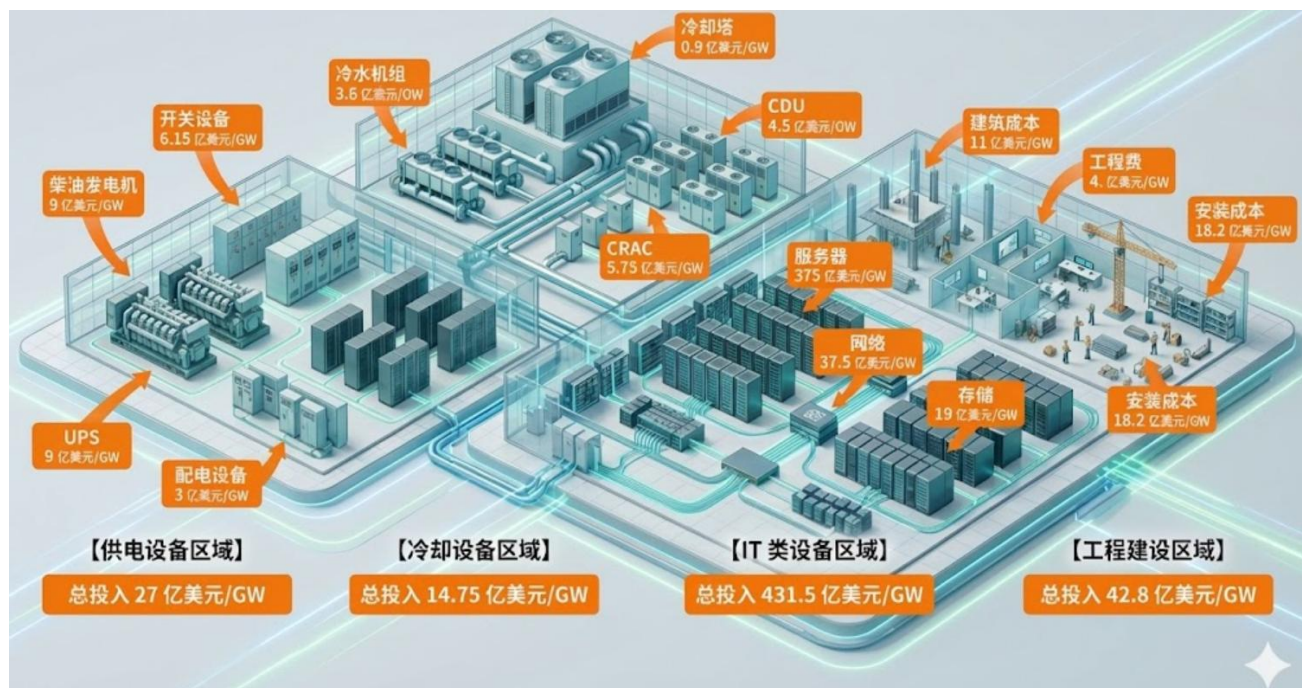
	核心驱动力	推荐关注细分
储能（ESS）	“表后”供电需要削峰填谷，保障 24/7 稳定性	大储集成商、液冷温控
电力设备	美国本土变压器产能不足，交付周期长达 3 年+	配电变压器、高压开关等
光伏（PV）	核电落地前的“快速部署”替代方案	拥有美国本土产能的组件厂
核电资源	全球科技巨头抢购核电导致铀资源短缺	锂矿、铀矿等稀有矿

数据来源：东北证券

3.2. “去英伟达化”的可能性：推理时代的供应链重构

在当前算力中心的建设投入结构中，IT 类设备无疑是成本占比的绝对大头。剖析当前千兆瓦级（1GW）算力中心的建设投入结构，IT 类设备（服务器、网络、存储）无疑是资金流向的绝对大头，构成了资本开支的核心主体。数据显示，在建设 1GW 数据中心的模型中，IT 类设备的投资高达 431.5 亿美元，其中仅服务器一项就占据了 375 亿美元。这一量级不仅形成了对供电设备（27 亿美元/GW）、冷却设备（14.75 亿美元/GW）及工程建设（42.8 亿美元/GW）的压倒性优势，更直接决定了项目的整体造价水位。尽管受 H100/Blackwell/TPU V7 等不同高性能算力卡价格成本有所区别的影响，具体数值会存在动态变化，但从行业通例来看，IT 类设备的投资额度通常稳居整体建设成本的 60% 以上，是算力中心经济模型中最为关键的变量。

