Token经济风起之时

投资评级：推荐（维持）

---计算机行业深度报告

华龙证券研究所 计算机行业

分析师：孙伯文

SAC执业证书编号：S0230523080004

邮箱：sunbw@hlzq.com

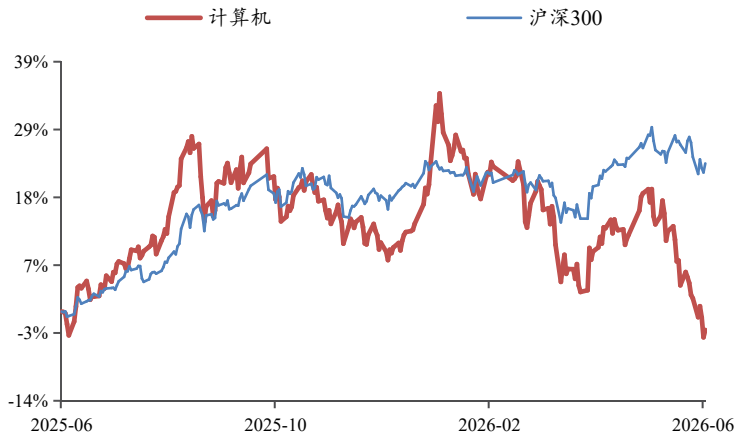
分析师：朱凌萱

SAC执业证书编号：S0230526030001

邮箱：zhulx@hlzq.com

2026年06月14日

最近一年市场走势



相对沪深300表现（2026.06.12） （单位：%）

表现	1M	3M	12M
计算机行业	-18.59	-16.29	-0.97
沪深300	-4.42	2.32	23.63

相关报告

《CPU重要性提升，智谱刷新模型API速度上限—计算机行业周报》2026.05.26
《AI产业高景气度持续验证，积极把握AI主线机会 —计算机2025年年报及2026年一季报业绩综述》2026.05.18
《看好算力通胀向下传导，关注国产算力投资机会—计算机行业周报》2026.04.14

报告摘要

摘要：

- **Agent与多模态支撑token攀升。**随着AI Agent处理任务日趋复杂，其推理深度与调用链路不断延伸，将驱动底层Token消耗呈数量级跃升。IDC预计，活跃Agent的数量将从2025年的约2860万，快速攀升至2030年的22.16亿。这意味着五年后，能够帮助企业执行任务的数字劳动力数量将是今天的近80倍，年复合增长率139%。数据显示，年度Token消耗量预计将从2025年的0.0005 Peta Tokens激增至2030年的152667 Peta Tokens，年复合增长率高达3418%。此外，多模态成为新的竞技场。字节于2026年2月发布第三代AI视频生成模型Seedance2.0，支持最长15秒视频时长，新增多模态输入；多镜头叙事、音画同步、角色一致性等核心能力领先于全球主流竞品。快手、Minimax等亦快速跟进新一代多模态模型，有望进一步支撑token消耗量增长。
- **国产token的性价比凸显。**根据OpenRouter的数据，在最新一周（2026年6月2日-6月8日）全球大模型调用量排名前五的模型中，中国占据四席，合计贡献了前五名总调用量的86.47%。我们认为，国产头部AI模型在tokens消耗量上已稳居全球第一梯队，部分国产模型凭借高性价比使得增长速率更加陡峭，更预示着持续的扩张潜力。未来，“每瓦特Token吞吐量”将成为衡量AI企业竞争力的核心指标。这意味着在固定的电力预算下，谁能以更高的能源效率生产更多Token，谁就拥有最低的生产成本和最强的市场竞争力。从电费角度看，中国工业用电0.4-0.6元/度，美国0.8-1.2元/度，叠加国内电力输送和响应速度优势，国产token具备显著的电价优势支撑。
- **Token经济的涨价传导。**上游算力硬件开启涨价潮，本轮涨价的核心驱动因素是AI算力需求的爆发。全球AI大模型和数据中心的加速扩张，导致供需失衡。CPU：英特尔和AMD 已分别于2026年3月底和4月起，通知客户将上调全系列CPU产品价格。此次涨价平均幅度在10%至15% 之间，部分产品涨幅更高。芯片代工：德州仪器宣布将于2026年4月1日启动近一年内的第三轮调价，涨价幅度最高达85%。与此同时，中芯国际、世界先进、华虹、力积电、晶合集成等芯片代工巨头纷纷确认或计划上调代工价格，涨幅普遍在5%至20%之间。存储：三星电子、SK海力士等内存大厂已于2025年开始上调存储芯片价格。自2025年3月至2026年5月，消费级DRAM 16GB DDR4价格从约200元暴涨至2000元，涨幅高达约900%，16GB DDR5涨幅达300%；NAND闪存方面，256GB和512GB产品价格普遍上涨了200%至250%。云服务：2026年一季度已经进入涨价实证期，主要是因为AI硬件成本上行叠加需求快速增加。算力租赁：大模型从训练走向推理，算力需求将进入7x24小时持续消耗。高端机型供需仍处于不平衡状态，预计未来1-2年仍有较强价格支撑。
- **投资建议：**AI Agent与多模态爆发正驱动Token消耗呈指数级激增，国产模型凭借极高性价比与电价优势占据全球流量主导，催生巨大算力需求；高端算力供需失衡推动上游算力硬件至下游云服务开启涨价潮，Token经济正值风起之时。维持计算机行业**“推荐”**评级，建议关注：（1）国产芯片：寒武纪（688256.SH）、海光信息（688041.SH）、中国长城（000066.SZ）；（2）云厂商及IDC：润泽科技（300442.SZ）、润建股份（002929.SZ）、优刻得-W（688158.SH）、首都在线（300846.SZ）、大位科技（600589.SH）、网宿科技（300017.SZ）；（3）算力租赁：协创数据（300857.SZ）、宏景科技（301396.SZ）。

