2026 年 01 月 17 日

计算机行业研究

买入（维持评级）

行业深度研究

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：李可夫（执业 S1130525120009） likefu@gjzq.com.cn
联系人：孙恺祈 sunkaiqi@gjzq.com.cn

算力依旧

本周观点

- **AI 入口之争，大厂投资力度再增强。**

  1）大模型迭代不断，中国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。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： 1 月 15 日千问 App 宣布全面接入淘宝、支付宝、淘宝闪购、飞猪、高德等阿里生态业务，实现点外卖、买东西、订机票等 AI 购物功能，引领 AI 行业从“聊天对话”迈入“办事时代”，1 月 16 日 OpenAI 宣布推出月费 8 美元的“ChatGPT Go”订阅套餐，并计划未来几周在美国的免费用户和 Go 用户中测试广告。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。③政策驱动需求爆发：工信部一月规定限制新建智算中心的 PUE，推动算力硬件向高效节能方向升级，“十五五”规划首次将“全国一体化算力网”纳入国家级基础设施体系。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。③海外云端算力管制加码，1 月 12 日美国众议院高票通过《远程访问安全法》，预计迫使国内 AI 厂商加速国产化替代进程。

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 3
二、国产卡持续突破，2026 年爆发之时 8
三、供需双高，国内算力斜率陡峭 11
四、相关标的 13
风险提示 13

图表目录

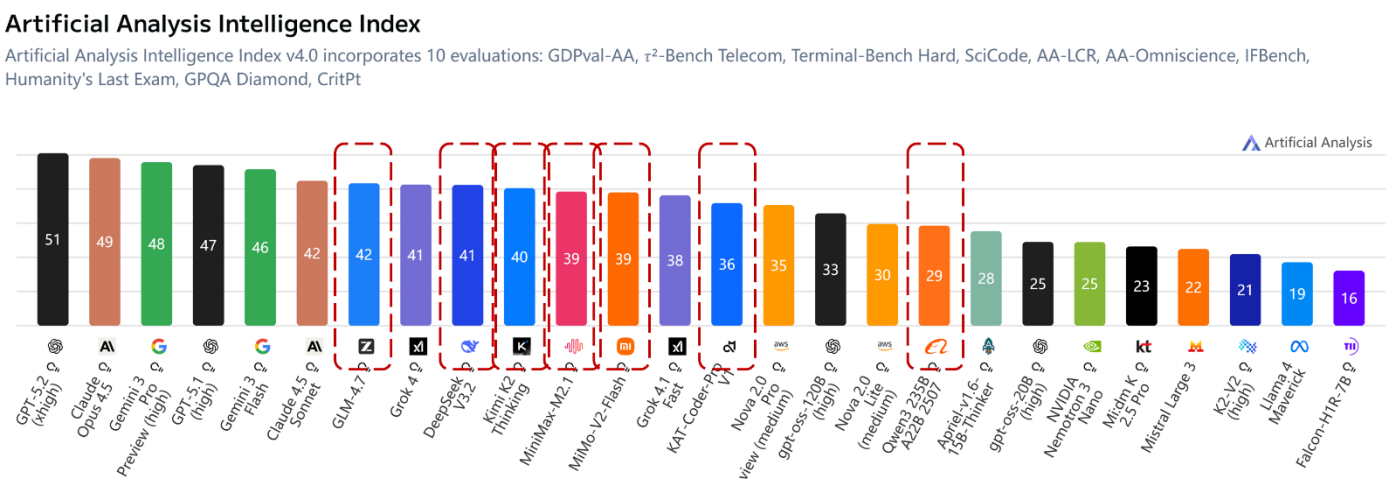
图表 1：中国 AI 在全球顶级模型排名中位居中段 3
图表 2：北美 AI 玩家持续引领；中国 AI 玩家加速追赶；欧洲创企表现不俗 3
图表 3：中国开源 AI 模型调用量约占全球市场的 70+%，DS 一家约占 40+% 4
图表 4：Hugging Face 上主要开源权重 LLMs 的累积下载量 4
图表 5：ChatBot Arena 中得分最高的开源权重模型,截止 12 月 4 日 5
图表 6：三大 Scaling 定律 5
图表 7：Interleaved Thinking 推理范式 6
图表 8：Hopper (H100)与 Blackwell(B200/GB300) 对比 7
图表 9：推理相对非推理解锁了 10X Token 消耗 7
图表 10：千问 APP 接入阿里生态业务 8
图表 11：2020-2028 年中国智能算力规模及预测 9
图表 12：2024 年中国 AI 加速芯片市场竞争格局 9
图表 13：国产通用 GPU 从“可用”向“好用”升级 9
图表 14：中芯国际产能/利用率持续提升 10
图表 15：腾讯云宣布全面适配主流国产芯片 10
图表 16：供需两高，国产算力斜率陡峭 11
图表 17：预计未来推理型服务器将逐渐成为市场主流 12
图表 18：2024 年我国 x86 服务器市场 CR6 12
图表 19：不同主体的 AIDC 的功能定位与布局诉求存在差异 12
图表 20：我国公有云大模型调用量快速增长 13