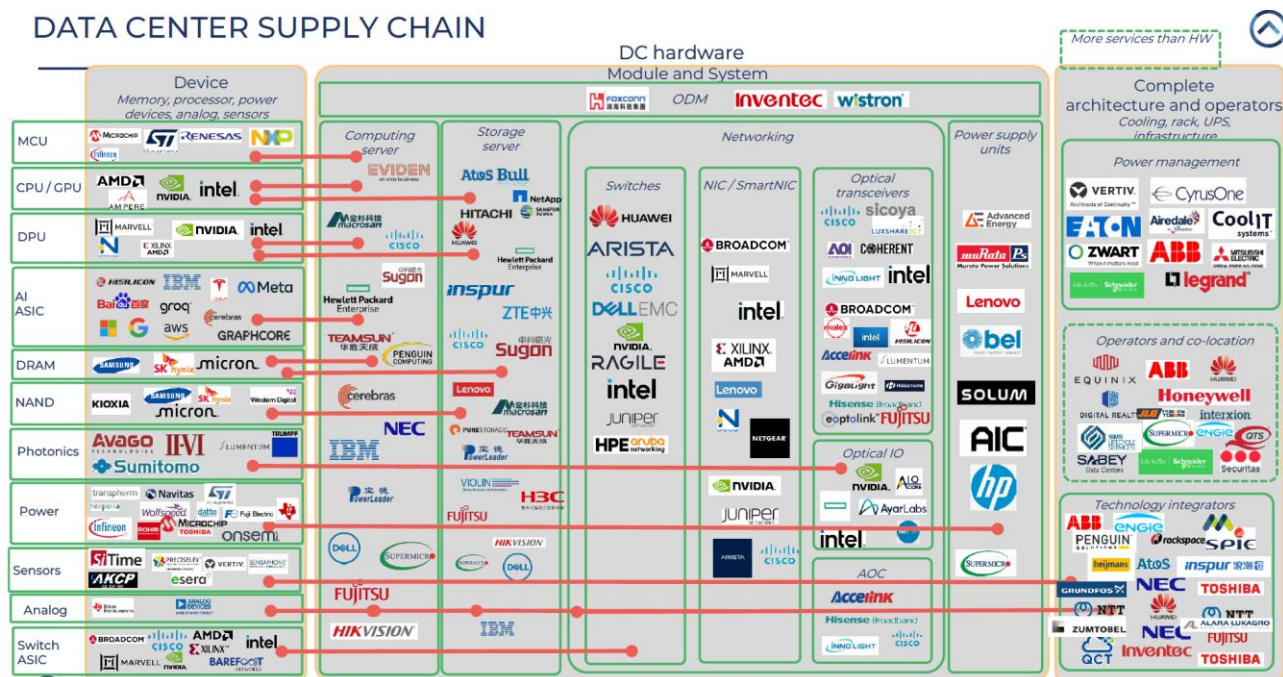
图 24：北美硬件商业投资闭环



数据来源：硅谷 101、东北证券

在追求极致算力密度的训练端，英伟达（NVIDIA）凭借其 H100/Blackwell 系列芯片的性能优势、CUDA 生态护城河以及供应链的产能封锁，依然是构建大模型底座不可替代的硬通货。英伟达 GPU 的强大，核心在于其构建了“算力性能、生态系统与迭代速度”三位一体的绝对壁垒，使其成为了 AI 时代的“物理基石”。在性能维度，从 Hopper 到即将量产的 Blackwell 架构，英伟达通过架构创新（如引入 Transformer 引擎）和对 CoWoS 先进封装、HBM 高带宽显存的极致应用，不断突破芯片的物理极限，H100/B200 始终定义着行业的最高算力标准。然而，更深的护城河在于其深耕近二十年的 CUDA 生态系统。这不仅是近 500 万开发者和数千个 AI 应用框架赖以生存的软件土壤，更形成了一种强大的网络效应——任何新的 AI 模型和算法几乎都首选在 NVIDIA GPU 上开发和优化，使得竞争对手即使在硬件参数上接近，也难以跨越软件兼容性的鸿沟。加之其堪称激进的产品迭代策略（如将发布周期从两年缩短至一年），英伟达始终保持着代际领先优势，从单一芯片供应商进化为涵盖网络、系统和软件的全栈式 AI 计算平台，确立了其在训练端几乎不可替代的霸主地位。因此我们认为，英伟达相关产业链的稳定性还将持续发展，在 2026 年，这些传统英伟达链（PCB、服务器、晶圆厂等）上的公司依旧可以有较强的产业生态。

