2026 年 03 月 28 日

计算机行业研究

买入（维持评级）

行业专题研究报告

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
分析师：鲍淑娴（执业 S1130526020002） baoshuxian@gjzq.com.cn

国内算力部分进入业绩临界点

本周观点

- **国产 Token 量破 140 万亿，国内算力厂商进入业绩临界点，CPU 涨价潮又起。**1）2026 年 3 月国内日均 token 调用量突破 140 万亿，两年增长超千倍；国产 AI 登上世界舞台，3 月 23-27 日全球 AI 大模型总调用量前十阵营中，中国 AI 大模型占据六席，国内算力需求持续扩圈；在此背景下，国产算力、算力商已经进入业绩临界点。2）CPU 涨价潮又起，25 年 10 月，英特尔率先上调前一代 PC 端处理器价格；26 年 1 月，英特尔与 AMD 年度服务器 CPU 库存提前告罄，并酝酿约 15%的提价计划；26 年 3 月 25 日报道，英特尔与 AMD 已正式通知客户全面上调全线 CPU 价格（平均涨幅 10-15%），且交货周期拉长至 8-12 周。国产 CPU 核心性能实现实质性跨越（如阿里玄铁 C950 刷新纪录、海光核心指标对标国际一线），有望承接由 Agent 范式重塑所驱动的 CPU 需求增量。除 CPU 外，本周国内 AI 编程平台 EazyDevelop 对会员订阅套餐提价，算力通胀实质性传导。
- **训推共振，算力需求极速释放。**我们判断，2026 年将是中国算力需求从“云端训练”向“训练+推理”双轮驱动转型的关键之年，算力缺口将在更多模态和更广场景的催化下，极速释放。1）训练侧：向高质量与多模态进阶。头部互联网厂商（字节、阿里、腾讯）持续迭代万亿参数级模型，智谱、DeepSeek 等新势力快速更新 MoE 架构。Scaling-law 在多模态领域延续性显著，以 Seedance 为代表的模型对视频、音频及文本的统一理解，推动底层算力需求从单一文本向高消耗的视频/3D 训练跃迁，对集群互联带宽与稳定性提出更高要求。2）推理侧：应用落地元年，需求斜率陡峭。 2026 年 AI 应用加速渗透，根据 QuestMobile 数据，豆包 APP 2025 年 12 月 MAU 已突破 2.26 亿；通义千问亦全面打通阿里生态。C 端流量与 AI 漫剧、编程等原生场景爆发，叠加 B 端垂类模型蓄势，共同驱动实时推理算力消耗大幅增长。我们预测，推理侧需求将成为拉动产业链增长的新引擎。
- **供给端外部边际改善，内部国产化放量。**我们认为，2026 年国内算力供给端将从单一的紧缺状态转向结构性平衡，充沛的算力资源将有效承接需求端的爆发，为算力产业链的业绩兑现奠定基础。1）外部方面，根据相关贸易许可清单的更新，NVIDIA H200（合规版）已正式获批进入中国市场，短期内将有效缓解头部互联网厂商在超大规模模型训练上的算力焦虑，助推模型迭代速度。2）与此同时，国产算力芯片的性能与生态建设已跨过“可用”向“好用”的拐点。华为昇腾系列、寒武纪思元系列以及海光深算系列在实战中快速迭代；大厂自研芯片战略进入收获期，百度昆仑芯、阿里平头哥及字节跳动自研芯片均开始大规模部署；而摩尔线程、沐曦股份、天数智芯等国产算力厂商成功上市或正在加速推进 IPO 进程，进一步丰富了市场供给。3）上游先进制程产能的扩充为芯片供应提供了底层保障。中芯国际刚刚发布的 2025 年全年财报显示，公司在先进制程及成熟制程的扩产上均取得突破性进展，全年产能利用率稳步回升，资本开支维持高位以确保新增产能的及时释放。
- **国产算力全链景气加速，有望量价齐升。**在供需双侧强逻辑的挤压下，我们预判 2026 年算力产业链将进入“全链通胀”周期，行业景气度将从核心芯片向 AIDC、云与算力服务、配套电力设备及服务器等环节全面外溢。投资策略上，鉴于字节跳动、阿里、腾讯等巨头较为明确的资本开支趋势，其供应链具有较高的业绩确定性，深度绑定头部互联网厂商（CSP）的供应链公司，或将获得显著的超额收益。

相关标的

相关标的：东阳光、寒武纪、海光信息、利通电子、豫能控股、协创数据、网宿科技、优刻得、润泽科技、亿田智能、华丰科技、神州数码、云天励飞、大位科技、润建股份、科华数据、中芯国际、华虹半导体、中科曙光、禾盛新材、奥飞数据、首都在线、云赛智联、瑞晟智能、浪潮信息、潍柴重机、欧陆通等。

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

图表目录

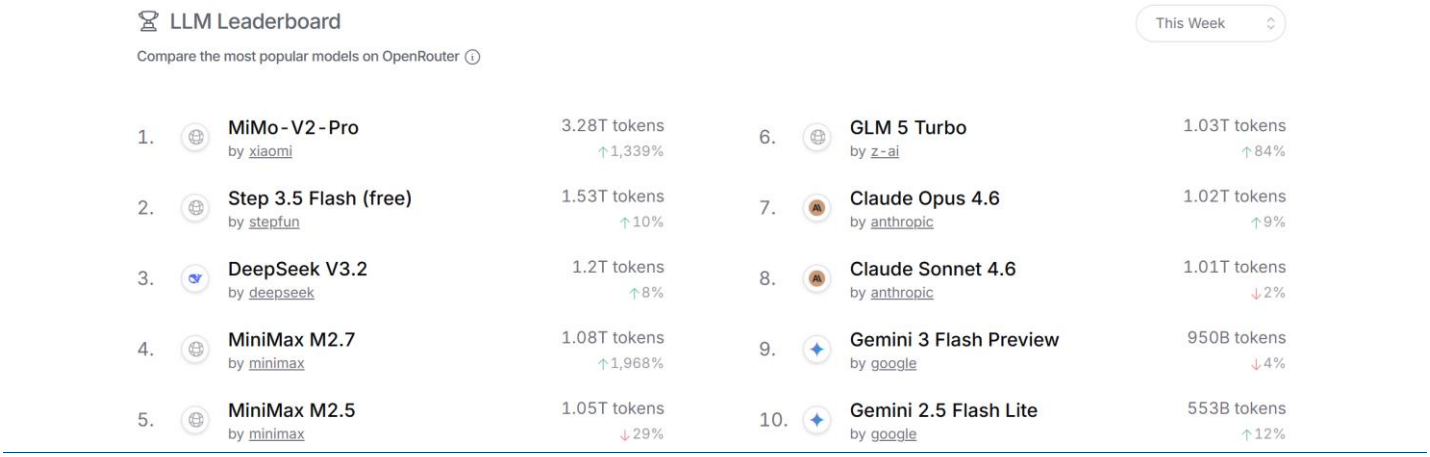
一、国产 Token 量破 140 万亿，CPU 涨价潮又起

1.1 国内日均 Token 调用量破 140 万亿，国产 AI 算力需求持续扩圈，国内算力厂商进入业绩临界点

国内日均 token 调用量两年增长超千倍，中国 AI 大模型调用量持续领跑全球。据国家数据局，2024 年初我国日均 Token 调用量为 1000 亿、2025 年底跃升至 100 万亿、2026 年 3 月国内日均 token 调用量已突破 140 万亿，实现两年超千倍增长。据 OpenRouter 数据，3 月 16-22 日全球 AI 大模型总调用量为 20.4 万亿 Token，上榜前十的 AI 大模型中，中国 AI 大模型周调用量为 7.36 万亿 Token，较前一周增长 56.9%，连续三周超越美国；3 月 23-27 日全球 AI 大模型总调用量 TOP10 阵营中，6 个国产模型上榜，持续保持领先水平。

