



计算机行业研究

买入（维持评级）

行业周报

证券研究报告

计算机组

分析师：刘高畅（执业 S1130525120005）
liugaochang@gjzq.com.cn

分析师：陈芷婧（执业 S1130525120008）
chenzhijing@gjzq.com.cn

联系人：孙恺祈
sunkaiqi@gjzq.com.cn

国内算力斜率陡峭

本周观点

- **AI 入口之争，大厂投资力度再增强。**
- 1) 中国 AI 全球存在感大幅提升，国内大模型迭代不断。据 AA，尽管 GPT-5.2 与 Gemini 3 Pro 仍占据塔尖，但中国国产模型已有效改变了北美单极主导的竞争格局。在全球 Top 10 阵营中，GLM-4.7、DeepSeek V3.2、Kimi K2 Thinking 已占据 3 席；若将观测范围扩大至 Top 15，中国企业席位达 6 个，2025 年中国开源 AI 模型调用量约占全球市场的 70+%。2) 推理需求激增：o1 类推理模型的出现，使推理阶段的计算量（Inference-time Compute）相对传统模型解锁了约 10 倍的潜力，算力需求已从单一的“训练驱动”转向“训练+推理双轮驱动”。3) 流量入口重塑：入口不再仅限于手机，而是演变为“OS 级智能体”与“超级 APP”层面较量。2025 年 12 月 24 日，字节跳动旗下 AI 应用豆包宣布日均活跃用户数（DAU）突破 1 亿；千问 App 近期持续扩大投流，截止 12 月 10 日（公测 23 天）月活已突破 3000 万，下成全球增长最快 AI 应用。豆包绕开传统接口，用“看屏幕+代操作”的方式在现有生态之上再搭一层“AI 操作系统”，AIOS 直接触碰微信、支付宝等超级 App 的商业命门，传统 App 时代的游戏规则面临挑战，微信、淘宝等超级 App 陆续封禁豆包手机权限。在此背景下，我们认为入口之争的背后是流量之争，直接指向大型互联网企业广告及电商业务的基本盘，2026 年各家在模型、AI 应用产品上的算力投入将进入白热化的锦标赛阶段。
- **国产卡持续突破，2026 年爆发之时。**
- 1) 智算中心持续扩容，国产替代加速：2020-2028 年中国智能算力规模预计保持 57% 复合增长率。2) 国产通用 GPU 从“可用”向“好用”升级：国产算力芯片在工艺与架构上持续突破，在处理大模型长文本、复杂算子融合等方面的表现已显著缩小与国际巨头的差距，以中芯国际为代表的本土晶圆厂产能/利用率持续高位运行，为国产 AI 芯片的产能提供了坚实保障。3) CSP 厂商加速适配，助力国产芯片生态建设：腾讯云宣布全面适配主流国产芯片，百度、阿里等企业加速适配国产芯片，推动“芯片-模型-应用”闭环形成。
- **供需双高，国内算力斜率陡峭。**
- 1) 需求侧：①下游推理需求伴随 AI 手机等终端 AI 进入落地放量临界点确定性增强；②大模型 Scaling Law 仍然有效，模型迭代持续加速提升训练需求的确定性。2) 供给侧：①国产 GPU 性能持续提升；②国内 CSP 厂商加速适配国产 GPU。上述背景下，GPU、服务器、柴发、电源、IDC、云 6 大产业环节供需两高，其中，AI 服务器整合 GPU、存储等直接影响模型效果的硬件基础；AIDC 承载算力与数据；AIDC 电源是 AIDC 的主要供电系统；柴油发电机是 AIDC 不可或缺的备用电力来源；大模型公有云本质上是算力的共享化与弹性化。

投资建议

- **相关标的：**
国内算力：寒武纪、东阳光、海光信息、协创数据、星环科技、神州数码、百度集团、大位科技、润建股份、华丰科技、中芯国际、华虹半导体、兆易创新、大普微、中微公司、兴森科技、中科曙光、禾盛新材、润泽科技、浪潮信息、东山精密、亿田智能、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、欧陆通、杰创智能。
海外算力/存储：中际旭创、新易盛、天孚通信、源杰科技、胜宏科技、景旺电子、英维克等；闪迪、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

风险提示

- 行业竞争加剧的风险；技术研发进度不及预期的风险；特定行业下游资本开支周期性波动的风险。



内容目录

一、AI 入口之争，大厂投资力度再增强	4
二、国产卡持续突破，2026 年爆发之时	11
三、供需双高，国内算力斜率陡峭	13
四、相关标的	17
风险提示	17

图表目录

图表 1: 中国 AI 在全球顶级模型排名中位居中段	4
图表 2: 北美 AI 玩家持续引领；中国 AI 玩家加速追赶；欧洲创企表现不俗	4
图表 3: 中国开源 AI 模型调用量约占全球市场的 70+%，DS 一家约占 40+%	5
图表 4: Hugging Face 上主要开源权重 LLMs 的累积下载量	6
图表 5: ChatBot Arena 中得分最高的开源权重模型,截止 12 月 4 日	6
图表 6: 三大 Scaling 定律	7
图表 7: Interleaved Thinking 推理范式	7
图表 8: Hopper (H100)与 Blackwell(B200/GB300) 对比	8
图表 9: 推理相对非推理解锁了 10X Token 消耗	9
图表 10: Groq 的芯片专为推理设计	10
图表 11: 微信淘宝等超级 APP 封禁豆包手机	10
图表 12: 豆包手机扩充 VIVO、联想、传音等合作商	10
图表 13: 终端 AI 有望依托 AI 手机生态持续繁荣	11
图表 14: 2020-2028 年中国智能算力规模及预测	11
图表 15: 2024 年中国 AI 加速芯片市场竞争格局	11
图表 16: 国产通用 GPU 从“可用”向“好用”升级	12
图表 17: 中芯国际产能/利用率持续提升	12
图表 18: 腾讯云宣布全面适配主流国产芯片	13
图表 19: 供需两高，国产算力斜率陡峭	14
图表 20: 预计未来推理型服务器将逐渐成为市场主流	14
图表 21: 2024 年我国 x86 服务器市场 CR6	14
图表 22: AIDC 电源是专为人工智能数据中心设计的供电系统	15
图表 23: AIDC 电源主要玩家一览	15
图表 24: 柴油发电机组在数据中心作为备用电源使用	15
图表 25: 2024 年国产 IDC 柴发机组份额约为 30%	15



图表 26: 不同主体的 AIDC 的功能定位与布局诉求存在差异	16
图表 27: 我国公有云大模型调用量快速增长	16
图表 28: 2025 上半年我国大模型公有云服务市场格局	17



一、AI 入口之争，大厂投资力度再增强

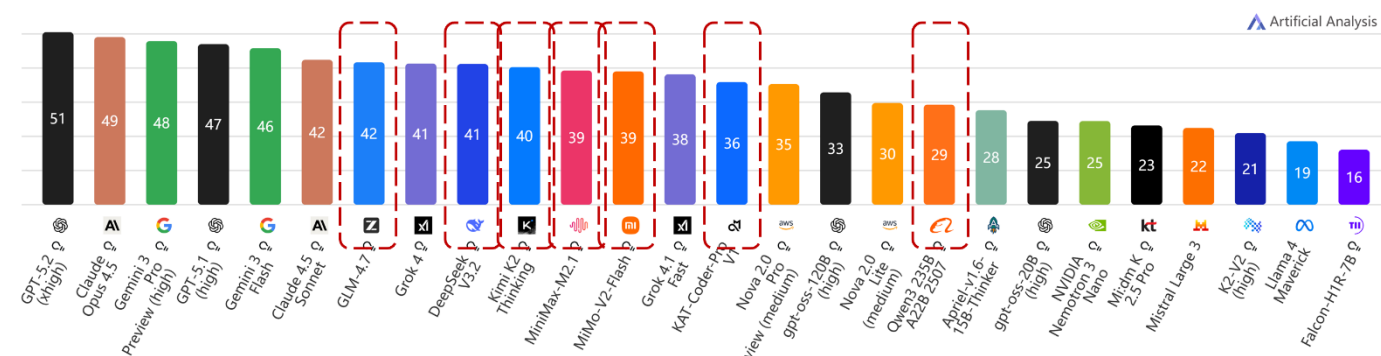
模型迭代持续精彩：中国 AI 全球存在感大幅提升，国内大模型迭代不断。

中国 AI 在全球顶级模型排名中位居中端。据 AA，尽管 GPT-5.2 与 Gemini 3 Pro 仍占据塔尖，但中国国产模型已有效改变了北美单极主导的竞争格局。在全球 Top 10 阵营中，GLM-4.7、DeepSeek V3.2、Kimi K2 Thinking 已占据 3 席；若将观测范围扩大至 Top 15，中国企业席位达 6 个。