图 25：全球数据中心供应链



数据来源：Yole、东北证券

表 3：传统英伟达链相关公司不完全整理

细分行业	核心公司名称
1. 服务器整机 & ODM	[A 股] 工业富联、浪潮信息、中科曙光 [美股] 戴尔 (DELL)、惠普企业 (HPE)、超微电脑 (SMCI)
2. PCB、板卡 & 载板	[A 股] 胜宏科技、沪电股份、深南电路、景旺电子
3. 光模块 & 光通信	[A 股] 中际旭创、新易盛、天孚通信 [美股] Lumentum (LITE)、Coherent (COHR)、博通 (AVGO)
4. 液冷 & 散热系统	[A 股] 英维克、高澜股份 [美股] 维谛技术 (VRT)
5. 电源管理系统	[A 股] 麦格米特、欧陆通、中恒电气
6. 晶元制造	[台股] 台积电 (TSM)、日月光 (ASX)、三星电子 (SSNLF)

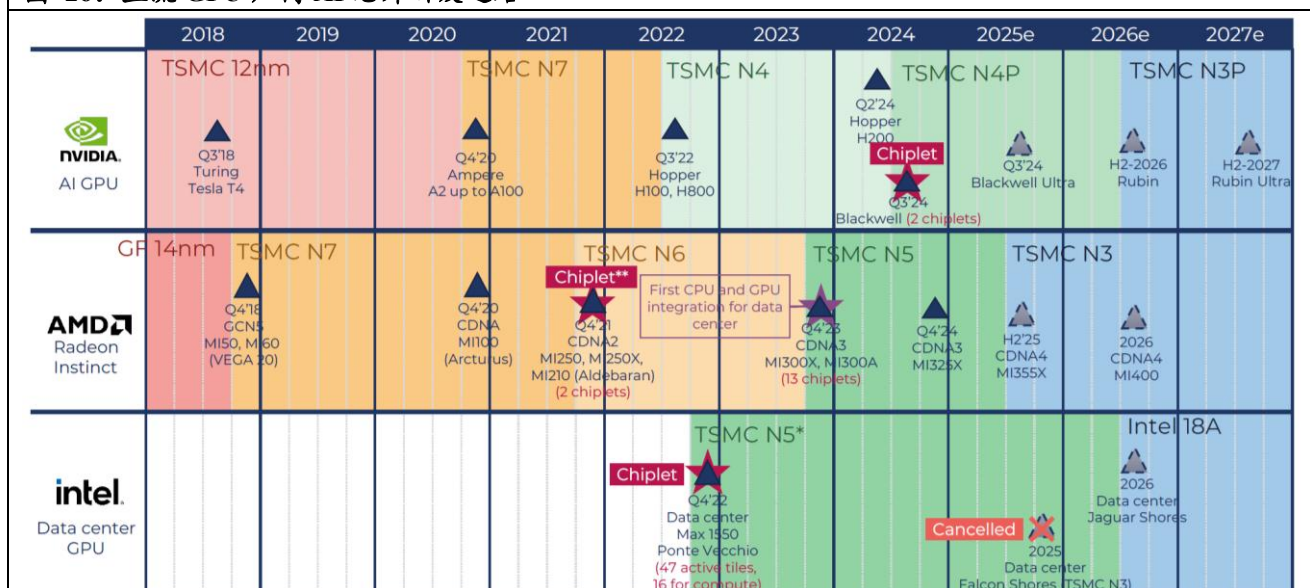
数据来源：东北证券整理

但同时，其他厂商的 AI 芯片研发也在继续。尽管英伟达凭借性能优势以及 CUDA 生态在训练端维持着绝对的统治力，但随着 AI 算力结构中推理需求占比突破 50% 的临界点，市场对于“极致性价比”与“供应链安全”的迫切诉求，正在这一垄断版图上撕开裂缝。