国产算力、算力租赁厂商已经进入业绩临界点。得益于国内算力需求陡峭上行，部分国产算力、算力租赁厂商已经进入业绩临界点，如寒武纪 2025 年实现营业收入 64.97 亿元，同比大幅增长 453.21%，实现归母净利润 20.59 亿元，首次实现了全年利润的扭亏为盈，印证国产 AI 芯片正加速实现规模化放量；受益于智算中心投建需求与算力持续的供需错配，利通电子预计 2025 年实现归母净利润 2.70 亿元至 3.30 亿元，同比增速区间高达 996.83%至 1240.57%；宏景科技预计 2025 年实现净利润 3000 万元至 4350 万元，较上年同期顺利扭亏为盈。

图表1：3 月 23-27 日全球 AI 大模型总调用量前十阵营中，中国 AI 大模型占据六席



来源：OpenRouter，国金证券研究所

1.2 CPU 与 AI 编程端相继提价，算力通胀实质性传导

CPU 短缺加剧，涨价周期来临。25 年 10 月，据外媒 TrendForce 报道，英特尔公司正计划对其第 13 代 Raptor Lake 和第 14 代 Raptor Lake Refresh 处理器进行价格调整，涨幅最高可达 10%。26 年 1 月，据外媒 Wccftech 报道，AMD 和英特尔今年各自的服务器 CPU 库存均已售罄，大部分需求来自超大规模企业，他们希望将最新的服务器 CPU 集成到现有机架架构中，这也是过去几个季度需求显著增长的原因，因此，据称 AMD 和英特尔都计划将服务器 CPU 价格提高多达 15%，以确保供应保持稳定。据日经亚洲 3 月 25 日报道，英特尔与 AMD 已各自通知客户，将分别于 3 月和 4 月起上调全系列 CPU 价格，平均涨幅达 10-15%，部分产品涨幅更高；同时，交货周期将从之前的 1-2 周大幅延长至 8-12 周，个别情况下甚至将长达 6 个月。AI 算力需求爆炸式增长，AI 芯片巨头占用大量原材料与产能，英特尔与 AMD 面临产能扩张瓶颈，叠加原材料价格上涨，CPU 供给端持续承压，供需错配加剧，CPU 价格进入上行通道。

图表2：算力全链共振，步入涨价周期

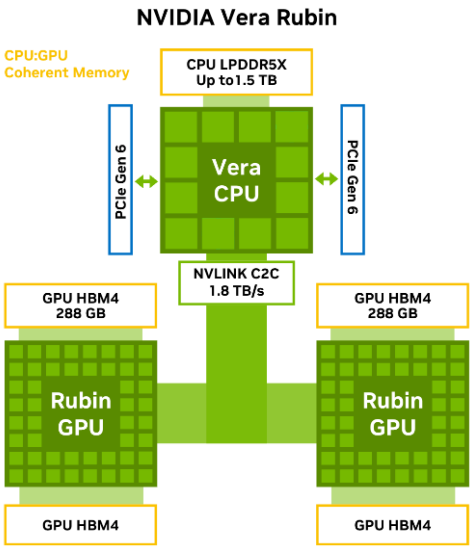
厂商	公告时间	具体内容
AWS	1 月 4 日	EC2 机器学习容量块价格上调约 15%，包括由 NVIDIA GPU 驱动的 P5en、P5e、P5、P4d，以及使用 AWS Trainium 的 Trn2 和 Trn1 实例，P5e.48xlarge 实例每小时费用由 34.61 美元涨至 39.80 美元，美国西部地区 P5e 实例价格从 43.26 美元/小时涨至 49.75 美元/小时。
三星电子	1 月 5 日 1 月 25 日	2026Q1 将 DRAM 价格较 2025Q4 提升 60%至 70%（包括向服务器、PC 及智能手机领域）。 2026Q1 NAND 闪存供应价格上调 100%以上。
SK 海力士	3 月 2 日 1 月 5 日 3 月 2 日	上调 2025Q2 DRAM 价格，DDR5 颗粒统一涨价 40%，部分产品涨幅高达 100% 。 2026Q1 将服务器 DRAM 价格较 2025Q4 提升 60%至 70%（向服务器、PC 及智能手机领域）。 上调 2025Q2 DRAM 价格，DDR5 颗粒统一涨价 40%，部分产品涨幅高达 100%。
谷歌云	1 月 27 日	对 Google Cloud、CDN Interconnect、Peering 以及 AI 与计算基础设施服务进行价格调整；北美地区数据传输价格从原来的 0.04 美元/GiB 上涨至 0.08 美元/GiB，涨幅达 100%；欧洲地区从 0.05 美元/GiB 上涨至 0.08 美元/GiB，涨幅为 60%；亚洲地区从 0.06 美元/GiB 上涨至 0.085 美元/GiB，涨幅约为 42%。

网宿科技	2 月 4 日	CDN 产品标准服务组流量上调 35%，CDN 产品快速回源通道流量上调 40%，对象存储产品存储空间上调 40%。
优刻得	2 月 11 日	对续签及新签用户的全线产品与服务进行价格上浮调整。
铠侠	2 月 14 日	2026Q1 NAND 闪存供应价格翻倍。
Hetzner	2 月 23 日	调高全线产品及服务报价，包括云服务、专用服务器、存储及负载均衡器等。云服务涨价最为显著，德国及芬兰的云服务价格涨幅在 30%到 38%之间，美国地区的 CCX 专用 vCPU 云服务器价格普遍涨幅在 30%水平。
腾讯云	3 月 11 日	GLM 5、MiniMax 2.5、Kimi 2.5 结束免费公测，转为按量计费。混元系列模型 Tencent HY2.0 Instruct 与 Tencent HY2.0 Think 的价格上调，其中，HY2.0 Instruct 输入价格从每千 tokens 0.0008 元调整为 0.004505 元，输出价格从 0.002 元调整为 0.01113 元；HY2.0 Think 输入价格从 0.001 元调整为 0.0053 元，输出价格从 0.004 元调整为 0.0212 元。
阿里云	3 月 18 日	平头哥真武 810E 等算力卡产品上涨 5%-34%，文件存储产品 CPFS（智算版）上涨 30%。
百度智能云	3 月 18 日	AI 算力相关产品服务上调约 5%-30%，并行文件存储等上调约 30%。
英特尔	3 月 19 日	计划 3 月起将全线 CPU 产品价格统一上调约 10%。
AMD	3 月 25 日	计划 4 月起将全系 CPU 价格上调约 10-15%。