图表1：中国 AI 在全球顶级模型排名中位居中段

Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, r^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

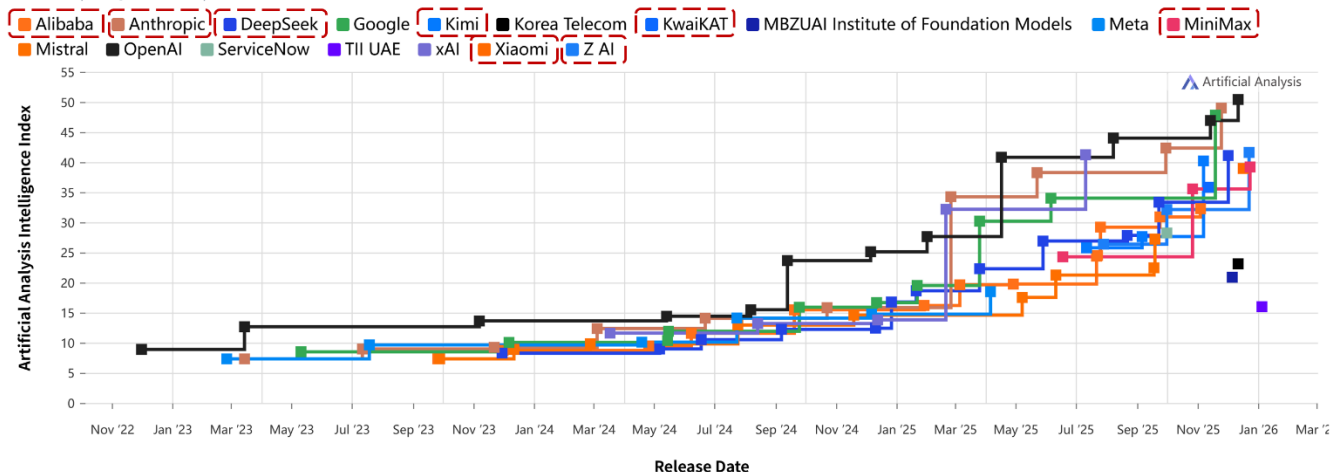


来源：Artificial Analysis 官网，国金证券研究所（注：①数据及信息截至 2026 年 1 月 9 日；红框内为中国公司；② Intelligence Index 综合考虑了 MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, r^2 -Bench Telecom 十类测评基准。）

图表2：北美 AI 玩家持续引领；中国 AI 玩家加速追赶；欧洲创企表现不俗

Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, r^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt



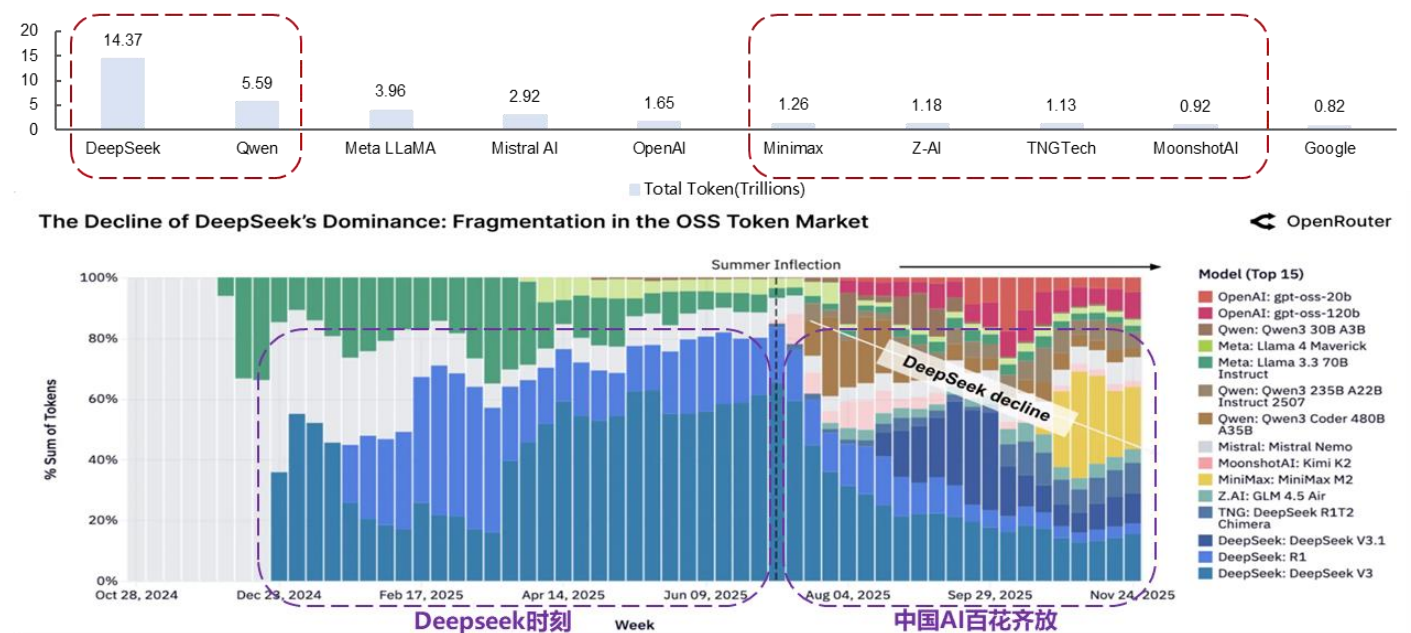
来源：Artificial Analysis 官网，国金证券研究所（注：①数据及信息截至 2026 年 1 月 9 日；红框内为中国公司；② Intelligence Index 综合考虑了 MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, r^2 -Bench Telecom 十类测评基准。）

2025 年中国开源 AI 模型调用量约占全球市场的 70+%。据《State of AI: An Empirical 100 Trillion Token Study with OpenRouter》，DeepSeek 与 Qwen 分别以 14.37 万亿及 5.59 万亿 Token 消耗量领跑开源市场，这一数据远超同口径下 OpenAI（1.65 万亿）与 Google（0.82 万亿）的表现。去年 8 月份之后，国产开源模型厂商竞相繁荣，驱动中国



厂商整体份额稳中有进，合计占据了全球超过 70% 的开源份额。

图表3：中国开源 AI 模型调用量约占全球市场的 70+%，DS 一家约占 40+%



来源：《State of AI: An Empirical 100 Trillion Token Study with OpenRouter》，国金证券研究所

DeepSeek V3.2 系列模型提出多项创新：延续使用 DSA 稀疏注意力机制、加码后训练、大规模合成数据。

DSA 全称 DeepSeek Sparse Attention，旨在降低计算复杂度，同时在上文场景中还能保持模型性能。举个例子，当你面前有一本超厚的书（假设页数上万页），现在需要从这本书里查询特定信息，要求只能一页一页地看，且需要记忆看过的内容，这就是传统 AI 模型处理长文本时的困境——计算量随文本长度呈平方级增长。而 DSA 的作用如同将这种“大海捞针”的方式转为“精准定位”，不是每页都要看，而是能够只关注与目标信息最相关的几页，大幅提升了模型记忆与理解效率。

此外，DeepSeek 意识到开源模型的后训练投入普遍不足，限制了模型任务表现，于是采用更激进的方式，将后训练算力预算调整到超过预训练成本的 10%。

DeepSeek 这次用到了大规模智能体数据合成，根本原因在于缺乏足够多样化的真实训练环境，无法让智能体具备任务泛化能力。从测试评估结果看，使用合成数据使模型泛化能力提升显著。