- **垂类突围：ASIC 的生存空间被打开。** 与训练端高度依赖 CUDA 不同，推理任务对通用生态的粘性显著降低。这为 Google TPU、AWS Trainium/Inferentia 以及 AMD 等自研芯片提供了巨大的战略窗口。云巨头推动这一进程的本质，是为了打破高昂的“英伟达税”，通过优化单位经济模型来实现商业闭环。
- **软件抹平：中间件重塑竞争规则。** 随着 PyTorch 2.0 和 OpenAI Triton 等标准化中间件的成熟，跨硬件迁移的摩擦成本大幅下降，“软件定义算力”正在成为现实。这意味着未来的行业赢家或许不再是单一的硬件寡头，而是那些能够屏蔽底层硬件差异、提供高效“异构算力调度”能力的平台型玩家。

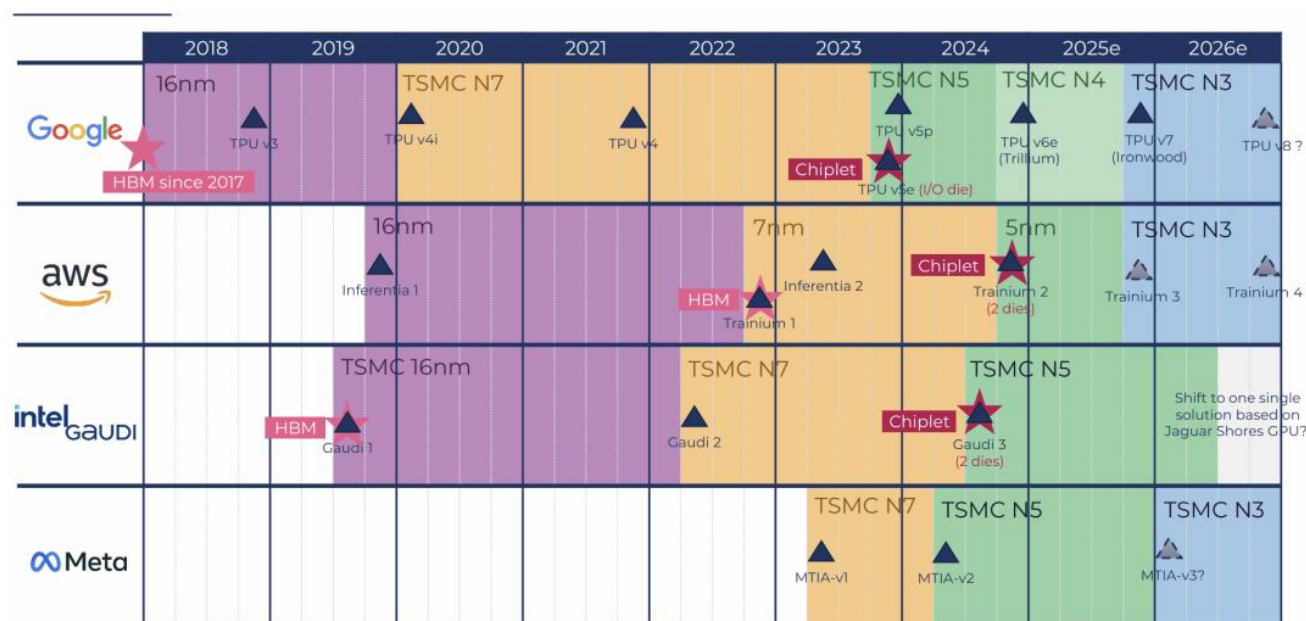
因此，其他厂商芯片“去英伟达化”在 2026 年还将加速。目前来看，AMD 是唯一的“通用军火商”：在这六家中，只有 AMD 是要把芯片卖给别人的（Dell, HP, 甚至卖给微软和 Meta）。MI355X 的 288GB 显存是其最大杀手锏，专门针对“显存墙”问题。：AWS 的 Trainium3 和 Google 的 Trillium 迭代速度非常快。AWS 甚至直接拿到了 3nm 产能，且深度绑定 Anthropic (Claude)，这构成了对 NVIDIA 最大的长期威胁。Meta 的 MTIA v2 非常务实，它不追求跑大语言模型（LLM）的通用能力，而是死磕“推荐算法”。这帮 Meta 省下了巨额的采购成本，因为推荐算法才是其最费算力的业务。相比 AWS 和 Google，微软的 Maia 芯片进展稍慢，二代芯片 Maia 200 已延期至 2026 年。