一、AI 入口之争，大厂投资力度再增强

模型迭代持续精彩，中国大模型全球存在感大幅提升。

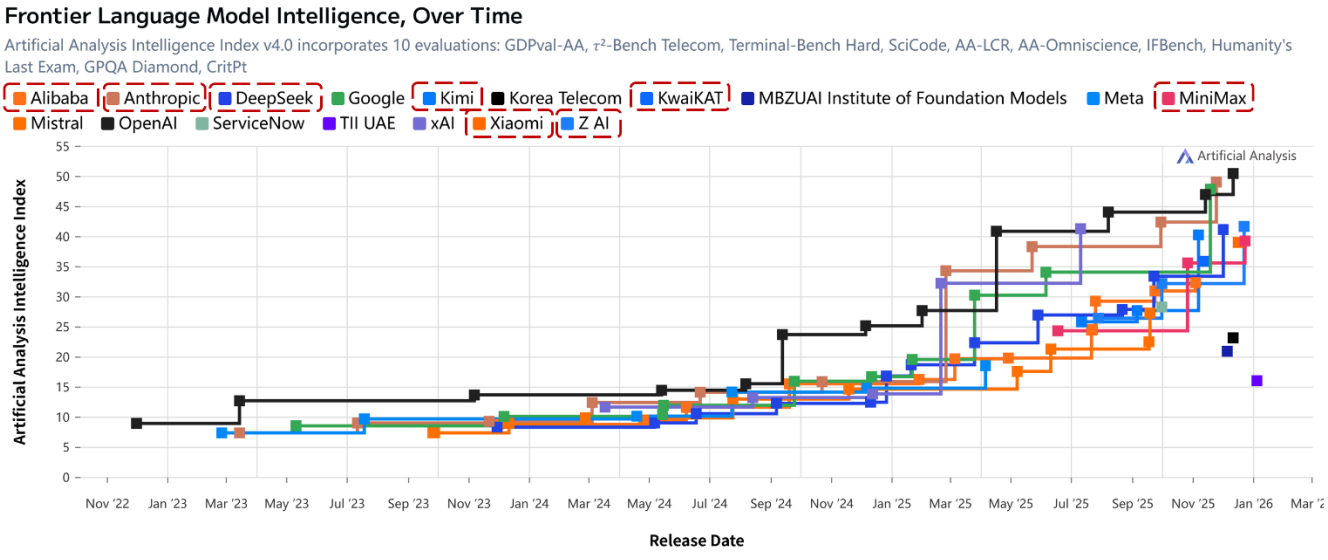
中国 AI 在全球顶级模型排名中位居中端。据 AA，尽管 GPT-5.2 与 Gemini 3 Pro 仍占据塔尖，但中国国产模型已有效改变了北美单极主导的竞争格局。在全球 Top 10 阵营中，GLM-4.7、DeepSeek V3.2、 Kimi K2 Thinking 已占据 3 席；若将观测范围扩大至 Top 15，中国企业席位达 6 个。

图表1：中国 AI 在全球顶级模型排名中位居中段



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, τ²-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