12 月 16 日斯坦福大学发布研究报告指出，在能力与采用率方面，中国的 AI 模型尤其是开放权重的大语言模型，已接近甚至部分领先于国际先进水平。

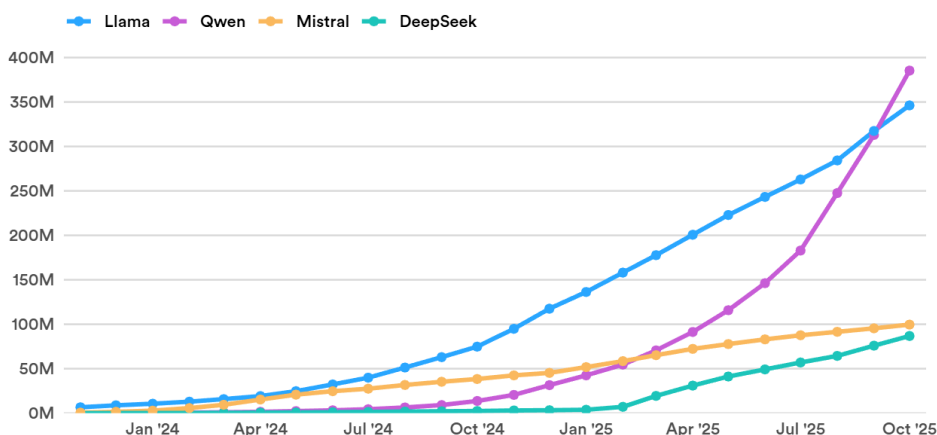
2025 年 9 月，阿里巴巴的 Qwen 模型家族超越 Llama，成为 Hugging Face 上下载量最多的 LLM 家族。2024 年 8 月至 2025 年 8 月期间，中国开源模型开发者占 Hugging Face 所有下载量的 17.1%，略超过美国开发者，后者占下载量的 15.8%。自 2025 年 1 月起，基于阿里巴巴和 DeepSeek 发布的开源模型的上传衍生模型已超过基于美国和欧洲主要模型的衍生模型。并且在 2025 年 9 月，中国微调或衍生模型占 Hugging Face 上发布的所有新微调或衍生模型的 63%。



图表4: Hugging Face 上主要开源权重 LLMs 的累积下载量

Figure 1: Cumulative downloads of major open-weight LLMs on Hugging Face (November 2023 – October 2025)

Source: Hugging Face via the ATOM Project



来源: 斯坦福报告, 国金证券研究所

中国的开放权重模型现在在主要基准和排行榜上表现接近顶尖水平, 涵盖通用推理、编码和工具使用。在 Chatbot Arena 这个著名的用户比较和评分排行榜上, 中国生产了顶尖的开放权重模型, 而且其中最好的模型几乎与美国公司领先的闭源模型不相上下。

图表5: ChatBot Arena 中得分最高的开源权重模型, 截止 12 月 4 日

Rank	Model	Lab	Country	Elo score*	Coding score	Vision score	AAII score†	MMLU-Pro score§
1	GLM-4.6	Z.ai	China	1442	1458		56	83.5
1	Kimi-K2-Thinking	Moonshot	China	1438	1450		66	84.8
2	GLM-4.5	Z.ai	China	1430	1448		54	83.5
2	Mistral Large 3	Mistral	France	1429	1450		40	81
2	Qwen3-VL-235B-A22B-Instruct	Alibaba	China	1427	1457	1246	49	82.8
2	DeepSeek-R1-0528	DeepSeek	China	1426	1436		57	84.9
2	DeepSeek-V3.2-Exp-Thinking	DeepSeek	China	1421	1438		58	85.1
2	LongCat-Flash-Chat	Meituan	China	1420	1460		49	82.7
2	Qwen3-235B-A22B-Instruct-2507	Alibaba	China	1418	1457		49	82.8
2	DeepSeek-V3.2-Exp	DeepSeek	China	1418	1431		47	83.6
2	DeepSeek-V3.1	DeepSeek	China	1418	1430		47	83.3
2	Qwen3-Next-80B-A3B-Instruct	Alibaba	China	1417	1456		57	82.4
2	DeepSeek-V3.1-Thinking	DeepSeek	China	1417	1437		58	85.1
2	Qwen3-235B-A22B-Thinking-2507	Alibaba	China	1416	1442		62	84.3
3	Qwen3-VL-235B-A22B-Thinking	Alibaba	China	1411	1432	1215	62	84.3
3	GLM-4.5-Air	Z.ai	China	1386	1410		47	81.5
3	Qwen3-30B-A3B-Instruct-2507	Alibaba	China	1382	1425		44	77.7
3	Kimi-K2-0905-Preview	Moonshot	China	1382	1403		49	82.4
3	Qwen-VL-Max-2025-08-13	Alibaba	China	1381	1440	1213		
3	Kimi-K2-0711-Preview	Moonshot	China	1380	1402		47	82.4
3	DeepSeek-V3-0324	DeepSeek	China	1377	1391		42	81.9
3	DeepSeek-R1	DeepSeek	China	1373	1382		48	84.4
3	Qwen3-235B-A22B	Alibaba	China	1369	1394		46	82.8
3	gpt-oss-120b	OpenAI	US	1368	1398		59	80.8
3	Qwen3-Coder-480B-A35B-Instruct	Alibaba	China	1358	1406		43	78.8

来源: 斯坦福报告, 国金证券研究所

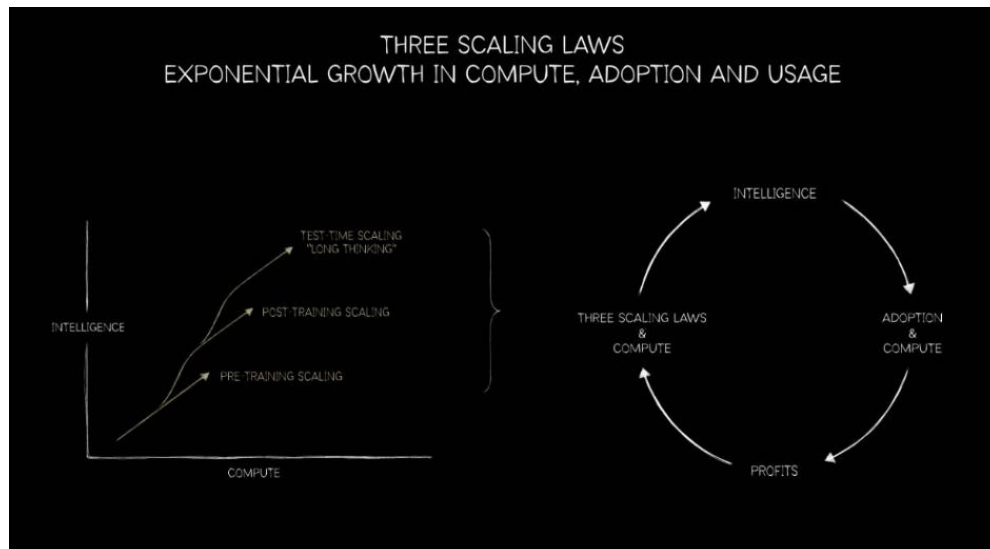
Scaling-law 仍然有效: 三大 Scaling 持续演绎, DS 发布 mHC 提升训练稳定性。

大模型三大 Scaling 定律: 预训练 (pre-training) / 后训练 (post-training) / 推理深度思考 (test-



time long thinking)。算力与应用循环促进：算力 Scaling—>智能提升（Intelligence）—>应用的广泛采用（Adoption）—>经济效益（Profits）—>算力 Scaling。GPTo1 之后，模型 Scaling law 从单一的预训练（pre-training）向三大 Scaling 转变。1）模型在回应用户之前产生很长的内部思考链，思考时间越长，答案的质量就越高；2）Multi-agent 进一步提升模型性能，但其 Token 消耗量往往达到对话聊天的数倍。