来源：InfoQ，腾讯新闻网，网宿科技官网，优刻得官网，Hetzner 官网，36 氪，腾讯云官网，阿里云官网，百度智能云官网等，国金证券研究所

Agentic AI 驱动 CPU 角色跃迁，需求倍增。1）计算范式演进驱动 CPU 地位重估：对话式 AI 场景下，CPU 主要负责 Token 化等边缘计算工作，工作量仅占约 5%；代理型 AI 模式下，CPU 承担工具调用、任务编排、实时决策等大量非 AI 原生计算，消耗量占 AI 工作流的 80-90%。随 Agent 应用更普及、任务场景更复杂、工具调用密度更高，对系统协同度的要求提升，CPU 系统调度与资源协调能力的重要性被显著放大。2）Agent 爆发，CPU 需求巨大：ARM CEO Hass 表示，随企业不断扩大由智能体驱动的应用规模，数据中心对每吉瓦（GW）功耗提供的 CPU 算力需求将增长至当前的 4 倍以上，腾讯科技报道，据 Creative Strategies 预测，数据中心 CPU 需求将从 2026 年的 250 亿美元增长至 2030 年的 600 亿美元，若叠加 AI 智能体的需求，这一数字将接近 1000 亿美元。智能体工作负载的延迟主要来自 CPU 的工具处理任务，Agentic AI 时代 CPU 从支持模型的辅助计算单元，跃升为驱动模型的资源调度与管理枢纽，驱动 CPU 迎来高速增长。

图表3：NVIDIA Rubin 平台中 CPU 连接 GPU、内存与高速互联网络，成为系统调度与资源协调的核心枢纽



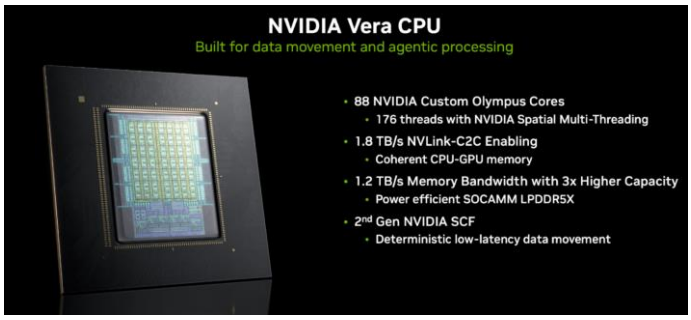
来源：NVIDIA Technical Blog，国金证券研究所

海外 CPU 龙头多款 CPU 新品密集发布。1）英伟达于 3 月 16 日发布专为智能体 AI 工作负载打造的 Vera CPU，搭载全新 Olympus 内核，在各类智能体应用场景和强化学习极端条件下提供更快的响应速度，较传统 x86 CPU 单线程性能提升 50%、每核心内存带宽提升至 3 倍、能效翻倍。2）ARM 于 3 月 24 日推出首个自研数据中心级 CPU，专为代理式 AI 工作负载设计，单机架性能可达 x86 平台的 2 倍以上、每 1GW 的 AI 数据中心算力资本支出节省高达 100 亿美元，已与 Meta、OpenAI、Cerebras、Cloudflare 等企业达成合作。科技巨头密集布局自研 CPU，以应对 Agentic AI 带来的 CPU 瓶颈，CPU 作为 Agent 场景的执行中枢，核心战略地位显著提升，市 场缺货与涨价潮并至，CPU 步入量价齐升周期。

国产 CPU 性能跃迁，跻身一线。1）阿里达摩院于 3 月 24 日发布新一代旗舰处理器玄铁 C950，为全球性能最高的 RISC-

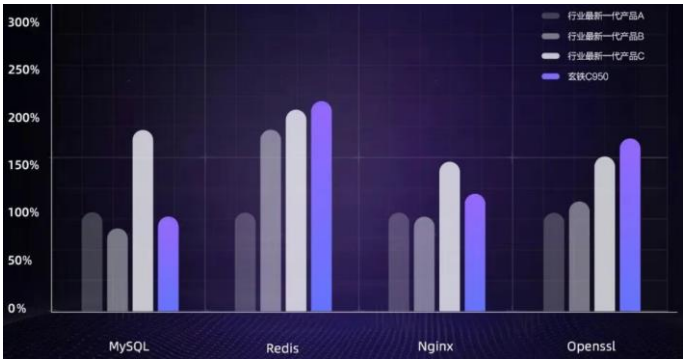
V CPU，在 SPEC Cint2006 测试中单核性能首次超过 70 分，搭载自研 AI 加速引擎，原生支持 Qwen3、DeepSeek V3 等千亿参数大模型。2）海光处理器的主要性能指标达到国际先进水平，具备卓越性能、安全可信以及完善成熟的生态，具有 7000，5000，3000 三大产品系列。

图表4：英伟达 Vera CPU 专为智能体 AI 工作负载打造



来源：NVIDIA Technical Blog，国金证券研究所

图表5：玄铁 C950 云计算典型场景性能达业界领先水平



来源：达摩院 DAMO 微信公众号，国金证券研究所

国产 AI IDE 开启提价，验证算力全链通胀向下游传导。EazyDevelop 为应对用户需求激增，进行计算资源池的全面扩容与架构升级，算力成本大幅增加，为保障用户体验，于 3 月 25 日对会员订阅套餐提价，个人标准版价格从 0.9 元/百万 tokens 调整为 1.99 元/百万 tokens、个人专业版价格从 0.598 元/百万 tokens 调整为 1.198 元/百万 tokens、团队标准版价格从 0.999 元/百万 tokens 调整为 1.999 元/百万 tokens、团队专业版价格从 0.7998 元/百万 tokens 调整为 1.3998 元/百万 tokens。AI 应用加速落地，AI 编程等平台并发请求量持续飙升，驱动算力消耗大幅攀升，EazyDevelop 此次提价正是算力通胀从基础设施层向应用层渗透的有力体现，再次印证国内算力斜率陡峭。

图表6：EazyDevelop 会员订阅套餐价格调整

类型	套餐名称	调整前价格(元/百万tokens)	调整后价格(元/百万tokens)
个人套餐	个人标准版	0.9	1.99
	个人专业版	0.598	1.198
团队套餐	团队标准版	0.999	1.999
	团队专业版	0.7998	1.3998

来源：卓易信息微信公众号，国金证券研究所

图表7：EazyDevelop 普惠专区各套餐定价

资源项目	EazyBot特惠	个人普惠版	个人标准版	个人专业版	团队普通版	团队专业版
价格	¥49	¥99	¥199	¥599	¥1,999	¥7,999
CPU	2.0 Core	2.0 Core	4.0 Core	8.0 Core	16.0 Core	64.0 Core
内存	4.0 GB	8.0 GB	16.0 GB	32.0 GB	64.0 GB	256.0 GB
存储	5.0 GB	10.0 GB	20.0 GB	50.0 GB	500.0 GB	2,048.0 GB
流量	5.0 GB	5.0 GB	10.0 GB	20.0 GB	1,024.0 GB	4,096.0 GB
知识库	5 GB	5 GB	5 GB	20 GB	50 GB	100 GB
AI大模型	20M Tokens	50M Tokens	100M Tokens	500M Tokens	1B Tokens	5B Tokens
数据库	1 个	1 个	1 个	4 个	5 个	10 个
MCP服务	5 个	5 个	5 个	20 个	50 个	100 个
Supabase	2 个	2 个	2 个	5 个	10 个	20 个
团队成员	-	-	-	-	5 个	20 个