- **风险提示：**所引用数据资料的误差风险；AI投资力度不及预期；AI产品竞争加剧；重点关注公司业绩不达预期；政策标准出台速度不及预期。

表：重点关注公司及盈利预测

股票代码	股票简称	2026/06/12	EPS（元）				PE				投资评级
		股价（元）	2025A	2026E	2027E	2028E	2025A	2026E	2027E	2028E	
000066.SZ	中国长城	15.51	-0.02	0.05	0.10	0.27	/	343.9	150.6	57.3	未评级
002929.SZ	润建股份	58.63	0.14	1.32	2.20	3.55	284.4	44.6	26.7	16.5	未评级
300017.SZ	网宿科技	14.31	0.33	0.36	0.45	0.58	37.4	39.8	31.8	24.7	增持
300442.SZ	润泽科技	73.53	3.00	2.03	2.55	3.17	17.6	36.3	28.8	23.2	未评级
300846.SZ	首都在线	21.27	-0.34	-0.004	0.27	0.11	/	/	78.8	195.0	增持
300857.SZ	协创数据	223.70	3.38	4.95	7.42	10.99	49.9	45.2	30.2	20.4	未评级
301396.SZ	宏景科技	172.56	0.17	2.01	4.13	9.95	385.5	85.9	41.8	17.3	未评级
600589.SH	大位科技	9.35	-0.01	0.07	0.09	0.13	/	132.4	101.5	70.4	未评级
688041.SH	海光信息	280.00	1.10	2.00	2.74	3.88	204.0	140.0	102.2	72.1	增持
688158.SH	优刻得-W	35.00	-0.16	0.11	0.31	0.53	/	318.2	112.9	66.0	增持
688256.SH	寒武纪	1240.00	4.93	11.00	17.19	27.01	275.0	112.7	72.1	45.9	增持

数据来源：Wind，华龙证券研究所，注：网宿科技（300017.SZ）、优刻得-W（688158.SH）盈利预测来源于华龙证券研究所；首都在线（300846.SZ）、海光信息（688041.SH）、寒武纪（688256.SH）2026-2027年盈利预测来源于华龙证券研究所，2028年盈利预测来源于Wind一致预期；其余所有公司盈利预测来源于Wind一致预期。