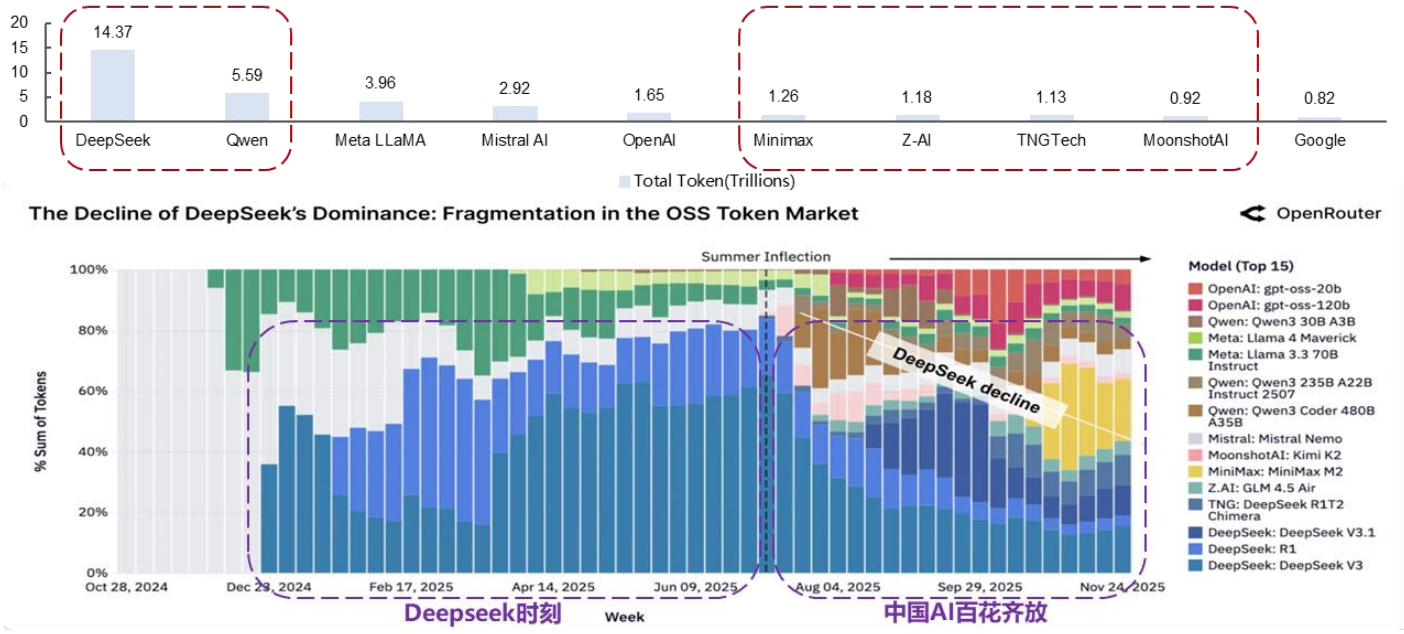


来源：Artificial Analysis 官网，国金证券研究所（注：①数据及信息截至 2026 年 1 月 9 日；红框内为中国公司；② Intelligence Index 综合考虑了 MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ²-Bench Telecom 十类测评基准。）