来源：卓易信息微信公众号，国金证券研究所

二、训推共振，算力需求极速释放

2.1 大模型“军备竞赛”并未降温，向更高质量、更多模态加速进步

头部互联网厂商的护城河效应日益显著，字节跳动（豆包系）、阿里巴巴（通义系）、腾讯（元宝系）近期密集发布了万亿参数级的新一代主力模型，以智谱 AI、DeepSeek（深度求索）、Minimax、科大讯飞为代表的“AI 新势力”亦在快速迭代其 MoE 架构。更为关键的是，Scaling-law 在多模态领域也已展现出延续性，多模态视频生成是典型的“算力吞噬兽”，进一步显著利好算力基础设施。

字节 AI 全栈革新。据科创板日报 2 月 12 日报道，字节跳动火山引擎初步确定 2026 年 2 月 14 日发布豆包大模型的一系列重要升级发布，本次模型发布涉及豆包大模型 2.0、音视频创作模型 Seedance 2.0、图像创作模型 Seedream 5.0 Preview。此次升级，豆包大模型 2.0 将正式发布，基础模型能力和企业级 Agent 能力将有大幅提升。据字节跳动官方公众号，2 月 12 日豆包视频生成模型 Seedance2.0 正式发布，现已全面接入豆包和即梦产品，并上线火山方舟体验中心。

Seedance 2.0 达业界 SOTA 水平。Seedance 2.0 在运动场景下的生成可用率达到业界 SOTA 水平，其在人物动作建模方面，有着自然连贯、遵循现实世界运动规律的显著优势。它能高保真地合成时序精密的复杂交互场景，也能在特写镜头中展现出高度逼真的细节与严密的物理逻辑，宛如实拍。且 Seedance 2.0 能精准还原复杂脚本，保持稳定的主体一致性。模型还具备一定的编导思维，分镜设计和运镜贴合叙事逻辑。此外，Seedance 2.0 新增了视频编辑和视

频延长能力，每位用户都能像导演一样掌控全场。在各项评测中，Seedance 2.0 的综合表现达到行业领先水平。

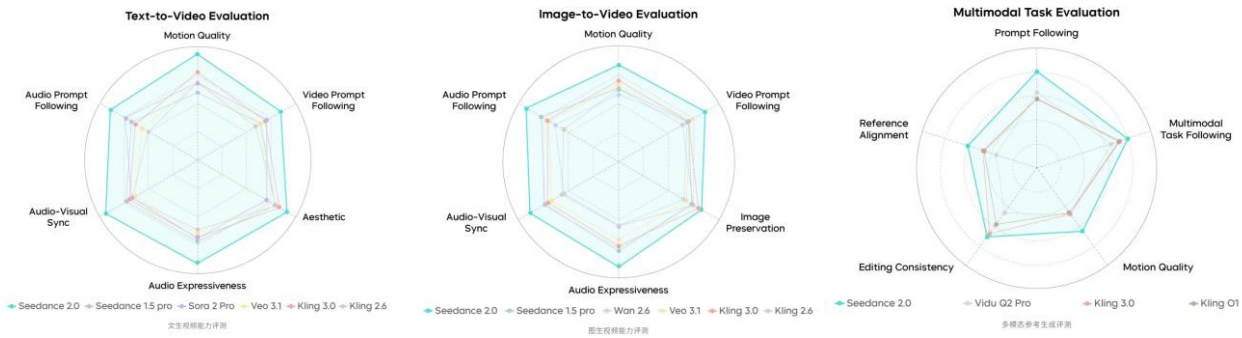
图表8：Seedance 2.0 生成视频展示（1）

来源：字节跳动官方公众号，国金证券研究所

图表9：Seedance 2.0 生成视频展示（2）

来源：字节跳动官方公众号，国金证券研究所

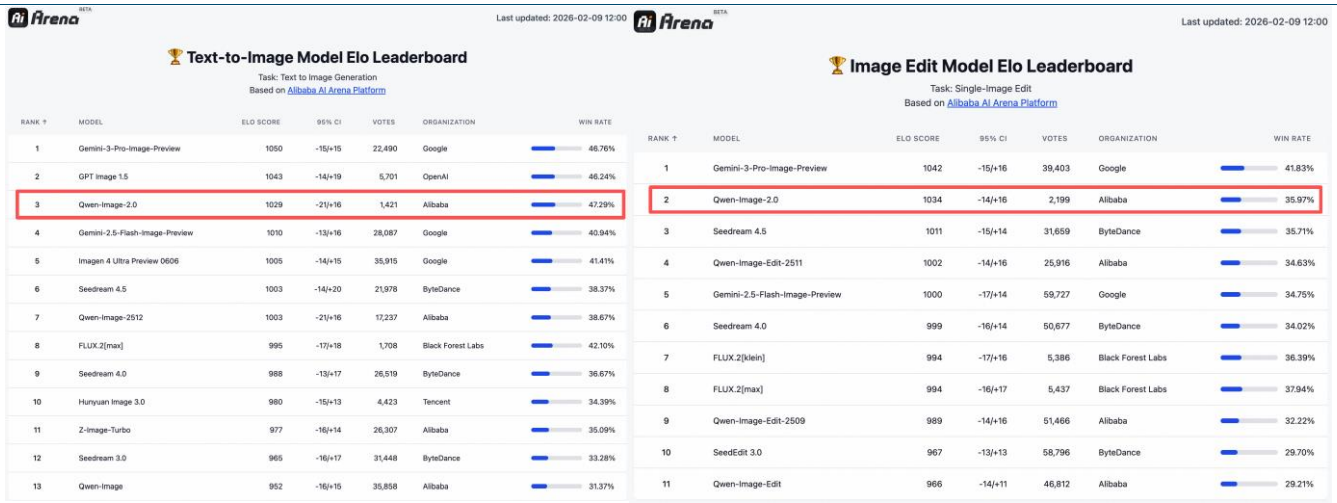
图表10：在各项评测中，Seedance 2.0 的综合表现达到行业领先水平



来源：字节跳动官方公众号，国金证券研究所

通义千问 Qwen-Image-2.0 发布，在文生图和图生图基准测试中获得优越性能。根据千问大模型官方公众号，2 月 10 日，阿里推出 Qwen-Image-2.0，新一代图像生成基础模型。Qwen-Image-2.0 主要特色包括：1）更专业的文字渲染：1k token 指令支持，直出专业信息图，包括 PPT/海报/漫画等。2）更细腻的真实质感：2k 分辨率支持，细腻刻画写实场景，包括人物/自然/建筑等。3）更强的语义遵循：理解生成一体化，生图编辑二合一。更轻量的模型架构：更小模型，更快速度。AI Arena 模型盲测数据显示，Qwen-Image-2.0 作为一个生图编辑二合一的模型，同一模型在文生图和图生图基准中获得优越性能。