目录

Token是什么？能够代表推理侧算力需求

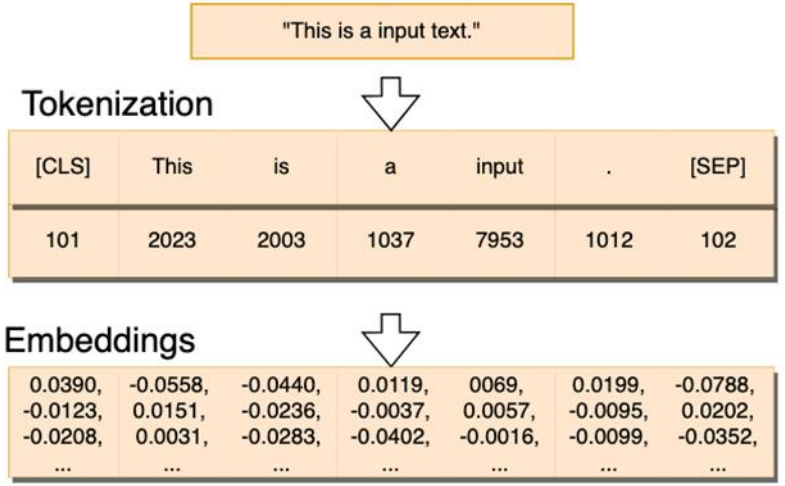
- Token是AI大模型处理文本的基础计算单元，可理解为模型理解和生成语言的“原子”。在技术上，它将输入的每个字、词或符号转化为数字向量进行处理，然后进行理解和生成；在商业上，它则是云服务对AI调用进行计费的核心依据。Token调用量的多少，直接对应着GPU算力资源的消耗规模。
- Token定义为词元，政策重视程度提升。2026年3月23日，Token的中文译名被正式表述为"词元"。同时，Token被定义为大模型处理信息的最小信息单元，具有可计量、可定价、可交易的特征，是连接技术供给与商业需求的结算单位。

表1：中英文文本内容大致对应的token数量

文本内容	大致对应token数量
英文	一个token可能是一个单词（如"apple"）、一个子词（如"un-"、"happy"）或一个标点符号
中文	一个token通常对应一个汉字或常见词组;通常对应1-1.5倍字数

资料来源：腾讯云社区，华龙证券研究所

图1：字符转化为token示例

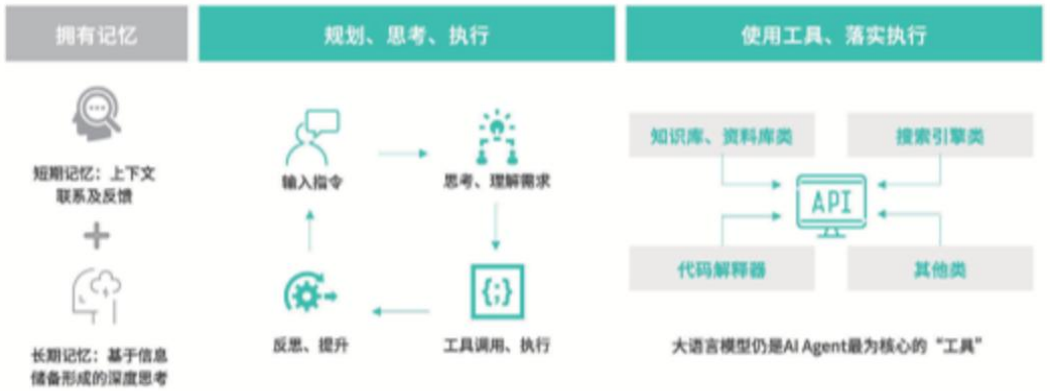


资料来源：华为云开发者社区，华龙证券研究所

Agent将推动token消耗量激增

- 随着AI Agent处理任务日趋复杂，其推理深度与调用链路不断延伸，将驱动底层Token消耗呈数量级跃升。IDC预计，活跃Agent的数量将从2025年的约2860万，快速攀升至2030年的22.16亿。这意味着五年后，能够帮助企业执行任务的数字劳动力数量将是今天的近80倍，年复合增长率139%，换言之，平均每年活跃Agent数量都将以超过一倍的速度增长。Agent真正干活的频率增长得更快，年执行任务数将从2025年的440亿次暴涨至2030年的415万亿次，年复合增长率高达524%。据IDC2025年统计数据显示，年度Token消耗量预计将从2025年的0.0005 Peta Tokens（1 Peta=1000万亿）激增至2030年的152667 Peta Tokens，年复合增长率高达3418%。