图表6：三大 Scaling 定律

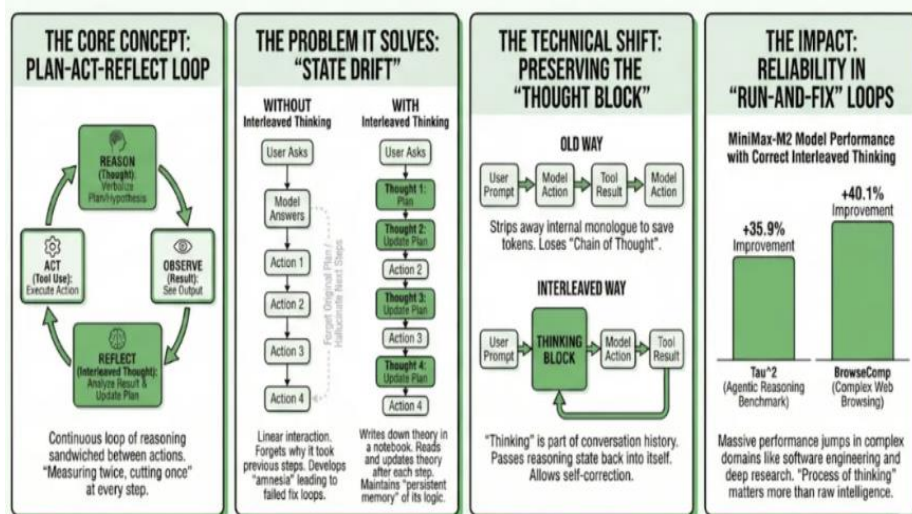


来源：英伟达官网，国金证券研究所

训练层面，预训练重启，后训练加码。预训练方面，DeepSeek 团队在论文中坦言，受训练算力约束，DeepSeek-V3.2 的世界知识广度仍落后于领先闭源模型（如 Gemini 3 Pro），后续将通过扩大预训练规模补齐能力上限。此外，V3.2 往往需要更多 token 才能逼近 Gemini-3.0-Pro 等模型的输出质量，未来将聚焦提升推理链的“智能密度”，以更少 token 达到同等效果；后训练投入强度提升：DeepSeek 认为开源模型后训练投入普遍不足并制约任务表现，因此将后训练算力预算上调至超过预训练成本的 10%。

推理方面，MiniMax M2、DeepSeek V3.2 等国产大模型，把“思考 → 行动 → 观察 → 再思考”的动态循环模式（Interleaved Thinking）融入推理流程，正成为 Agent 模型标配。我们认为，国内 Interleaved Thinking 范式渗透，能够显著提高 Agent 的准确性和规划能力，2026 年 Token 消耗与推理需求有望加速爆发。

图表7：Interleaved Thinking 推理范式



来源：英伟达官网，国金证券研究所

我们认为模型的进步目前并未到瓶颈，2026 年模型的进展会更值得期待：

1）靠 Scaling law 提升模型能力的路径依旧有效。



DeepSeek V3.2 后训练规模扩展到预训练规模 10%，预计未来比例还会提高，而且 DeepSeek 团队在论文表示由于训练算力有限，DeepSeek-V3.2 的世界知识广度还是落后于 Gemini 3 pro 这样的顶尖闭源模型，团队计划未来进一步扩大预训练规模。同时 DeepSeek 大量使用合成数据有效说明不用担心数据会遇到瓶颈。

2) DeepSeek-V3.2 提出的 DSA 机制展示出强大算法创新能力，不必担心大模型技术创新已经到达瓶颈。

3) 大模型训练的硬件基础正从英伟达的 Hopper 架构转向 Blackwell 架构，Blackwell 相比前代在单卡算力、显存带宽、显存容量、以及集群互联都大幅提升，这对大模型训练的意义一方面是加速和降低成本，另一方面是可以使用更大的 Batch Size（模型更新学习内容前一次性处理的训练样本数量），这对训练稳定性有帮助，更大的 Batch Size 能够更准确地估计整个数据集的梯度，从而使学习过程更加稳定，而较小的批次则会产生噪声过大且特征过于明显的信号，这可能导致模型的学习路径出现不稳定的跳跃。

图表8: Hopper (H100)与 Blackwell(B200/GB300) 对比

特征	NVIDIA HOPPER (H100)	NVIDIA BLACKWELL (B200/GB300)
建筑概论	Hopper (H100 Tensor Core, 2022)	Blackwell (B200/GB300 Tensor Core, 2024/25)
每个GPU的SM数量	约80个SM (GH100)	每个芯片约 80 个 SM (B200), 2 个芯片共封装 (nm)
晶体管 (总数)	~80B (整体式) (nm)	~208B (2×芯片组) (nm)
进程节点	台积电 4N (定制 5nm)	台积电 4NP (新型 4nm 变体) (nm)
芯片封装	单GPU芯片	双GPU芯片组, 带10 TB/s NVLink芯片间链路 (nm)
每个SM的共享/L1内存	256 KB (可配置)	128 KB (更小, 统一) (nm)
统一的 L2 缓存	总计 50 MB (分块)	65 MB 单体式 (nm)
指令缓存	每个SM的L0缓存	更大的统一 L0 (减少指令获取延迟)
HBM 容量	80 GB HBM3 (nm)	192 GB HBM3e (B200) / 288 GB (GB300) (nm) (nm)
HBM带宽	3.35 TB/s (nm)	8.0 TB/s (nm)
NVLink (芯片间)	NVLink4 - 900 GB/s	NVLink5 - 1.8 TB/s (nm)
NVSwitch (DGX pod 中的链接)	~50 TB/s (在 16 路 GPU 的 DGX A100 上)	每个 72-GPU NVL72 pod 的处理能力为 130 TB/s (nm)
张量核心精度 (密集)	FP16/BF16、FP8、INT8	FP16/BF16、FP8、FP6、FP4 (nm) (nm)
FP8/密集型 TFLOPS	峰值约为 4-5 PF	~9 PF (nm) (每个 GPU)
FP4/密集型 TFLOPS	(不支持)	~18 PF (nm) (每个 GPU, 新模式)
FP64 TFLOPS (张量)	60 TF (张量) (nm)	较低 (弱化) (nm)
峰值功率 (TGP)	约700 瓦 (nm)	约 1,200-1,400 瓦 (nm)

来源: intuitionlabs.ai, 国金证券研究所

2026 年 1 月 1 日，DeepSeek 在 AI 开源社区 HuggingFace 和研究分享平台 arXiv 发布论文，提出了名为 mHC (Manifold-Constrained Hyper-Connections) 的新型神经网络架构优化方案。mHC 通过数学上的流形约束 (Manifold Constraint) 强制信号在跨层传播中保持范数守恒。这成功解决了传统超连接 (HC) 架构虽能带来性能提升，但因信号发散导致训练不稳定的痛点。扩展残差流会增加显存和通信的压力，同时流形约束数学理论的落地也较复杂，DeepSeek 进行了大量底层算子和通信优化，再一次展现了极高的软硬结合能力。mHC 解决了超宽残差流的训练稳定性难题，为训练更大参数规模的基模扫清了架构障碍，我们预计 DeepSeek 下一代更大的基模正在酝酿中。

推理需求高增：推理模型解锁 10X 算力需求，英伟达或将收购 Groq 布局推理算力技术。

推理相对非推理解锁了 10X Token 消耗。随着大模型参数边际效应的初现，技术竞争的关键变量开始向后训练发生结构性转移。Post-training 将训练从通用的参数堆叠转向了基于高质量私有数据的监督微调与人类反馈强化学习，训练过程需要引入海量由高算力生成的合成数据进行反复迭代。在这一阶段，模型在响应用户指令之前，会先在内部生成思维链，自主进行策略试错、逻辑推演与自我纠错，从而显著提升了解决复杂数学与逻辑问题的准确率。据 AA，推理模型相对非推理模型在 Token 消耗量上出现 10X 提升。

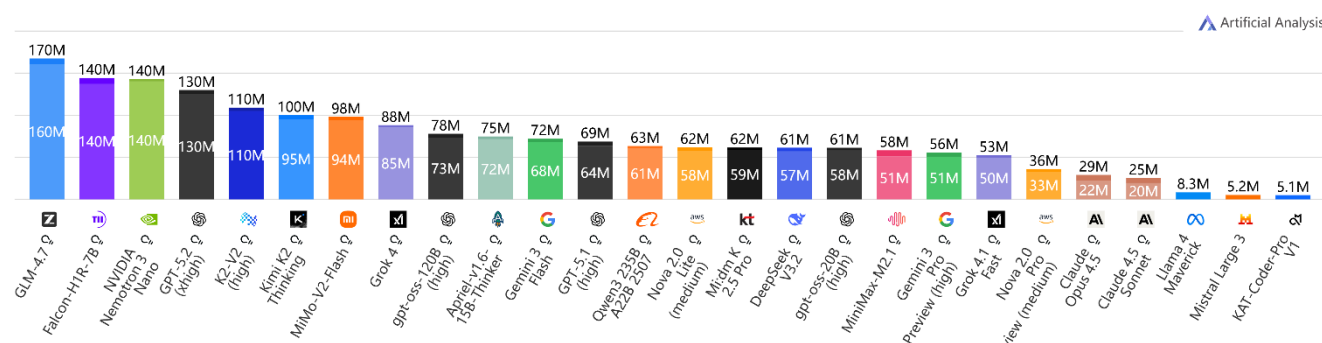


图表9：推理相对非推理解锁了 10X Token 消耗

Output Tokens Used to Run Artificial Analysis Intelligence Index

Tokens used to run all evaluations in the Artificial Analysis Intelligence Index

Answer Tokens Reasoning Tokens



来源：Artificial Analysis 官网，国金证券研究所

据界面新闻报道，美东时间 2025 年 12 月 24 日，AI 芯片初创公司 Groq 宣布已与英伟达就其推理技术达成非独家许可协议。根据协议条款，Groq 创始人 Jonathan Ross、总裁 Sunny Madra 及其他核心团队成员将加入英伟达，共同推进授权技术的升级与规模化应用。