图表11：AI Arena 模型盲测数据

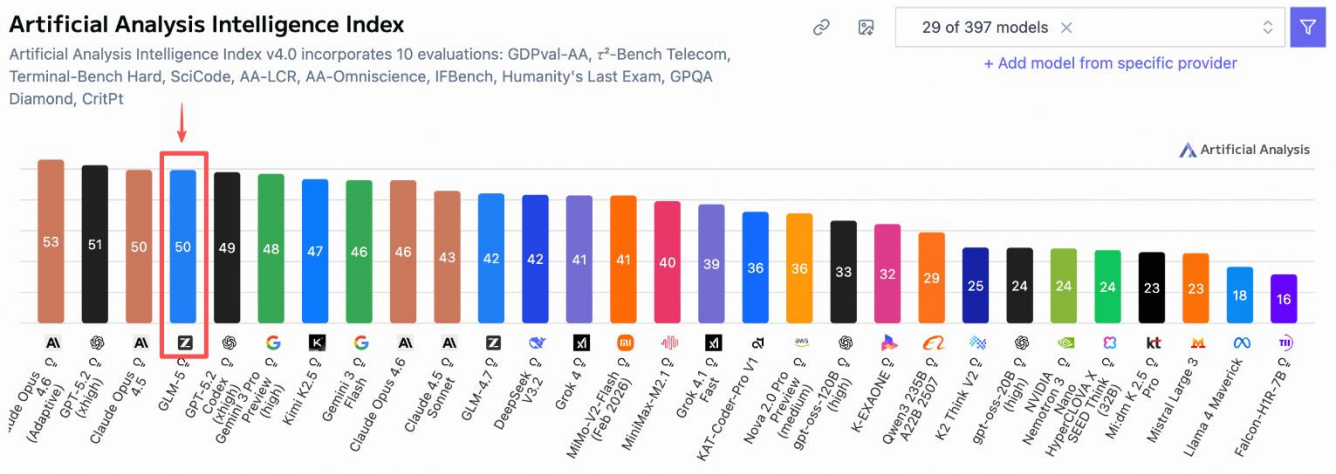


来源：千问大模型官方公众号，国金证券研究所

Deepseek 模型更新，上下文长度升级、知识库更新。据科创板日报 2 月 11 日报道，DeepSeek 在网页端和 APP 端进行了版本更新，支持最高 1M（百万）Token 的上下文长度，而去年 8 月发布的 DeepSeekV3.1 上下文长度拓展至 128K；同时，知识库更新至 2025 年 5 月。

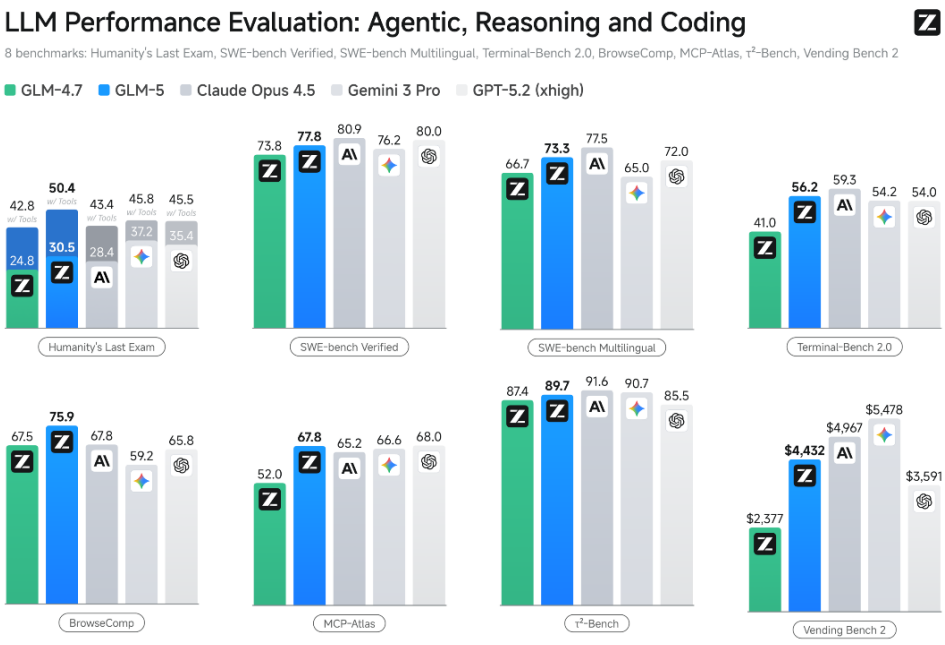
智谱发布新一代旗舰模型 GLM-5，在 Coding 与 Agent 能力上取得开源 SOTA。2 月 12 日，智谱上线并开源 GLM-5，其在 Coding 与 Agent 能力上，取得开源 SOTA 表现，在真实编程场景的使用体感逼近 Claude Opus4.5，擅长复杂系统工程与长程 Agent 任务。在全球权威的 Artificial Analysis 榜单中，GLM-5 位居全球第四、开源第一。GLM-5 基座能力全面演进：1）参数规模扩展：从 355B（激活 32B）扩展至 744B（激活 40B），预训练数据从 23T 提升至 28.5T，更大规模的预训练算力显著提升了模型的通用智能水平；2）异步强化学习：构建全新的“Slime”框架，支持更大模型规模及更复杂的强化学习任务，提升强化学习后训练流程效率；提出异步智能体强化学习算法，使模型能够持续从长程交互中学习，充分激发预训练模型的潜力；3）稀疏注意力机制：首次集成 DeepSeek Sparse Attention，在维持长文本效果无损的同时，大幅降低模型部署成本，提升 Token Efficiency。

图表12：在全球权威的 Artificial Analysis 榜单中，GLM-5 位居全球第四、开源第一



来源：智谱官方公众号，国金证券研究所

图表13：GLM-5 Coding 能力及 Agent 能力取得开源 SOTA



来源：智谱官方公众号，国金证券研究所

2.2 推理算力需求正以超预期的斜率上升

以字节、阿里、腾讯三家大厂为例，随着大厂围绕 AI 入口展开高强度竞争，流量获取与生态打通形成合力，推动 AI

从工具属性向高频服务入口跃迁。在用户规模与使用深度双提升的背景下，推理侧算力需求正以超预期斜率快速抬升。

- 字节：QuestMobile 数据显示，豆包 2025 年用户规模持续增长，第四季度月均活跃用户高达 2.3 亿户，连续两个季度登顶行业榜首，月均下载用户数也连续 3 个季度位居行业第一，而在豆包 APP 强势登陆央视春晚后，凭借全民级曝光，其推广程度有望再上台阶。
- 阿里：千问借助阿里巴巴生态的资源优势，打通了淘宝闪购、飞猪、盒马、大麦、高德、支付宝等多款应用，融合生态内的交易体系、地理位置服务、出行资源、电商资源等能力，让 AI 能够丝滑地实现点外卖、购物、订机票等相对复杂的操作，真正化身为用户身边的“全能管家”。数据显示，上线两个月，千问 C 端（消费者端）月活跃用户数已突破 1 亿，在学生和白领人群中增长迅猛。