图2：Agent的思考方式与运行逻辑



资料来源：中国信息协会，《2025年全球AI Agent行业洞察报告》，华龙证券研究所

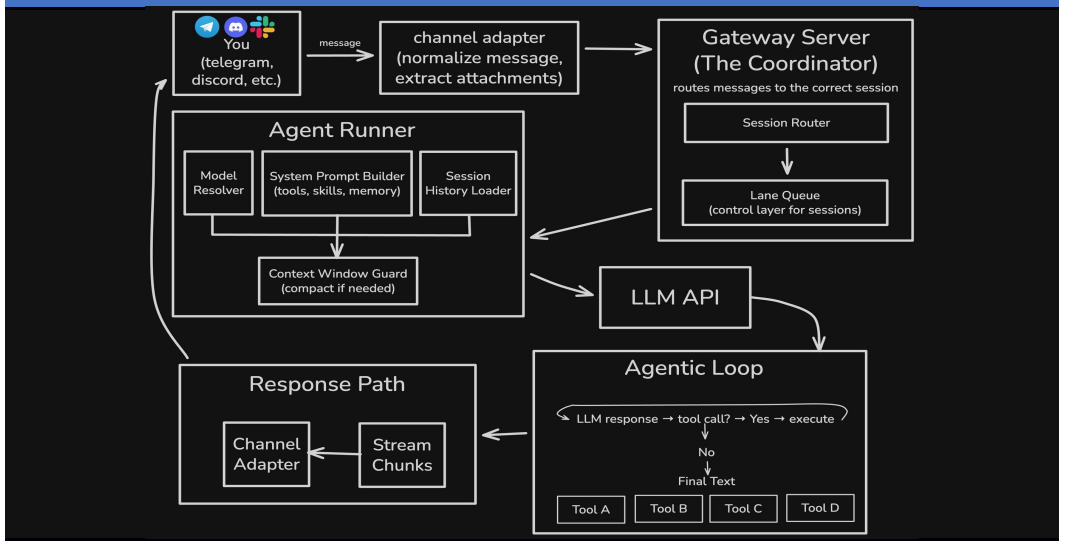
图3：全球企业活跃Agent数量与年度执行量预测（统计时间：2026年1月）

资料来源：IDC，华龙证券研究所

Agent爆款产品：OpenClaw开启“龙虾时刻”

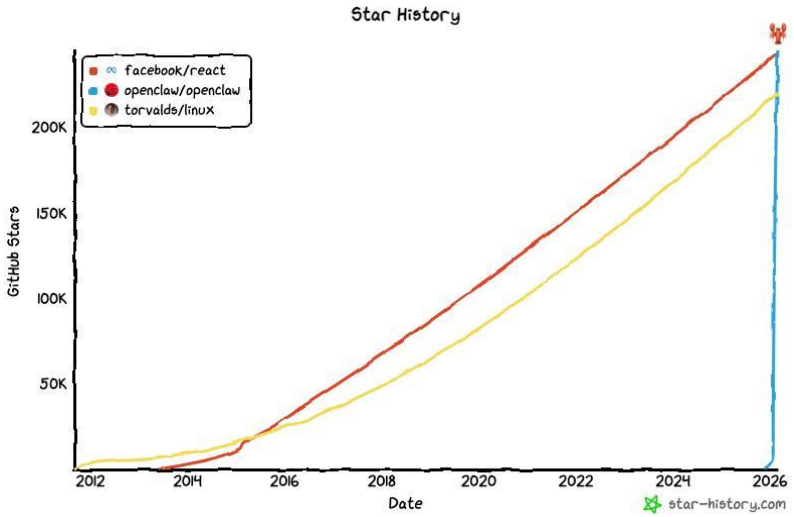
- OpenClaw是由奥地利开发者Peter Steinberger发布的一款开源个人AI智能体项目，上线后在极短时间内迅速走红。2026年1月28日，腾讯云与阿里云相继宣布上线 OpenClaw云端极简部署及全套云服务。国内云厂商优刻得、首都在线等也已上线该服务。截至2026年3月8日，OpenClaw在GitHub上的星标数量达27.9万，成为增长速度最快的开源AI项目之一，并超越了如Linux和React等众多超大型基建项目，登顶GitHub开源星标榜。我们认为，OpenClaw在“云端大脑+本地操作”的混合模式，既创造了海量的API Token消耗需求，又为Agent赋能产业打开了巨大的想象空间。未来，能将AI的能力与行业Know-how深度结合的解决方案提供商，将拥有更高的商业价值。