图 26：主流 GPU 厂商 AI 芯片研发之路



数据来源：Yole、东北证券

图 27：海外云计算厂商 AI 芯片研发之路



数据来源：Yole、东北证券

建议关注 2026 年同北美云厂商合作的新供应商。长期来看，出于对成本控制和供应链安全的渴求，下游市场将不可避免地撕开一道口子，为非英伟达阵营在推理侧

和垂直场景中留出巨大的生存空间。建议短期看封装产能的弹性（谁能拿到台积电以外的票）；中期看 ASIC 的放量（博通/联发科订单转化率）；长期看从北美新 AI 芯片的逐步放量中找到的新合作伙伴的未来市值增长。

3.3. 应用端爆发的潜力：System 2 带来的价值倍增

短期研判：幻觉与碎片的双重制约，新兴应用短期难以看到大规模爆发。在未来 1-2 年的短期窗口内，出现类似移动互联网时代微信、抖音那样“大一统”的国民级 AI 应用难度极大。这主要受限于两大结构性瓶颈：

- **信任鸿沟：**现阶段大模型普遍存在的“幻觉”问题，构成了其进入核心业务流的巨大阻碍。在金融交易、临床决策、法律合规等“零容错”场景中，企业无法完全放权给一个“概率性正确”的黑盒。因此，短期的应用形态将局限于“Human-in-the-loop（人在环路）”的辅助模式，而非完全的自动化替代。
- **需求碎片化：**不同于消费互联网标准化的“吃穿住行”需求，产业 AI 的需求呈现极度的非标化特征。一家律所的文档审查流程与一家医院的病历分析流程截然不同。这种“长尾且碎片”的需求结构，注定了短期内市场将被无数个针对特定场景优化的垂直 SaaS 所瓜分，而非由单一巨头垄断。

因此，短期来看，更多的应用方向还是去做传统人力的替代和提高效率，目前比较火的主要是在以下两个赛道。

- **多模态突围：**相比于对逻辑严谨性要求极高的文本场景，视频生成、创意设计、游戏制作等容错率较高的多模态领域，将率先迎来爆发。
- **垂直领域深耕：**在编程、药物研发等“规则明确、可验证”的专业领域，AI 将通过专项微调率先实现从辅助到替代的价值闭环，又比如 Top Apps 里面大部分都是 AI 编程也能说明 AI 在部分理科为主的行业中的优势。

图 28：火遍苏超的 AI 视频



数据来源：抖音、东北证券

不过随着 AI 产业步入深水区，应用端的锚点正逐步范式转移。从移动互联网时代迷信的“流量逻辑（DAU/MAU）”，彻底转向 AI 时代的“价值创造逻辑（ROI）”。

这一转变的核心引擎，正是具备 System 2（慢思考）能力的 AI 所带来的新价值创造。

- **高价值场景的闭环：**过去两年的 AI 应用多集中在“生成内容（AIGC）”，而未来的增量在于“解决问题”。具备 System 2 慢思考能力的智能体，能够深入医疗诊断、法律合规、工业控制等垂直领域，完成从“辅助人”到“替代人”的价值跃迁。这种高价值替代将彻底打通 B 端的付费意愿，验证算力投入的 ROI。
- **推理算力的乘数效应：**应用爆发不再仅仅意味着用户数的增加，更意味着单次交互算力消耗的指数级增长（OpenAI o1 模式）。因此，那些掌握核心数据资产、并能通过 Agent 重构工作流的垂直软件巨头，将承接算力基建后的最大一波价值红利。