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+%

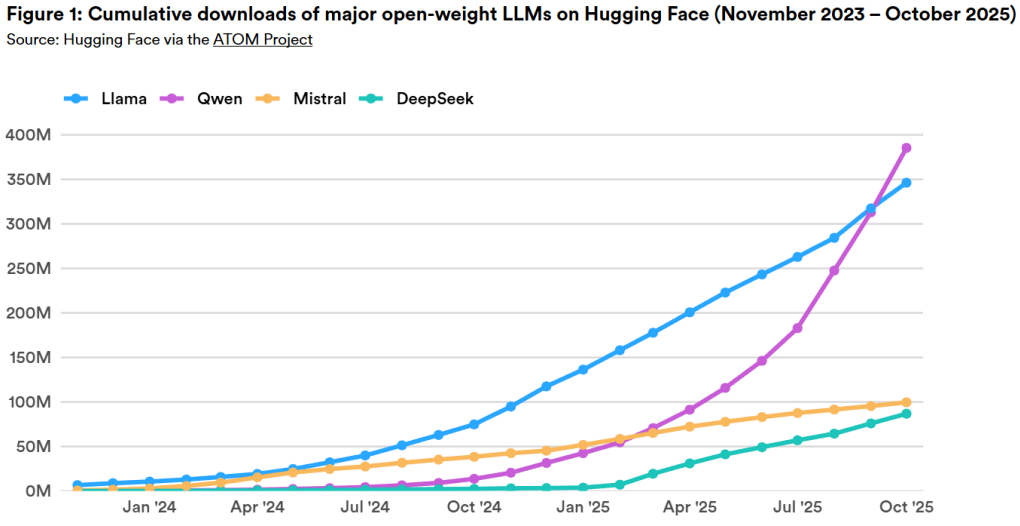


来源：《State of AI: An Empirical 100 Trillion Token Study with OpenRouter》，国金证券研究所

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



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

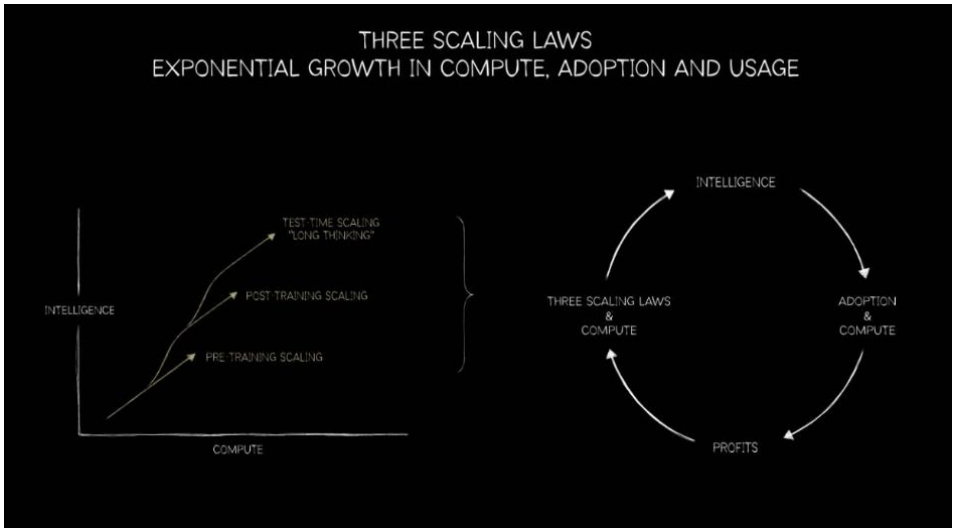
Rank	Model	Lab	Country	Elo score*	Coding score	Vision score	AAII score†	MMLU-Pro score§
1	GLM-4.6	Z.ai	China	**1442**	1458		56	83.5
1	Kimi-K2-Thinking	Moonshot	China	**1438**	1450		66	84.8
2	GLM-4.5	Z.ai	China	**1430**	1448		54	83.5
2	Mistral Large 3	Mistral	France	**1429**	1450		40	81
2	Qwen3-VL-235B-A22B-Instruct	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-Think-ing-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-In-struct	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。