腾讯：2026 年开年以来，腾讯在 AI 领域动作频频：先是启动社交 AI“元宝派”内测，接着狂撒 10 亿元红包为元宝派拉新，引发 AI 圈红包大战。

除了模型本身的入口流量，AI 漫剧、AI 编程等原生应用的快速爆发，AI 医疗、智能制造等垂类模型的蓄势待发。

多模态驱动视觉质变及技术红利释放，动态漫、AI 漫剧已成“爆款”。 2025 年多模态技术收敛推动视觉表现力跨越式提升。供给端，AI 大幅压降制作周期，实现低成本批量化产出；需求端，紧凑节奏与高情绪密度精准击中用户痛点。供需双侧适配，驱动赛道从边缘迈向主流。根据短剧自习室公众号统计，从 2025 年度抖音端漫剧累计播放量 TOP100 榜单来看，共有 52 部沙雕漫、28 部 AI 漫剧、17 部动态漫、3 部解说漫上榜，分别占比 52%、28%、17%、3%。其中仅有 1 部突破 5 亿播放——AI 漫《斩仙台下，我震惊了诸神》，10 月份上线，截至 26 年 1 月初累计播放量 10.6 亿，成为当之无愧的 2025 年度“剧王”。

图表14：2025 年累计抖音播放量 TOP10 漫剧统计，前十中 AI 漫及动态漫居多

2025年累计抖音播放量TOP100漫剧

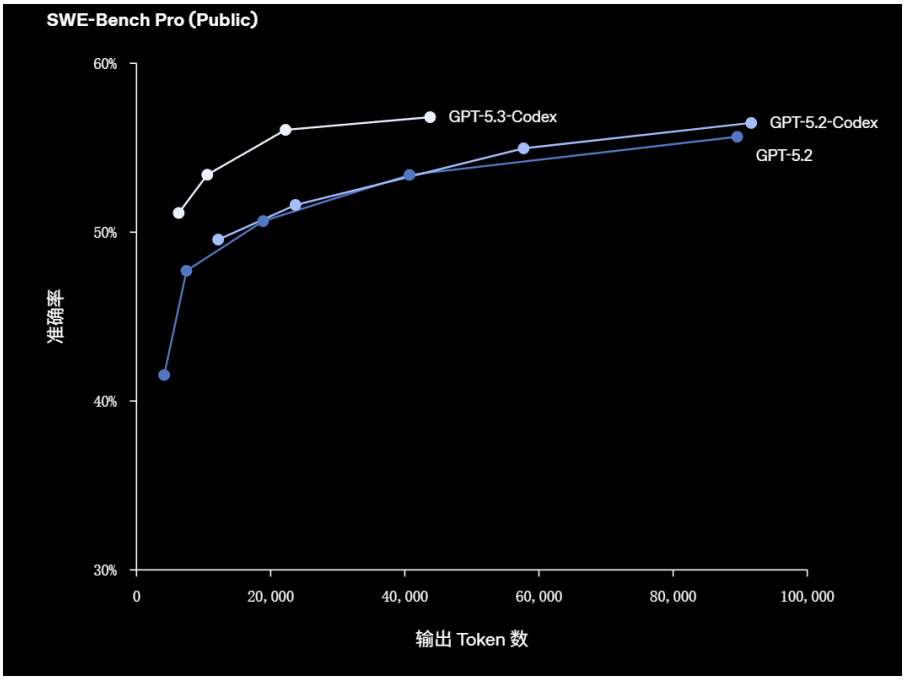
短剧自习室综合整理 | 统计期：2025.1.1-2025.12.31

排名	剧名	累计播放量	类型
1	斩仙台下 我震惊了诸神!	1011362087	AI漫
2	全民诡异 开局掌握零元购【第一季】	365575519	动态漫
3	气运擂台 华夏诸神震惊全世界!	337751415	AI漫
4	洪荒 代管截教 忽悠出了一堆圣人	279649175	AI漫
5	我在末世当老板 员工全是S级变异体	266314072	AI漫
6	冰冰巨鱼	259313489	解说漫
7	大明贤婿【第一季】	247600354	沙雕漫
8	地狱模式 你能活到第几关【第一季】	247131902	动态漫
9	边关不用我守	244932308	AI漫
10	无限突破 敌人越强我越强【第一季】	221857292	动态漫

来源：短剧自习室公众号统计，国金证券研究所

AI 编程从辅助工具迈向原生入口，开发工作流进入 Agent 时代。随着模型在长上下文理解与多文件协作能力上的突破，AI 编程正由“代码补全插件”升级为具备规划、生成与验证能力的编程 Agent。2 月 2 日 OpenAI 推出编码助手 Codex 独立 App，并发布底层 GPT-5.3-Codex 模型，独立应用程序 Codex 在推出后的第一周下载量就超过了 100 万次且 Codex 用户总数环比增长 60%。相比于之前的 AI 编程大模型，GPT-5.3-Codex“超越编程”，旨在支持软件生命周期中的所有工作，包括调试、部署、监控、编写公关需求文档（PRD）、编辑文案、用户研究、测试、指标分析等，可以在几天时间内从零开始构建功能高度复杂的游戏和应用程序，在 SWE-Bench Pro 和 Terminal-Bench 上创下了行业新高，并在 OSWorld 和 GDPval 上表现强劲。

图表15：GPT-5.3-Codex 在 SWE-Bench Pro 上达到了顶尖(state-of-the-art)水平



来源：OpenAI 官网，国金证券研究所

全球巨头抢滩，专业化闭环重塑医疗生态。1）爆款验证 C 端刚需：2025 年 12 月 25 日，蚂蚁集团战略升级“蚂蚁阿福”，实现从单一工具向“AI 健康伙伴”的形态跃迁。发布当日即冲至苹果应用商店总榜 TOP3，现象级表现有力验证了 C 端用户在严肃健康场景下强烈的交互需求与信任交付。2）垂直闭环构筑壁垒：不同于通用大模型的泛化问答，阿福依托专业医疗底座，深度打通“智能问诊-在线挂号-医保支付”全链路。这种将信息流直接转化为服务流的闭环能力，确立了其在医疗垂直场景下不可复制的竞争优势。3）产业趋势全球共振：2026 年 1 月 8 日，OpenAI 跟进推出 ChatGPT Health。继蚂蚁阿福之后，全球 AI 领军者不约而同切入医疗赛道，进一步强化了该细分领域“高壁垒、高单价、高粘性”的产业逻辑，板块配置价值显著提升。