图4：OpenClaw架构图解



资料来源：CoworkAI，华龙证券研究所

图5：OpenClaw登顶GitHub开源软件星标榜



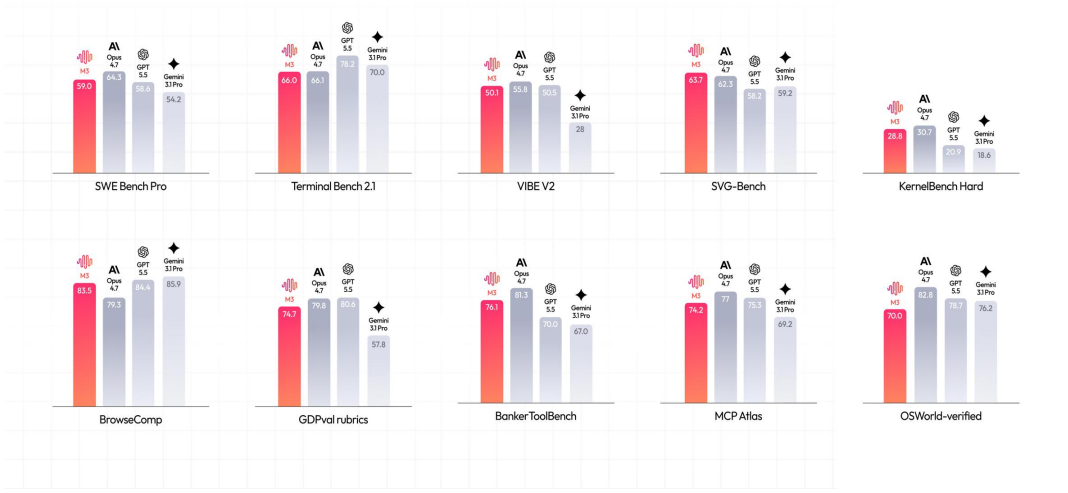
资料来源：OpenClaw，新智元，华龙证券研究所

Agent与多模态支撑token攀升

多模态成为新的竞技场

- 字节：2026年2月，第三代AI视频生成模型Seedance2.0，支持最长15秒视频时长，新增多模态输入；多镜头叙事、音画同步、角色一致性等核心能力领先于全球主流竞品。
- 快手：2026年2月5日，快手上线可灵3.0模型，同样支持最长 15 秒的视频时长及全模态输入输出，在一致性、照片级输出等方面进行了全新升级。
- MiniMax：2026年6月1日，MiniMax M3正式发布。作为原生多模态模型，M3在多模态测试集OmniDocBench上，得分超过Gemini3.1 Pro。

图6：MiniMax M3测试集得分



资料来源：MiniMax，华龙证券研究所

图7：Seedance 2.0 与主流竞品对比（统计时间：2026年2月）

能力维度	Seedance 2.0	Sora 2 (OpenAI)	Veo 3 (Google)	可灵 2.6 (快手)
最大视频时长	15秒 领先	15秒 (Pro版25秒)	8秒	10秒
多模态输入	12个文件 领先	有限支持	文本+图片	文本+图片
多镜头叙事	原生支持 领先	较弱	不支持	不支持
音画同步	原生支持 领先	部分支持	支持	较弱
角色一致性	优秀 领先	良好	一般	良好
物理模拟	优秀 持平	优秀	良好	优秀
画质分辨率	1080p / 2K	1080p	4K	1080p