Groq 将继续作为独立公司运营，Simon Edwards 将接任首席执行官职务。Groq 云服务将保持正常运行，不受此次合作影响。

值得注意的是，此前有报道称，英伟达将以大约 200 亿美元收购人工智能芯片初创公司 Groq，这将是英伟达迄今为止规模最大的一笔收购。但英伟达很快表示，并没有收购 Groq，只是达成了技术许可，使用 Groq 的推理技术而已。

公开资料显示，成立于 2016 年的 Groq 由谷歌 TPU 核心开发者 Jonathan Ross 创立，公司自研的 LPU 推理芯片是本次合作的核心价值所在。

区别于英伟达通用型 GPU，LPU 专为 AI 推理场景深度优化，凭借确定性架构、片上 SRAM 内存设计等核心技术，实现了超低延迟、超高能效与极速推理速度。在主流大语言模型运行中，LPU 推理速度可达英伟达 H100 GPU 的 5 至 18 倍，首 token 响应时间仅 0.2 秒，还能有效降低算力成本，解决传统 GPU 在推理环节的“内存墙”与高延迟问题。这家明星初创企业曾在 9 月完成一轮 7.5 亿美元的融资，投后估值达 69 亿美元，累计融资超 30 亿美元。

我们认为英伟达与 Groq 的合作是对推理算力的前瞻布局，反映出 AI 算力需求正由以训练为中心向训练+大规模推理并重演进，印证大模型应用落地节奏加快、商业化前景乐观。



图表10: Groq 的芯片专为推理设计



来源: groq 官网, 国金证券研究所

入口不再仅限于手机，而是演变为“OS 级智能体”与“超级 APP”层面较量。1) 超级 APP 的入口之争已经打响。2025 年 12 月 24 日，字节跳动旗下 AI 应用豆包宣布日均活跃用户数（DAU）突破 1 亿；千问 App 近期持续扩大投流，截止 12 月 10 日（公测 23 天）月活已突破 3000 万，下成全球增长最快 AI 应用。阿里巴巴在内部沟通中提及，千问 C 端事业群的首要目标是将千问打造成为一款超级 APP，成为 AI 时代用户的第一入口；2) “入口之战”从流量分发层推到了系统权限层，已成必争之地。豆包绕开传统接口，用“看屏幕+代操作”的方式在现有生态之上再搭一层“AI 操作系统”，AIOS 直接触碰微信、支付宝等超级 App 的商业命门，传统 App 时代的游戏规则面临挑战，微信、淘宝等超级 App 陆续封禁豆包手机权限。但值得注意的是，华为、小米、荣耀、OPPO、vivo 等品牌均在旗舰机型里注入 AI Agent 能力，模型厂商仍有望通过开源或广泛结盟来绕过 APP 厂商的封锁。

在此背景下，我们认为入口之争的背后是流量之争，直接指向大型互联网企业广告及电商业务的基本盘，2026 年各家在模型、AI 应用产品上的算力投入将进入白热化的锦标赛阶段。

图表11: 微信淘宝等超级 APP 封禁豆包手机

豆包手机刚上线，就被微信淘宝“拉黑”了

中国企业家杂志 · 2025年12月04日 15:36

字节为何敢于打破行业默契？

“ 电商比价、浏览微信，字节为何敢于打破行业默契？

字节的硬件野心，从未加以掩饰。但当它踏入了手机和超级App们的领地，甚至希望重新分配移动互联网的控制权时，很快撞上了阻力墙。

12月1日，豆包发布手机助手，在中兴nubia M153系列手机上，获得了操作系统层面的高权限。用户通过唤醒“豆包”，可以完成跨应用比价、购物、撰写报告等任务；也可以打开微信，自动回复消息，玩小程序游戏等。

图表12: 豆包手机扩充 VIVO、联想、传音等合作商

豆包手机扩充合作厂商 vivo、联想、传音在列

2025年12月19日 12:10 手机之家

据界面新闻今日报道，字节跳动正与 vivo、联想、传音等硬件厂商推进 AI 手机合作计划，拟通过预装 AIGC 插件获取用户入口。有 vivo 员工证实，字节跳动已与 vivo 确立合作意向，具体方案仍在商讨中。截至IT之家发稿，相关企业未对合作事宜作出回应。



来源: 新浪财经, 国金证券研究所

GUI Agent AI 手机的出现有望驱动形成繁荣的 AI 终端生态。因为手机作为超级个人计算终端，承载了几乎所有 C 端的信息流，且是物理世界的超级入口，所以只有当 AI 击穿手机之后，其他 AI 终端才有可能依托 AI 手机持续繁荣，并且由于 AI 手机的信息流处理复杂度较高，所以



AI 手机一定是“大云小端”的，进一步拉动算力需求。

图表13：终端 AI 有望依托 AI 手机生态持续繁荣

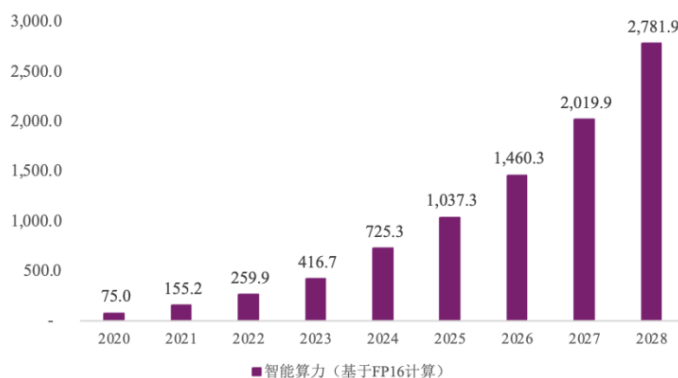


来源：国金证券研究所绘制

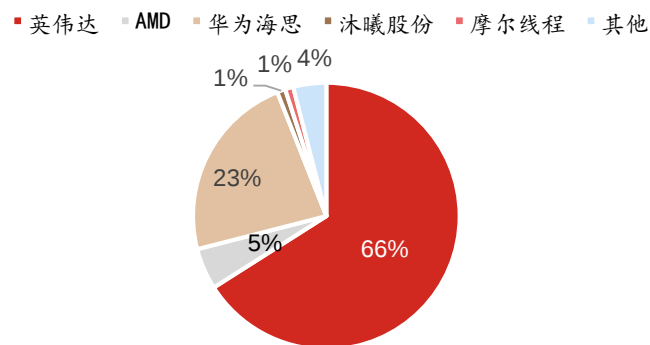
二、国产卡持续突破，2026 年爆发之时

智算中心持续扩容，国产替代加速。根据 IDC 数据，2020 年中国智能算力规模为 75.0 EFLOPS，到 2028 年预计将达到 2,781.9 EFLOPS，预计 2020-2028 年复合增长率达到 57.1%。随着地缘政治紧张局势推动中国企业寻求本地替代方案，以及国产芯片技术的稳步提升，国内云服务提供商正在加速构建异构环境（将不同类型的芯片结合使用，如 CPU、GPU、国内替代芯片）。例如，腾讯云已将其异构计算平台全面兼容所有主流国产芯片，以满足内部开发和客户对 AI 算力的需求。根据 Bernstein Research，2024 年中国 AI 加速芯片市场中，英伟达、AMD 市场份额分别为 66%、5%，合计占比达 71%。但受益于国产替代趋势及供应链安全需求，国内计算芯片公司正迅速提升。其中，华为海思/沐曦/摩尔市场份额已分别达到 23%/1%/1%。