图表16：蚂蚁阿福“健康陪伴”功能

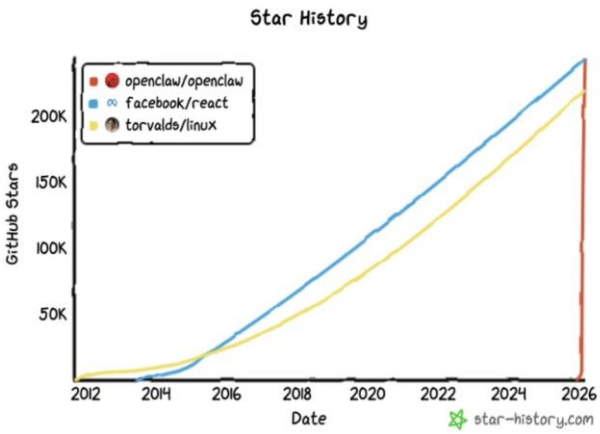
来源：21 世纪经济报道，国金证券研究所

Agent 生态持续扩张，业界创新频出。1）Anthropic 上调营收预测，Claude Code 助推业绩增长。Anthropic 预计今年销售额将增长四倍至 180 亿美元，而明年将达 550 亿美元，远高于公司去年夏天的预测；AI 编码助手 Claude Code 去年 11 月的年化收入已超过 10 亿美元，约占当时总年化收入的 14%，帮助该公司在去年年底实现了超过 90 亿美元的年化收入。2）K2.5 实现 Agent 集群。K2.5 能根据任务需求，现场调度多达 100 个分身，并行处理 1500 个步骤，且所有的角色分配与任务拆解，无需预设，全由 K2.5 现场决策。3）OpenClaw 爆火带动 Token 消耗爆发。AI Agent 工具 OpenClaw 能在本地设备上 7×24 小时自主执行任务，其高频自主调用机制使得推理 Token 消耗量较传统对话式 AI 呈现指数级跃升，Kimi、Minimax 等国内大模型厂商，通过与 OpenClaw 实现生态合作，均由 OpenClaw 带来较高的 Token 消耗量。

图表17：Kimi K2.5 模型使用多个角色的 agent 集群完成文献综述

来源：月之暗面 kimi 公众号，国金证券研究所

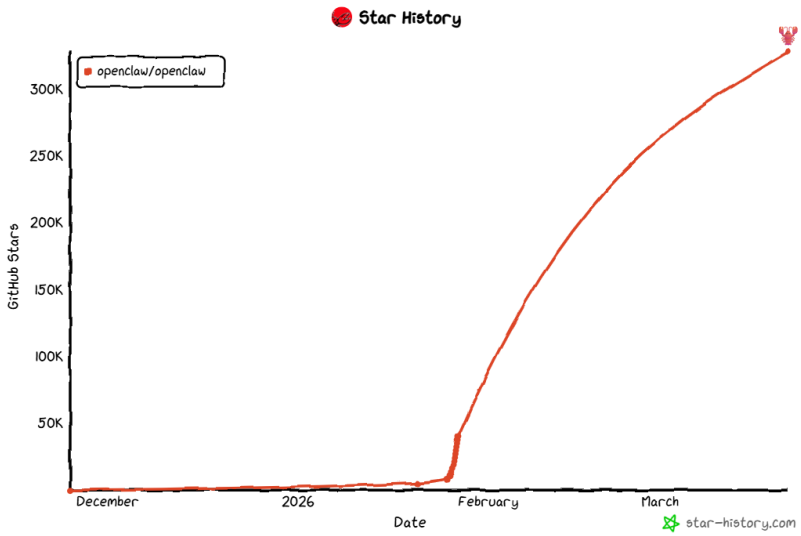
图表18：OpenClaw 项目正式登顶 Github 榜首



来源：OpenClaw X 平台官方账号，国金证券研究所

从 Prompt 到长 Agent 变迁。据英伟达 GTC 2026 大会博客：人工智能正从简单的、基于 Prompt 的工具发展成为能够推理、规划和行动的智能、长期运行的系统。这些自主 Agent 不仅能生成文本，还能编写代码、调用工具、分析数据、模拟结果并持续改进。大模型聚合平台 OpenRouter 的报告也提到：LLM 的使用正从单回合交互转向智能推理，模型需要进行规划、推理和执行，并跨越多个步骤。它们不再生成一次性响应，而是协调工具调用、访问外部数据，并迭代优化输出以达成目标。早期证据表明，多步骤查询和链式工具使用正在增加。随着这种范式的扩展，评估标准将从语言质量转向任务完成度和效率。下一个竞争前沿是模型执行持续推理的有效性，这一转变最终可能会重新定义大规模智能推理在实践中的意义。科创板日报 2026 年 3 月 3 日报道，发布仅四个多月的开源智能体项目 OpenClaw 创造了历史——以超过 24.8 万的 GitHub 星标数正式登顶星标榜，超越 Linux 成为 GitHub 平台上最受欢迎的开源项目。OpenClaw 的爆炸性扩散，标志着长运行 Agent 从实验阶段进入大规模生产部署。

图表19：OpenClaw 的 GitHub 星标增长趋势



来源：star-history，国金证券研究所

Agent 上下文长度结构性增长。Anthropic 发表的测试数据表示：智能体通常比聊天交互消耗的令牌多约 4 倍，而多智能体系统比聊天消耗的令牌多约 15 倍。 英伟达 2026 年 1 月的技术博客对下一代 AI 工厂的计算需求做出了明确定性：为了大规模地提供这些功能，下一代 AI 工厂必须处理数十万个输入标记，以提供智能推理、复杂工作流程和多模态管道所需的长期上下文，同时在功率、可靠性、安全性、部署速度和成本的限制下维持实时推理。可见，Agent 任务中模型所需处理的 Token 数量往往远超传统问答场景。

三、供给端外部边际改善，内部国产化加速放量

外部方面，NVIDIA H200（合规版）正式获批进入中国市场，短期内将有效缓解算力焦虑。据观察者网，1 月 13 日，美国特朗普政府正式批准英伟达对华出口 H200 人工智能（AI）芯片，根据美国商务部发布声明，商务部下属机构工

业与安全局（BIS）正在修订对某些半导体向中国出口的许可审查政策——从推定拒绝改为逐案审查。H200 短期内将有效缓解头部互联网厂商在超大规模模型训练上的算力焦虑，助推模型迭代速度。

国产算力芯片的性能与生态建设已跨过“可用”向“好用”的拐点。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20：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

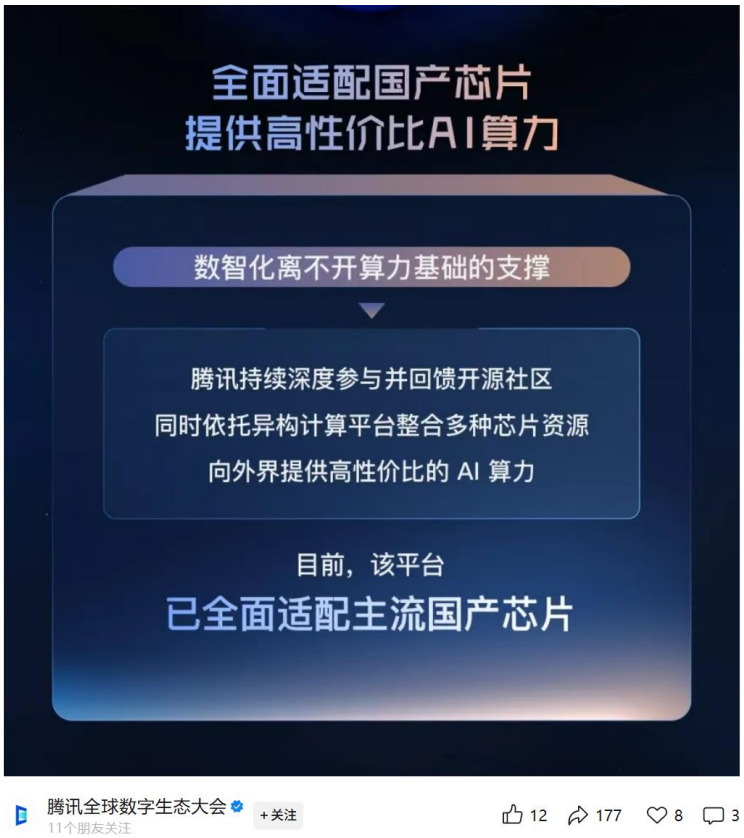
供给侧：上游先进制程产能的扩充为芯片供应提供了底层保障。中芯国际刚刚发布的 2025 年全年财报显示，2025 年第四季度公司营收为 24.89 亿美元，环比增长 4.5%；在本季度增加了 1.6 万片 12 英寸晶圆产能的基础上，公司产能利用率保持在 95.7%；整体 8 英寸产能利用率超过 100%，整体 12 英寸接近满载，这主要是产业链重构和迭代效应持续作用的结果。2026 年全年的指引为：营收增长预计将高于同市场产业界的平均水平，资本开支预计与 2025 年大致持平。