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

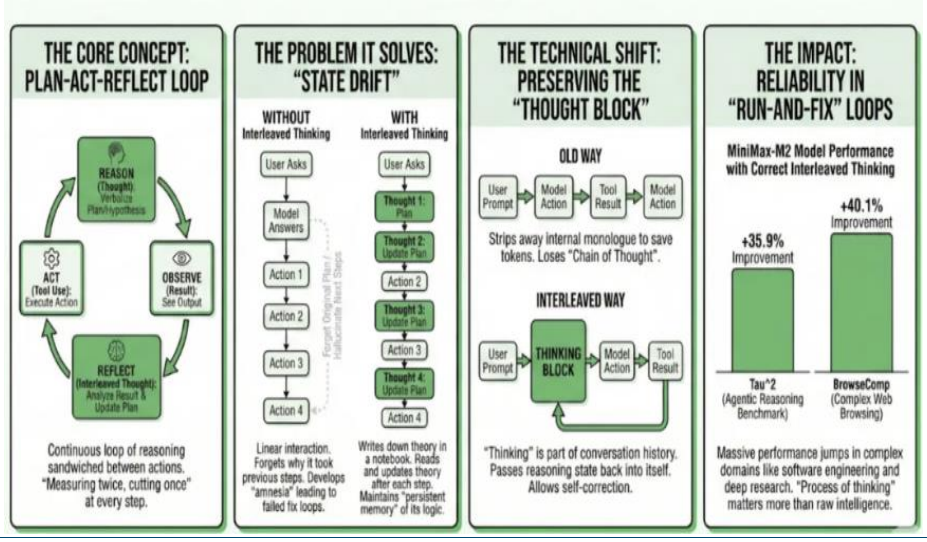


来源：英伟达官网，国金证券研究所

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

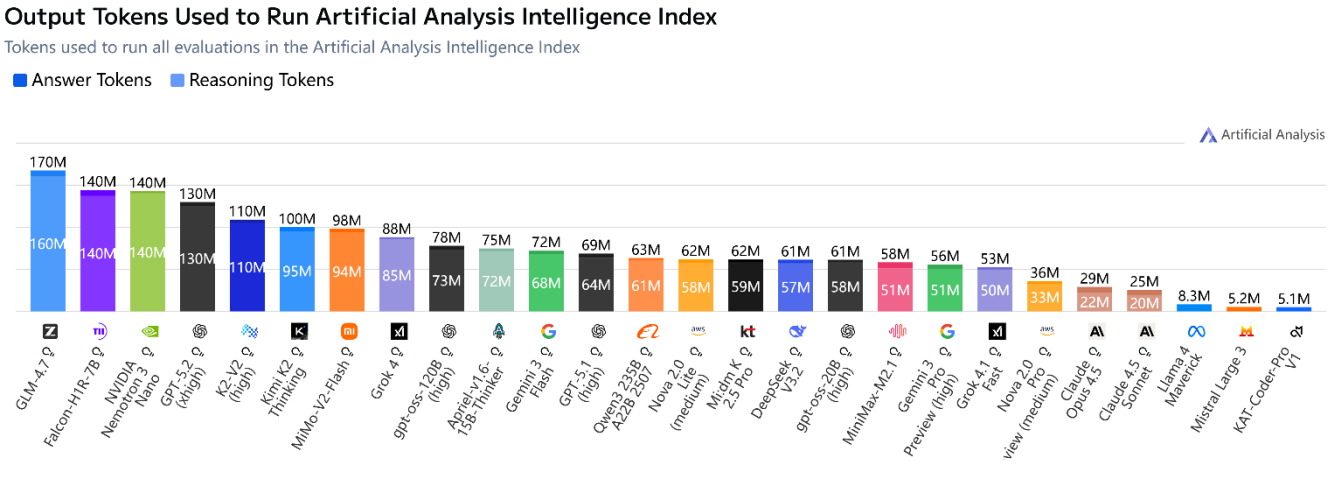
特征	NVIDIA HOPPER (H100)	NVIDIA BLACKWELL（B200/GB300）
建筑概论	Hopper（H100 Tensor Core，2022）	Blackwell（B200/GB300 Tensor Core，2024/25）
每个GPU的SM数量	约80个SM（GH100）	每个芯片约 80 个 SM（B200），2 个芯片共封装 [119]
晶体管（总数）	~80B（整体式）[120]	~208B（2×芯片组）[119]
进程节点	台积电 4N（定制 5nm）	台积电 4NP（新型 4nm 变体）[119]
芯片包装	单GPU芯片	双GPU芯片组，带10 TB/s NVLink芯片间链路 [119]
每个SM的共享/L1内存	256 KB（可配置）	128 KB（更小，统一）[14]
统一的 L2 缓存	总计 50 MB（分块）	65 MB 单体式 [14]
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 [17]	192 GB HBM3e (B200) / 288 GB (GB300) [2] [17]
HBM带宽	3.35 TB/s [18]	8.0 TB/s [18]
NVLink（芯片间）	NVLink4 - 900 GB/s	NVLink5 - 1.8 TB/s [21]
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 [21]
张量核心精度（密集）	FP16/BF16、FP8、INT8	FP16/BF16、FP8、*FP6*、*FP4* [10] [13]
FP8/密集型 TFLOPS	峰值约为 4–5 PF	~~9 PF [1] （每个 GPU）
FP4/密集型 TFLOPS	（不支持）	~~18 PF [1]（每个 GPU，新模式）
FP64 TFLOPS（张量）	60 TF（张量）[12]	较低（弱化）[1]
峰值功率（TGP）	约700瓦 [17]	约 1,200–1,400 瓦 [22]

来源：intuitionlabs.ai，国金证券研究所

推理需求高增：推理模型解锁 10X 算力需求。

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，会先在内部生成长思维链，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



来源：Artificial Analysis 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。

超级入口生态之战开启。1 月 15 日阿里召开产品发布会，千问 App 宣布全面接入淘宝、支付宝、淘宝闪购、飞猪、高德等阿里生态业务，实现点外卖、买东西、订机票等 AI 购物功能，并向所有用户开放测试。，此次升级将上线超 400 项 AI 办事功能，让千问 App 成为全球首个能完成真实生活复杂任务的 AI 助手，引领 AI 行业从“聊天对话”迈入“办事时代”。

图表10：千问 APP 接入阿里生态业务



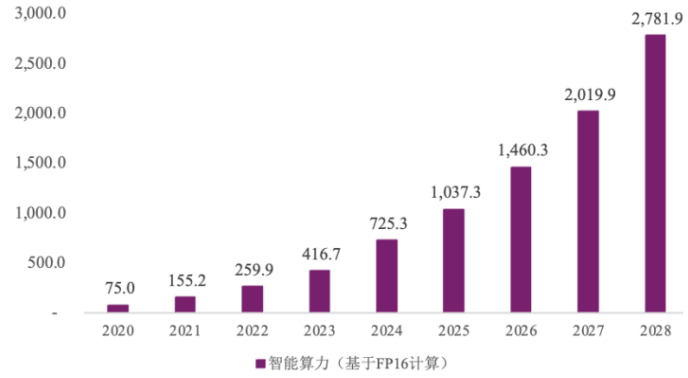
来源：南方日报，国金证券研究所

ChatGPT 将加入广告，大模型商业化进展加速。当地时间 1 月 16 日，OpenAI 宣布，将把月费 8 美元的“ChatGPT Go”订阅套餐推广到 ChatGPT 服务覆盖的所有地区。该套餐最早在印度上线，目前扩展至美国市场。与 ChatGPT Go 相比，Plus 的订阅费用要 20 美元/月；相应地，OpenAI 计划未来几周在美国的免费用户和 Go 用户中测试广告，Plus、Pro 等付费较高的套餐将保持无广告体验。