图表14：2020-2028 年中国智能算力规模及预测



图表15：2024 年中国 AI 加速芯片市场竞争格局



来源：沐曦招股说明书，国金证券研究所

来源：沐曦招股说明书，国金证券研究所

国产通用 GPU 从“可用”向“好用”升级。国产 GPU 在性能指标、软件生态、应用适配等方面与 NV 最先进一代仍有差距，但已基本追平 H20、A100 等，且在本地化服务、政策支持、成本控制等方面具备优势。随着资本持续注入，国产企业有望在细分场景实现突破，逐步扩大市场份额。1) 算力指标上：国内多数头部企业主流在售产品的 FP16/BF16 在 100-300 TFLOPS 左右，处于英伟达 A100 产品阶段，少数厂商通过先进封装等方式实现接近英伟达 H100 产品的算力，为国内最先进水平；2) 显存方面：国内企业结合自身产品特点，分别选择 HBM2e、HBM2、GDDR 等显存类型，显存带宽在 0.5-2 TB/s 左右。



图表16：国产通用 GPU 从“可用”向“好用”升级

参数 / 厂商	平头哥	NV		华为	壁仞
型号	PPU	A800	H20	昇腾 910B	104P
显存容量	96G	80G	96G	64G	32G
显存类型	HBM2e	HBM2e	HBM3	HBM2	HBM2e
片间带宽 (GB/s)	700	400	900	392	256
PCIe	5.0 × 16	4.0 × 16	5.0 × 16	4.0 × 16	5.0 × 16
功耗 (W)	400	400	550*	350	300

来源：芯东西微信公众平台，国金证券研究所

供给侧：中芯国际作为中国集成电路领导者，产能/利用率持续提升。11 月 13 日，中芯国际披露 2025 年第三季度财报，2025Q3 中芯国际营收 171.62 亿元，环比增长 6.9%，毛利率 25.5%；月产能为 102.28 万片(折合 8 英寸)，同比增加产能约 13.85 万片(折合 8 英寸)；产能利用率达 95.8%，环比增长 3.3 个百分点。中芯国际产能全球第三，约为台积电的三分之一。公司资本支出维持高位，未来产能或将继续增加。

图表17：中芯国际产能/利用率持续提升

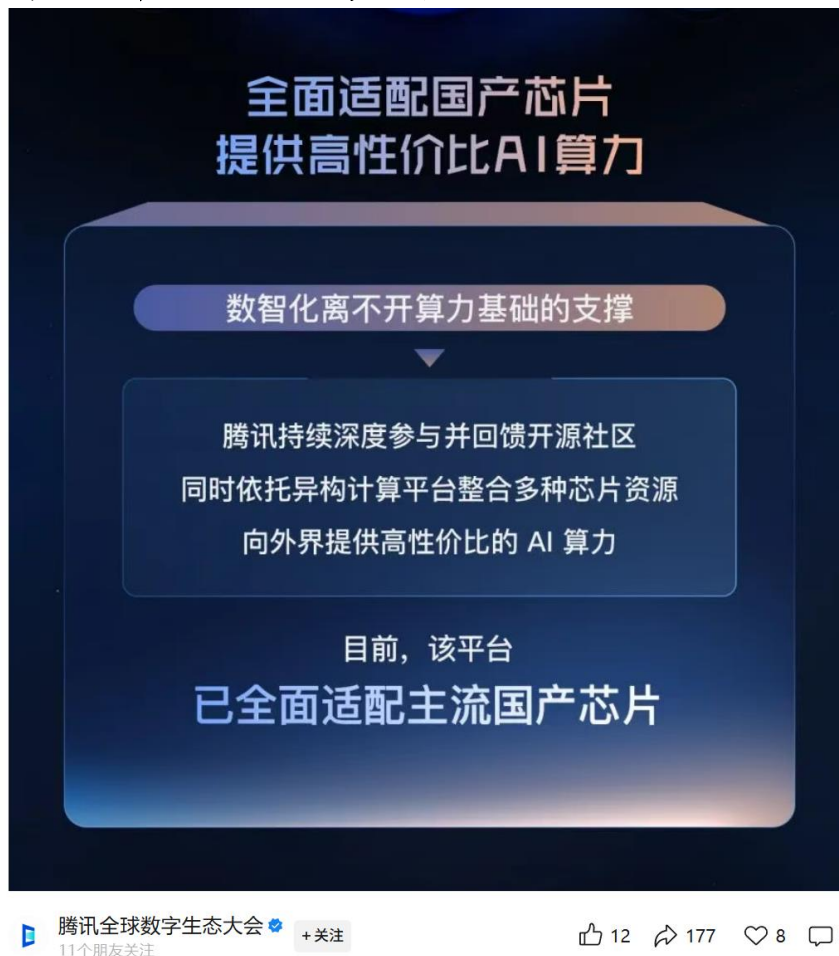


来源：TrendForce，国金证券研究所

CSP 厂商加速适配，助力国产芯片生态建设。英特尔的 X86 生态、英伟达的 CUDA 生态之所以难以撼动，核心在于形成了“芯片-软件-应用”的闭环。而当前国产阵营中，华为昇腾、阿里平头哥、壁仞科技等芯片厂商各有技术路线，生态分散问题显著。腾讯集团高级执行副总裁汤道生在交流中坦言，不同参数规模的 AI 模型需要适配不同芯片配置，当前只能通过与多家厂商合作实现场景覆盖。百度、阿里等企业加速适配国产芯片，推动“芯片-模型-应用”闭环形成。



图表18：腾讯云宣布全面适配主流国产芯片



来源：腾讯全球数字生态大会，国金证券研究所

三、供需双高，国内算力斜率陡峭

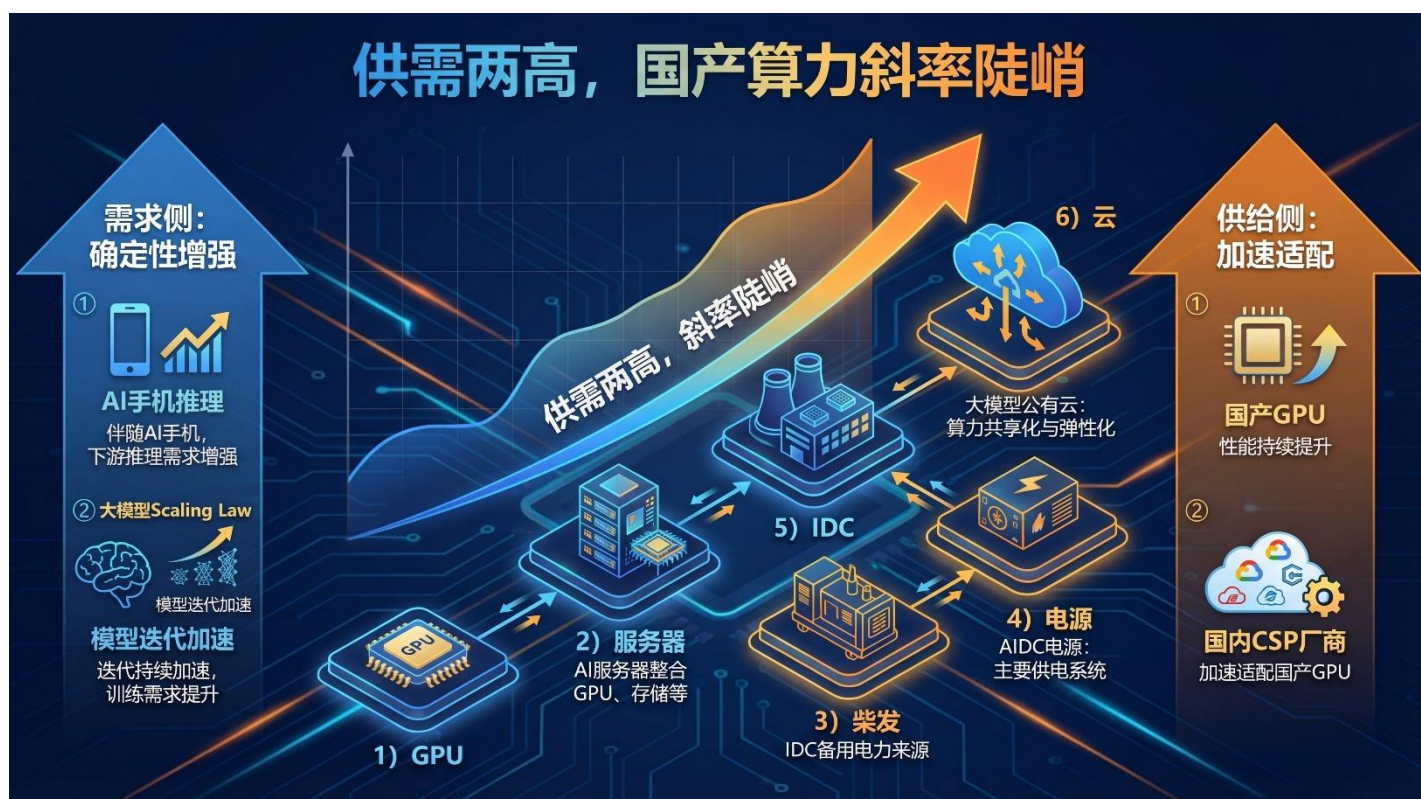
供需双高，国内算力斜率陡峭。

- 1) 需求侧：①下游推理需求伴随 AI 手机等终端 AI 进入落地放量临界点确定性增强；②大模型 Scaling Law 仍然有效，模型迭代持续加速提升训练需求的确定性。
- 2) 供给侧：①国产 GPU 性能持续提升；②国内 CSP 厂商加速适配国产 GPU。