图表21：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22：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

四、国产算力全链通胀，有望量价齐升

在供需双侧强逻辑的挤压下，我们预判 2026 年算力产业链将进入“全链通胀”周期，行业景气度将从核心芯片向 AIDC、云与算力服务、配套电力设备及服务器等环节全面外溢。

- CPU 涨价：自 25Q4 起，部分 CPU 大厂已步入涨价周期。25 年 10 月，据外媒 TrendForce 报道，英特尔公司正计划对其第 13 代 Raptor Lake 和第 14 代 Raptor Lake Refresh 处理器进行价格调整，涨幅最高可达 10%；26 年 1 月，据外媒 Wccftech 报道，AMD 和英特尔今年各自的服务器 CPU 库存均已售罄，大部分需求来自超大规模企业，他们希望将最新的服务器 CPU 集成到现有机架架构中，这也是过去几个季度需求显著增长的原因，因此，据称 AMD 和英特尔都计划将服务器 CPU 价格提高多达 15%，以确保供应保持稳定。
- 算力/云厂商涨价：1 月 23 日，亚马逊云科技（AWS）近日上调其 EC2 机器学习容量块（Capacity Blocks for ML）价格约 15%，其中 p5e.48xlarge 实例每小时费用由 34.61 美元涨至 39.80 美元；1 月 27 日，谷歌云正式官宣涨价，自 2026 年 5 月 1 日起，对 Google Cloud、CDN Interconnect、Peering 以及 AI 与计算基础设施服务进行价格调整；2 月 12 日，智谱宣布对 GLMCoding Plan 套餐价格体系进行结构性调整，整体涨幅自 30%起。

图表23：智谱宣布 GLMCoding Plan 价格调整

GLM Coding Plan 价格调整函

尊敬的客户：
您好！

感谢您长期以来对智谱的支持。

近期，GLM Coding Plan 市场需求持续强劲增长，用户规模与调用量快速提升。为保障高负载下的稳定性与服务质量，我们同步加大算力与模型优化投入，产品能力持续升级。

基于实际使用情况与资源投入变化，**我们决定对 GLM Coding Plan 套餐价格体系进行结构性调整。**

调整内容如下：

- 取消首购优惠，保留按季按年订阅优惠
- 套餐价格进行结构性调整，整体涨幅自 30% 起
- 已订阅用户价格保持不变

生效时间：2026 年 2 月 12 日

我们将持续投入，保障产品能力与服务稳定性。

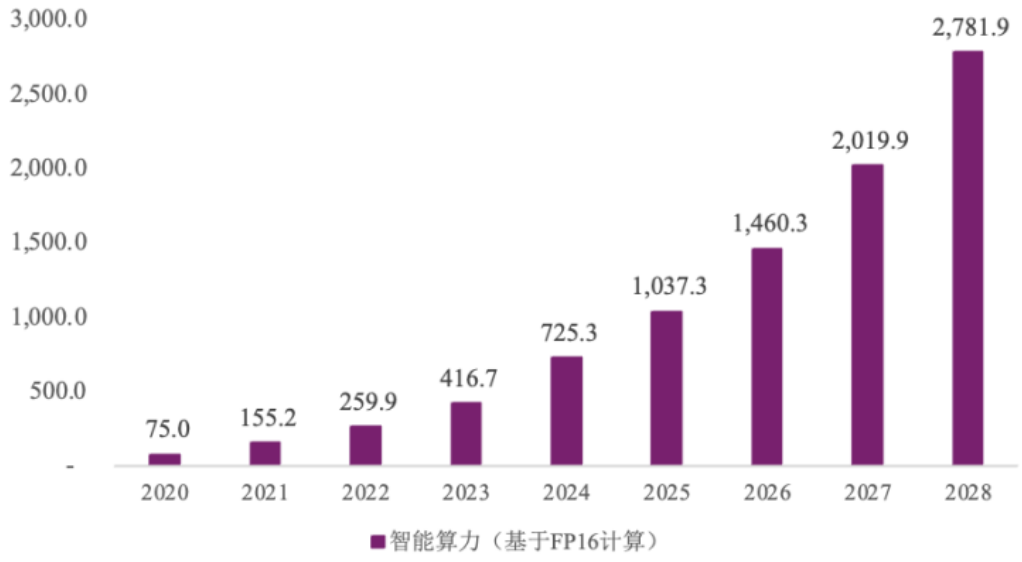
北京智谱华章科技股份有限公司
2026 年 2 月 12 日

z.ai

来源：智谱官方公众号，国金证券研究所

- AIDC 投建力度持续高景气：1）海内外大厂 CapEx 持续高增，硅谷四大科技巨头 2026 年 CapEx 将高达 6500 亿美元，AI 军备竞赛进一步加剧，具体看：亚马逊成为四家中投入规模最大的企业，将 2026 年资本支出目标定在 2000 亿美元；Alphabet 的资本支出计划高达 1750 亿美元-1850 亿美元，同比接近翻倍；Meta 预计全年资本支出将增至 1350 亿美元，同比增幅或达 87%；微软同期公布其第二季度资本支出同比增长 66%，预计其截至 6 月的财年资本支出将逼近 1050 亿美元。2）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 75.0EFLOPS，到 2028 年预计将达到 2,781.9EFLOPS，预计 2020-2028 年复合增长率达到 57.1%。在多维度数据与产业动态的交叉印证下，AI 算力基础设施投建力度维持高位，AIDC 环节呈现持续高景气扩张态势。

图表24：2020-2028 年中国智能算力规模及预测（EFLOPS，基于 FP16 计算）



来源：沐曦招股说明书，国金证券研究所

五、相关标的

相关标的：东阳光、寒武纪、海光信息、利通电子、豫能控股、协创数据、网宿科技、优刻得、润泽科技、亿田智能、华丰科技、神州数码、云天励飞、大位科技、润建股份、科华数据、中芯国际、华虹半导体、中科曙光、禾盛新材、奥飞数据、首都在线、云赛智联、瑞晟智能、浪潮信息、潍柴重机、欧陆通等。

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。

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。

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