二、国产卡持续突破，2026 年爆发之时

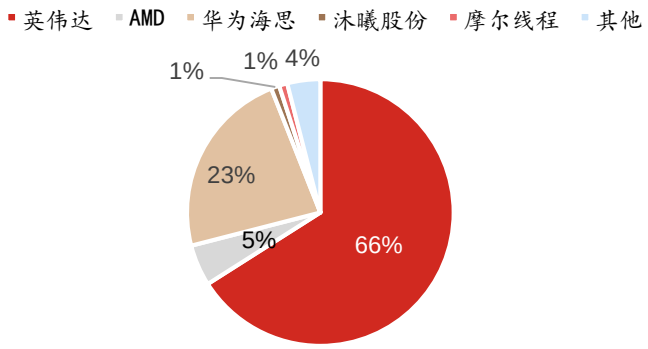
智算中心持续扩容，国产替代加速。根据 IDC 数据，2020 年中国智能算力规模为 75.0 EFLOPS，到 2028 年预计将达到 2,781.9 EFLOPS，预计 2020-2028 年复合增长率达到 57.1%。随着地缘政治紧张局势推动中国企业寻求本地替代方案，以及国产芯片技术的稳步提升，国内云服务提供商正在加速构建异构环境（将不同类型的芯片结合使用，如 CPU、GPU、国内替代芯片）。例如，腾讯云已将其异构计算平台全面兼容所有主流国产芯片，以满足内部开发和客户对 AI 算力的需求 。根据 Bernstein Research，2024 年中国 AI 加速芯片市场中，英伟达、AMD 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11：2020-2028 年中国智能算力规模及预测



来源：沐曦招股说明书，国金证券研究所

图表12：2024 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

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13：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。

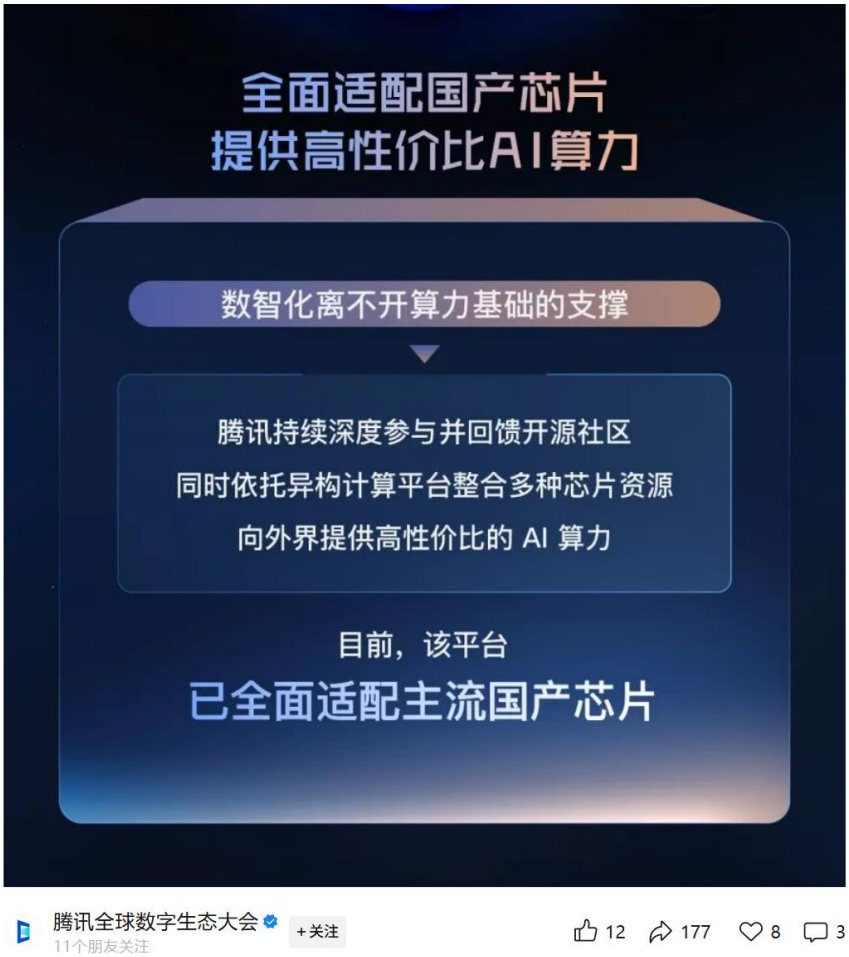
图表14：中芯国际产能/利用率持续提升



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15：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