因此建议重点关注以下具备高 ROI 潜力和重构工作流能力的新赛道与垂直行业：

- **高门槛垂直赛道的价值闭环：**在医疗诊断、法律合规及金融风控等高专业壁垒、高人工成本的领域，AI 正从单纯的“信息辅助”进化为具备逻辑推演能力的“决策代理”，通过完成从检索到深度风控建议的飞跃，实现高 ROI 的价值替代。
- **传统 SaaS 向 Agent 化的范式转移：**掌握核心行业数据资产的垂直软件巨头（如 CRM、法律软件、工业管理系统）正通过 Agent 重构工作流。这种转变将 AI 从边缘插件变为底层的生产力引擎，显著提升研发与运营效率，承接算力基建后的最大红利。
- **高收益应用点的规模化落地：**应用端的爆发重点将聚焦于“解决问题”而非“生成内容”。通过在智能客服、工业预防性维护及供应链优化等场景投入更高频次的“慢思考”算力，企业将获得显著的降本增效成果，从而打通 B 端市场的规模付费意愿。

图 29：具备高 ROI 潜力和重构工作流能力的新赛道与垂直行业



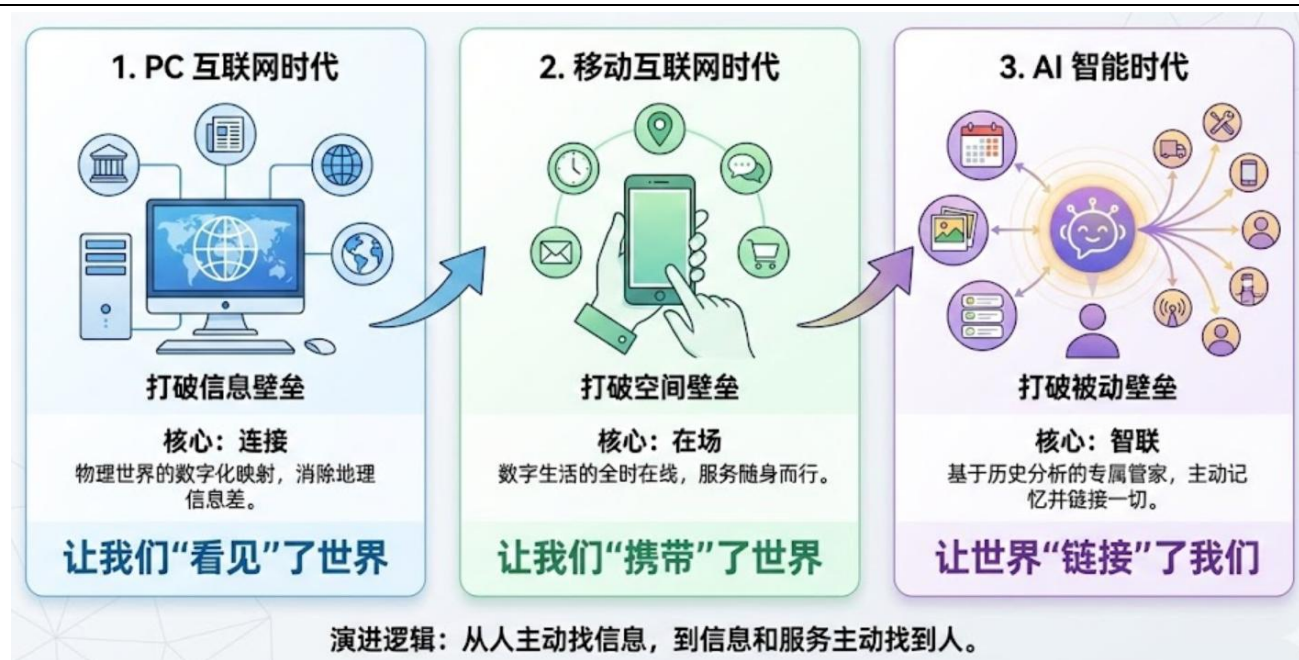
数据来源：东北证券制作

同时我们认为，新载体催生新物种，硬件变革是“普适化”的前提。历史规律表明，交互范式的革命往往滞后于技术突破，且必须依赖新型硬件载体的普及。目前的 AI 主要是以“插件”或“对话框”的形式寄生于手机和 PC 这两个旧时代的载体上，这极大地限制了其潜力的释放。