上述背景下，GPU、服务器、柴发、电源、IDC、云 6 大产业环节供需两高，其中，AI 服务器整合 GPU、存储等直接影响模型效果的硬件基础；AIDC 承载算力与数据；AIDC 电源是 AIDC 的主要供电系统；柴油发电机是 AIDC 不可或缺的备用电力来源；大模型公有云本质上是算力的共享化与弹性化。



图表19：供需两高，国产算力斜率陡峭

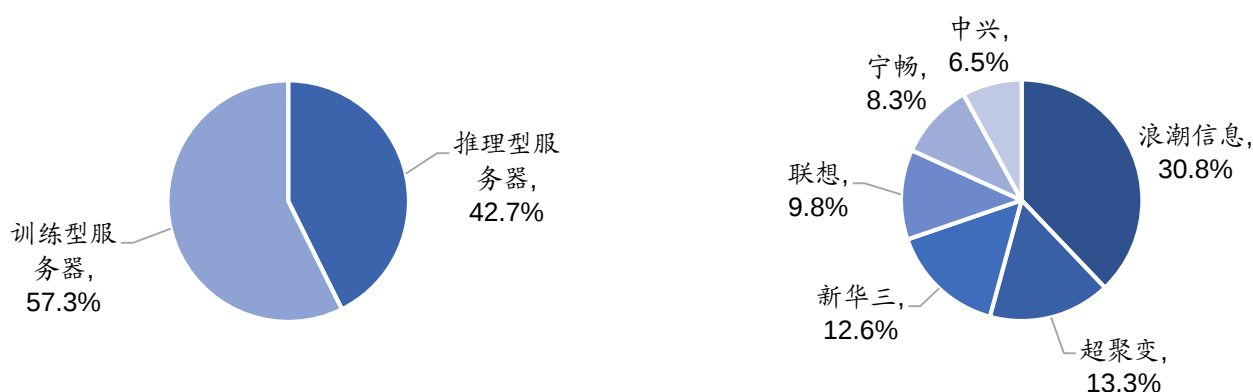


来源：国金证券研究所绘制

AI服务器是指专为AI应用设计的高性能计算机设备，能够支持大规模数据处理、模型训练、推理计算等复杂任务。与传统服务器相比，AI服务器通常配备强大的处理器（如CPU、GPU、FPGA、ASIC等）、高速内存、大容量且高速的存储系统以及高效的散热系统，以满足人工智能算法对计算资源的高需求。据IDC，目前AI服务器市场中57.33%为训练型服务器，推理型服务器占比达42.67%，预计未来推理型服务器将逐渐成为市场主流。2024年，我国x86服务器市场CR6分别为浪潮信息/超聚变/新华三/联想/宁畅/中兴，2024年我国x86服务器市场规模约为393亿美元，同比增长49.7%。

图表20：预计未来推理型服务器将逐渐成为市场主流

图表21：2024年我国x86服务器市场CR6



来源：中投未来产业研究中心转引自IDC，国金证券研究所

来源：芯语转引自IDC，国金证券研究所

AIDC电源是专为人工智能数据中心设计的供电系统。其核心功能是将电网交流电高效、稳定地转换为服务器、GPU等高性能计算设备所需的直流电压，确保供电环节的高可靠性、高效率和高功率密度。国际巨头如台达电子占据AIDC电源市场主导地位，施耐德等在高端市场具有优势。国内企业如麦格米特、欧陆通、中恒电气等凭借技术突破和成本优势，市场份额逐步提升，其中麦格米特成为英伟达指定供应商，中恒电气是阿里核心HVDC供应商。据华经产业研究院，2024年全球AIDC电源市场规模已达约49亿美元。



图表22: AIDC 电源是专为人工智能数据中心设计的供电系统

	主要 AIDC 电源分类
UPS 不间断电源	当市电输入正常时, UPS 将市电稳压后供应给负载使用, 同时向机内电池充电; 当市电意外中断时, UPS 立即将电池的直流电能, 通过逆变器切换转换的方法向负载继续供应电能, 使负载维持正常工作并保护负载软、硬件不受电网波动而造成损坏。
AC-DC 电源模块	将电网的交流电转换为适合服务器使用的直流电, 如 48V 或 50V 直流电。其内部包含电源供应单元(PSU)和电源管理控制器(PMC), PSU 为 CPU、GPU 等组件提供电源, PMC 则用于管理和优化服务器电源。
DC-DC 电源模块	进一步将较高电压的直流电降至芯片可接受的低电压, 如将 48V 或 50V 直流电降至 0.8V 等, 以满足服务器内部不同组件的供电需求。

来源: 观研天下, 国金证券研究所

图表23: AIDC 电源主要玩家一览

	主营业务	AIDC 电源产业布局
麦格米特	工业电源、新能源汽车零部件	800V 直流输出产品, 是英伟达指定车零部件等供应商, 参与 Blackwell 架构设计
欧陆通	开关电源的研发、生产和销售	浸没式液冷电源, 为阿里定制液冷供电系统, 合作浪潮、新华三等
中恒电气	电力电子设备及系统集成	800V HVDC 研发, 巴拿马电源系统, 是阿里核心 HVDC 供应商, 覆盖 BAT 数据中心
雄韬股份	化学电源、新能源储能、动力电池等	是阿里 HVDC 锂电备电唯一供应商, 主导 BBU 市场

来源: 华经产业研究院, 国金证券研究所

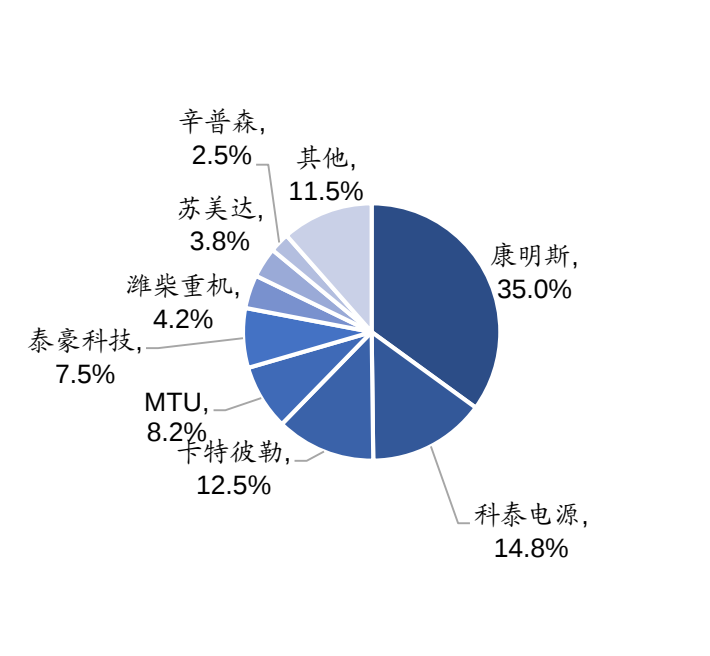
柴油发电机是数据中心不可或缺的备用电力来源。AIDC 备用电源可分为超级电容(瞬间响应, 应对短时大功率冲击)、BBU(快速响应, 提供机柜级持续供电)、UPS(迅速介入, 保障设备持续运行)和柴油发电机(较慢响应, 应对长时间停电)四类, 实际应用中通常采用组合式供电方案协同运作。柴油发电机组约占数据中心基建成本比重的 23%。由于从柴油发动机集成至柴发机组技术壁垒较低, 国内厂商与外资柴油发动机龙头合作, 具备相关制造能力。2024 年中国数据中心用柴发机组中, 国产品牌科泰电源、泰豪科技、潍柴重机、苏美达市占率合计约为 30%。据华经产业研究院, 2024 年我国数据中心用 2MW 柴发机组市场规模或达 82 亿元。