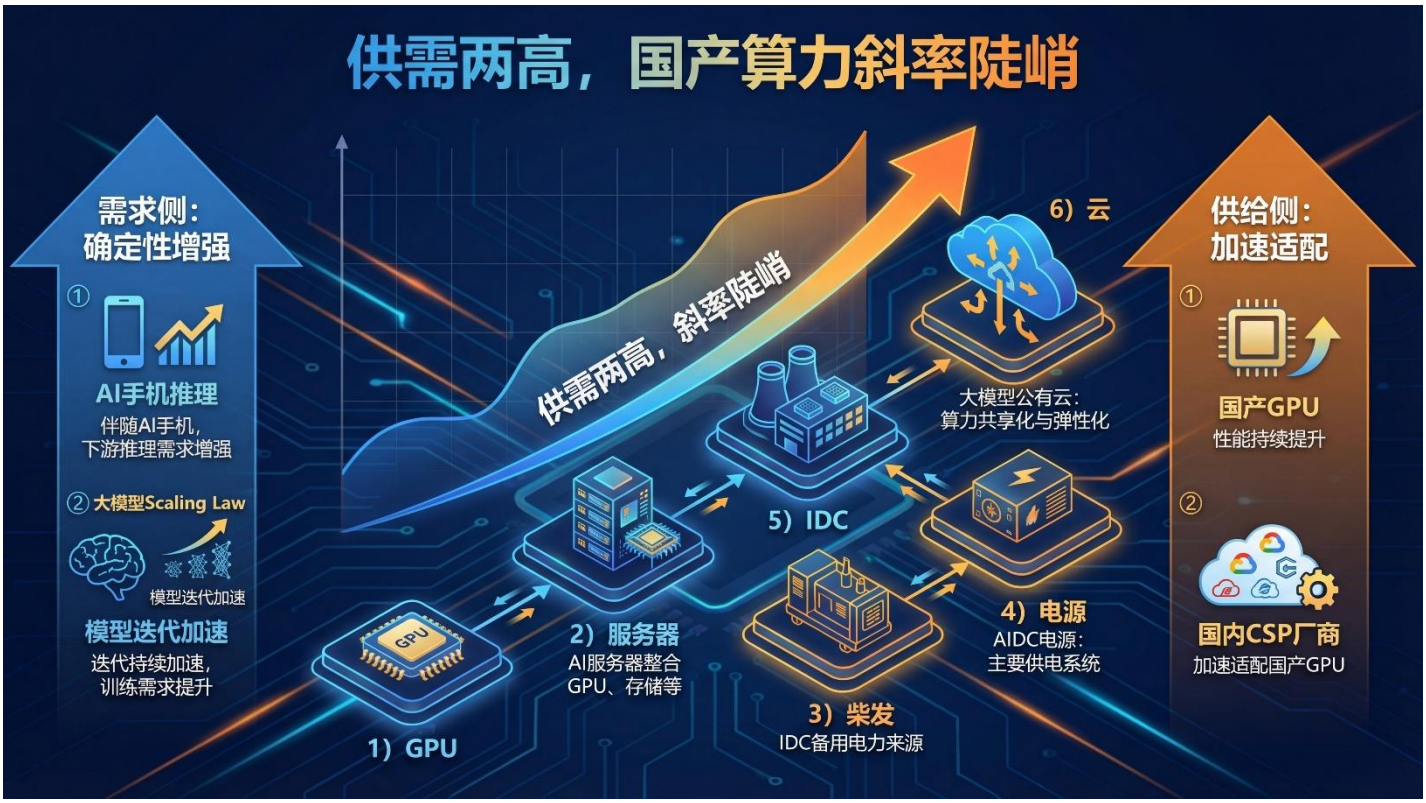
供需双高，国内算力斜率陡峭。

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。

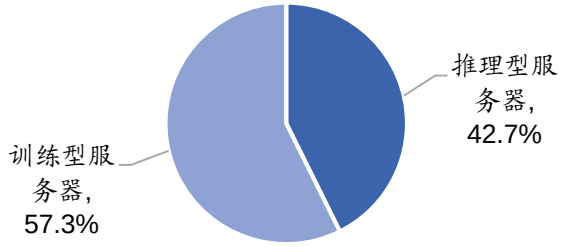
图表16：供需两高，国产算力斜率陡峭



来源：国金证券研究所绘制

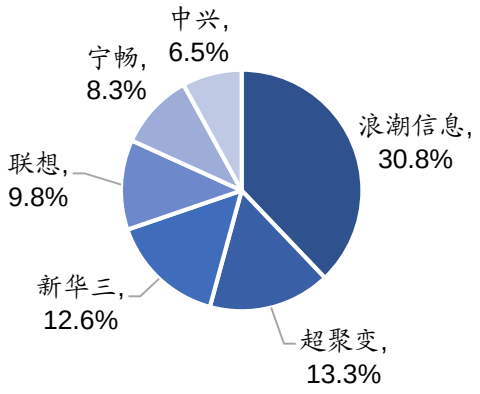
AI 服务器是指专为 AI 应用设计的高性能计算机设备，能够支持大规模数据处理、模型训练、推理计算等复杂任务。与传统服务器相比，AI 服务器通常配备强大的处理器（如 CPU、GPU、FPGA、ASIC 等）、高速内存、大容量且高速的存储系统以及高效的散热系统，以满足人工智能算法对计算资源的高需求。据 IDC，目前 AI 服务器市场中 57.33%为训练型服务器，推理型服务器占比达 42.67%，预计未来推理型服务器将逐渐成为市场主流。2024 年，我国 x86 服务器市场 CR6 分别为浪潮信息/超聚变/新华三/联想/宁畅/中兴，2024 年我国 x86 服务器市场规模约为 393 亿美元，同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

图表17：预计未来推理型服务器将逐渐成为市场主流



来源：中投未来产业研究中心转引自 IDC，国金证券研究所

图表18：2024 年我国 x86 服务器市场 CR6



来源：芯语转引自 IDC，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。

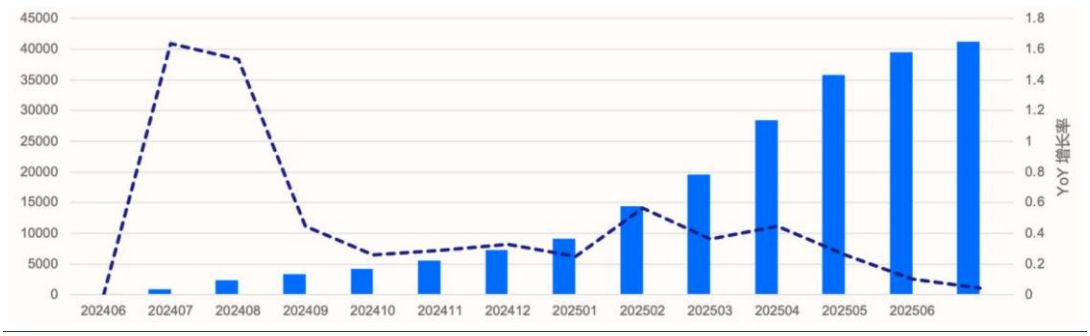
图表19：不同主体的 AIDC 的功能定位与布局诉求存在差异

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。

图表20：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