- **载体定义的交互革命：**
 - **PC 时代：**键盘+鼠标+固定屏幕，通过有线网络创造了门户网站与搜索需求；

- **移动互联网时代：** 多点触控+GPS+摄像头+随身携带，创造了 LBS（Uber/美团）、短视频（TikTok）等原生需求。
- **AI 时代：** 如果要出现真正的普适性 AI 应用，必须摆脱“低头戳屏幕”的交互桎梏。AI 眼镜、AI 吊坠（Pin）、具身机器人等新型硬件，将把交互从“二维屏幕”扩展到“三维物理世界”。

图 30：科技进化的三个阶段



数据来源：东北证券制作

新硬件创造新业务： 只有当硬件具备了“全天候感知”和“主动环境交互”的能力，AI 才能从一个“问答机器”进化为时刻在线的“全能管家”。这种“端侧硬件感知 + 云端 System 2 思考”的全新组合，将创造出当前手机上无法想象的全新下游需求（如实时视听翻译、物理世界导航、情绪感知伴侣等），从而真正引爆 AI 时代的“iPhone 时刻”。

表 4：AI 端侧相关公司不完全统计

细分行业	核心公司名称
1. 端侧 AI 芯片（算力底座）	【A 股】瑞芯微、晶晨股份、乐鑫科技、炬芯科技、寒武纪、恒玄科技、芯原股份 【美股】豪威集团（OmniVision）、高通（Qualcomm）
2. 端侧智能终端（落地载体）	【A 股】中兴通讯、歌尔股份、闻泰科技、蓝思科技、传音控股
3. 端侧算法 & 系统（软件适配）	【美股】Meta（AI 眼镜）、微软（AR 设备） 【A 股】中科创达、虹软科技、科大讯飞
4. 端侧光学组件（视觉感知）	【A 股】水晶光电、韦尔股份（旗下豪威 CIS）
5. 端侧通信 & 模组（连接支撑）	【A 股】美格智能、移远通信、广和通
6. 端侧配套支撑（散热 / PCB）	【A 股】中石科技（散热）、鹏鼎控股（PCB）、科翔股份（PCB）

数据来源：东北证券

4. 风险提示

1. 技术迭代不及预期

- **Scaling Law 边际效应递减：** 尽管报告指出 Scaling Law 目前尚未失效，但也提到单纯堆砌参数和算力的“暴力美学”投入产出比正在急剧下降。如果未来 Scaling Law 遭遇数学层面的天花板，或者数据枯竭导致模型能力无法通过算力线性提升，将直接打击整个行业的算力扩容逻辑。
- **推理侧“慢思考”成本过高：** 报告强调 2026 年是推理算力爆发的一年，OpenAI o1/o3 等模型通过消耗大量推理算力来换取智能。如果这种“昂贵的二次思考”模式无法大幅降低成本（目前 o3 单任务推理成本高达 20 美元），可能会导致商业化落地受阻，用户付费意愿不足。

2. 商业化落地不及预期

- **巨额资本开支的回报缺口：** 北美四大云厂商（Microsoft, Meta, Google, Amazon）单季资本开支超千亿美元，且处于“不计成本”的抢货阶段。市场普遍担忧存在巨大的 ROI 缺口。如果下游 AI 应用无法产生足够的收入来覆盖昂贵的折旧和运营成本，可能会导致巨头在 2026 年后大幅缩减资本开支，引发产业链剧震。
- **B 端应用信任鸿沟：** 现阶段大模型普遍存在的“幻觉”问题是进入金融、医疗等核心业务流的巨大阻碍。如果“幻觉”问题短期内无法解决，企业端（B 端）的付费意愿将局限于辅助角色，难以实现大规模的人力替代和高价值变现。

3. 配套设施建设不及预期

- **电力供应短缺与时间错配：** 北美 AI 数据中心电力需求激增，预计到 2035 年占比将达 70%。然而，电网扩容和输电网络建设需要 5-7 年，远慢于数据中心建设速度（18-24 个月）。这种时间错配可能导致已建成的算力中心因缺电而无法满负荷运转。
- **核电与新能源建设受阻：** 科技巨头纷纷押注核电（如三里岛重启、SMR）和天然气发电。但核电重启面临严格的监管审批、公众反对以及铀资源短缺的风险；天然气发电则面临碳中和政策的压力。如果能源解决方案落地受阻，将直接锁死算力增长的物理上限。