图表24: 柴油发电机组在数据中心作为备用电源使用

	原理	优势	劣势
电网主力电源	直接将市电引入数据中心, 为数据中心的各类设备供电。	系统简单, 建设成本低, 维护方便。	需要配备其他备用电源来提高供电的可靠性。
不间断电源(UPS)短时间备用电源	市电中断时, 继续为设备供电, 确保设备不会因市电中断而停机。	为数据中心设备提供高质量的电力保障。可在毫秒级时间内切换到电池供电。	UPS 电池组的容量有限, 且价格较高。
柴油发电机长时间备用电源	当市电中断且 UPS 电池组电量即将耗尽时, 柴油发电机启动, 为数据中心设备供电。	功率较大, 能满足数据中心在市电停电时的大部分甚至全部电力需求。可长时间持续供电。	启动时间相对较长, 一般需要几十秒到几分钟才能到达稳定供电状态。

来源: 观研天下, 国金证券研究所

图表25: 2024 年国产 IDC 柴发机组份额约为 30%



来源: 华经产业研究院, 国金证券研究所

AIDC 的参与主体众多, 涵盖政府、CSP、三方 IDC 服务商、上游服务器或芯片渠道商等等, 不同主体的 AIDC 的功能定位、盈利模式以及布局诉求均存在较大差异。据《中国智算中心产业发展白皮书》, 2024 年互联网及云厂商/基础电信运营商/地方政府分别占我国 AIDC 算力规模份额的 35.0%/25.6%/14.2%。AIDC 商业模式主要包括 IaaS(包括传统的机房托管与算力租赁两种细分模式)、PaaS(提供 AI 应用开发平台及工具)、MaaS(模型定制、微调、部署等全流程服务)、SaaS(直接落地应用)4 类, 目前算力租赁为 AIDC



主要商业模式，主要玩家包括上游芯片厂商、下游 AI 企业、地方政府上市公司新进入者、云厂商五大力量。

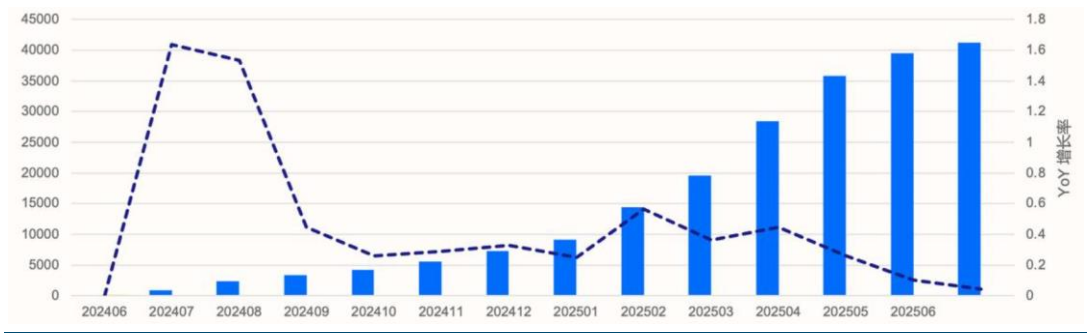
图表26：不同主体的 AIDC 的功能定位与布局诉求存在差异

主体类别	战略目标	优势	劣势	代表案例
地方政府	响应国家战略，推动政策落地；促进区域产业发展	政策、资金、产业园区客户	缺乏 IT 技术、供应链经验、运营能力不足	北京/上海/广州/武汉/天津/重庆/长沙/南京/西安/成都/合肥等 30 多个城市
互联网及云厂商、基础电信运营商	满足自身大模型训练需求；拓展算力业务	充足的资金、技术和客户资源，丰富的软件供应链资源，较成熟的云计算业务模式	自有大模型与大模型训练客户形成竞争互斥	腾讯合肥智算中心/百度智算中心/阿里飞天云智能华东算力中心等；中国电信武清智算中心/中国联通芜湖智算中心/中国移动武汉智算中心等
第三方 IDC 服务商	寻求业务增长	丰富的 IDC 机房资源，IDC 一体化建设运营能力	缺乏 IT 供应链资源和算力客户	润泽国际信息港 A-11 云数据中心、博大数据深圳前海智算中心
服务器厂商/芯片渠道商	纵向一体化	AI 芯片、服务器等硬件资源获取能力	缺乏 IDC 供应链和机房建设运营能力	协鑫智算（上海）中心、浪潮新疆克拉玛依智算中心
AI 企业、应用企业	纵向一体化	算法及相关软件能力，客户资源和应用场景积累	缺乏 AI 芯片资源、IDC 资源	商汤临港 AIDC、理想汽车智算中心、小鹏“扶摇”智算中心
跨界企业	战略转型/发展新业务	一般与芯片厂商、渠道商合作，获得芯片资源	缺乏 IDC 资源和持续稳定的客户	英博数科北京 AI 创新赋能中心、威星智能贵安智算中心、恒润股份芜湖智算中心

来源：《中国智算中心产业发展白皮书》，国金证券研究所

大模型公有云本质上是算力的共享化与弹性化，其核心是实现了算力发展的规模效应。据 IDC，2024 年我国公有云大模型调用量达 114 万亿 Tokens；2025 年上半年我国公有云上大模型调用量达 536.7 万亿 Tokens，实现同比近 400% 的增长，其中，火山引擎/阿里云/百度智能云分别取得 49.2%/27.0%/17.0% 的市场份额。

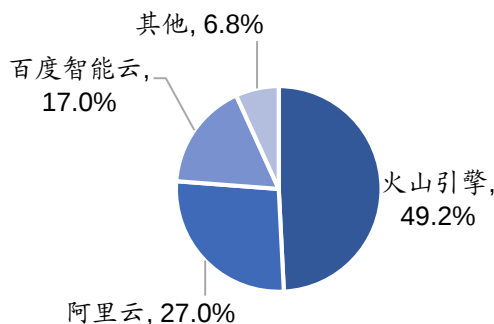
图表27：我国公有云大模型调用量快速增长



来源：量子位转引自 IDC，国金证券研究所



图表28：2025 上半年我国大模型公有云服务市场格局



来源：量子位转引自 IDC，国金证券研究所

四、相关标的

国内算力：寒武纪、东阳光、海光信息、协创数据、星环科技、神州数码、百度集团、大位科技、润建股份、华丰科技、中芯国际、华虹半导体、兆易创新、大普微、中微公司、兴森科技、中科曙光、禾盛新材、润泽科技、浪潮信息、东山精密、亿田智能、奥飞数据、云赛智联、瑞晟智能、科华数据、潍柴重机、金山云、欧陆通、杰创智能。

海外算力/存储：中际旭创、新易盛、天孚通信、源杰科技、胜宏科技、景旺电子、英维克等；闪迪、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

风险提示

■ 行业竞争加剧的风险：

在信创等政策持续加码支持计算机行业发展的背景下，众多新兴玩家参与到市场竞争之中，若市场竞争进一步加剧，竞争优势偏弱的企业或面临出清，某些中低端品类的毛利率或受到一定程度影响。

■ 技术研发进度不及预期的风险：

计算机行业技术开发需投入大量资源，如果相关厂商新品研发进程不及预期，表面层面将呈现出投入产出在较长时期的滞后特征。

■ 特定行业下游资本开支周期性波动的风险：

部分计算机公司系顺周期行业，下游资本开支波动与行业周期性相关性较强，或在个别年份对于上游软件厂商的营收表现产生扰动。



行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。



特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海	北京	深圳
电话：021-80234211	电话：010-85950438	电话：0755-86695353
邮箱：researchsh@gjzq.com.cn	邮箱：researchbj@gjzq.com.cn	邮箱：researchsz@gjzq.com.cn
邮编：201204	邮编：100005	邮编：518000
地址：上海浦东新区芳甸路 1088 号 紫竹国际大厦 5 楼	地址：北京市东城区建国门内大街 26 号 新闻大厦 8 层南侧	地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806



【小程序】
国金证券研究服务



【公众号】
国金证券研究