据金融界报道，2026 年 1 月算力硬件行业迎来多项权威政策与产业利好，为行业发展注入强劲动力：其一，政策层面强化支撑，工信部于 1 月发布《数据中心绿色低碳发展行动计划》，明确规定新建智算中心的 PUE 不得高于 1.2，超算中心的 PUE 不得高于 1.15，而液冷技术作为满足该标准的核心方案，直接推动算力硬件向高效节能方向升级，刺激相关硬件需求爆发。

其二，国家级基建布局落地，“十五五”规划首次将“全国一体化算力网”纳入国家级基础设施体系，国家数据局布局 7 个区域功能节点、25 个城市业务节点，年底将扩容至 50 个，覆盖 80%省区市，构建起全国算力网络骨架，直接拉动 AI 服务器、算力集群等硬件采购需求。

其三，产业需求持续高增，据行业数据显示，2026 年 1 月全球 AI 大模型落地速度较 2025 年翻倍，国内 AI 服务器市场需求同比增长 50%，当月新增 10 个智算中心，算力缺口仍达 30%，叠加液冷渗透率快速提升至 35%，算力硬件供需格局持续优化。

海外云端算力管制加码：据财新报道，当地时间 1 月 12 日美国众议院高票通过《远程访问安全法》（Remote Access Security Act）草案，计划更新美国的出口管制规则，限制通过云来远程访问人工智能（AI）芯片等受出口管制的物项。具体而言，该草案将受管制物项的违规使用范围扩展至一种新情形，即外国人士通过互联网或云计算服务，在受控物项实际所在地及其境外对物项进行远程访问。相比之下，现行美国出口管制法主要侧重于实物出口与转运的监管，以及软件销售和售后运维服务的提供。我们认为，云端算力限制加码将迫使国内 AI 厂商更坚决地回归本土算力底座，加速国产化替代进程。

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，表观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
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
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