4. 行业竞争加剧与市场格局变化风险

- **ASIC 替代不及预期：** 报告提到云厂商正在推行“去英伟达化”，自研 ASIC（如 TPU、Trainium、Maia）。如果自研芯片的性能迭代（如软件栈兼容性、显存带宽）不及预期，或者英伟达通过加速产品迭代（Tick-Tock）巩固垄断，那么试图通过 ASIC 降低成本的策略将失败，导致算力成本居高不下。
- **价格战风险：** 中国市场已出现“价格崩塌”，DeepSeek 等模型通过极致性价比引发算力通缩。如果全球范围内模型厂商陷入长期的价格战，将压缩整个产业链的利润空间，导致中小厂商出清。

研究团队简介：

吴源恒：武汉大学遥感工程学士，伯明翰大学国际商务硕士，曾在湖北省测绘局从事测绘地理信息数据处理、遥感地理信息解译等工作，对卫星遥感应用产业有较深的认识。2020 年加入万联证券 TMT 团队，2022 年加入东北证券数字经济团队。目前主要研究 5G、卫星互联网、遥感、物联网、光通信等数字经济相关领域。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师。本报告遵循合规、客观、专业、审慎的制作原则，所采用数据、资料的来源合法合规，文字阐述反映了作者的真实观点，报告结论未受任何第三方的授意或影响，特此声明。

投资评级说明

股票 投资 评级 说明	买入	未来 6 个月内，股价涨幅超越市场基准 15%以上。	投资评级中所涉及的市场基准： A 股市场以沪深 300 指数为市场基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为市场基准；香港市场以摩根士丹利中国指数为市场基准；美国市场以纳斯达克综合指数或标普 500 指数为市场基准。
	增持	未来 6 个月内，股价涨幅超越市场基准 5%至 15%之间。	
	中性	未来 6 个月内，股价涨幅介于市场基准-5%至 5%之间。	
	减持	未来 6 个月内，股价涨幅落后市场基准 5%至 15%之间。	
	卖出	未来 6 个月内，股价涨幅落后市场基准 15%以上。	
行业 投资 评级 说明	优于大势	未来 6 个月内，行业指数的收益超越市场基准。	
	同步大势	未来 6 个月内，行业指数的收益与市场基准持平。	
	落后大势	未来 6 个月内，行业指数的收益落后于市场基准。	

重要声明

本报告由东北证券股份有限公司（以下称“本公司”）制作并仅向本公司客户发布，本公司不会因任何机构或个人接收到本报告而视其为本公司的当然客户。

本公司具有中国证监会核准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，本公司对这些信息的准确性和完整性不作任何保证。报告中的内容和意见仅反映本公司于发布本报告当日的判断，不保证所包含的内容和意见不发生变化。

本报告仅供参考，并不构成对所述证券买卖的出价或征价。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证券买卖建议。本公司及其雇员不承诺投资者一定获利，不与投资者分享投资收益，在任何情况下，我公司及其雇员对任何人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。

本公司或其关联机构可能会持有本报告中涉及到的公司所发行的证券头寸并进行交易，并在法律许可的情况下不进行披露；可能为这些公司提供或争取提供投资银行业务、财务顾问等相关服务。

本报告版权归本公司所有。未经本公司书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用。如征得本公司同意进行引用、刊发的，须在本公司允许的范围内使用，并注明本报告的发布人和发布日期，提示使用本报告的风险。

若本公司客户（以下称“该客户”）向第三方发送本报告，则由该客户独自为此发送行为负责。提醒通过此途径获得本报告的投资者注意，本公司不对通过此种途径获得本报告所引起的任何损失承担任何责任。

东北证券股份有限公司

网址：<http://www.nesc.cn> 电话：95360,400-600-0686 研究所公众号：dbzqyanjiusuo

地址	邮编
中国吉林省长春市生态大街 6666 号	130119
中国北京市西城区锦什坊街 28 号恒奥中心 D 座	100033
中国上海市浦东新区杨高南路 799 号	200127
中国深圳市福田区福中三路 1006 号诺德中心 34D	518038
中国广东省广州市天河区冼村街道黄埔大道西 122 号之二星辉中心 15 楼	510630