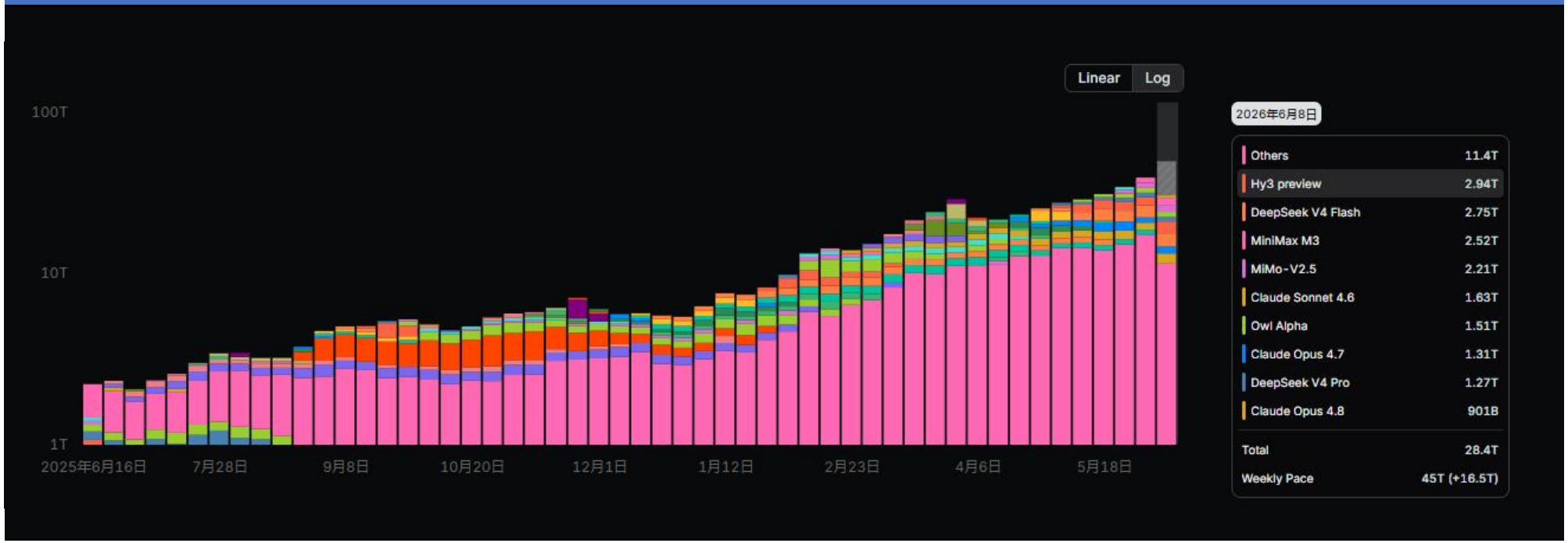
资料来源：即梦，华龙证券研究所

目录

模型作为AI应用的基础，奠定了国产token的竞争力

- 根据OpenRouter的数据，在最新一周（2026年6月2日-6月8日）全球大模型调用量排名前五的模型中，中国占据四席，分别是腾讯Hy3 preview、DeepSeek V4 Flash、Mini Max M3和小米MiMo-V2.5，合计贡献了前五名总调用量的86.47%。
- 我们认为，国产头部AI模型在tokens消耗量上已稳居全球第一梯队，部分国产模型凭借高性价比使得增长速率更加陡峭，更预示着持续的扩张潜力。反映出国产AI模型的发展趋势正从技术追赶转向市场引领，通过优异的产品力和用户体验在全球开源生态中占据重要位置，整个行业的竞争已进入规模化应用和增长效率比拼的新阶段。

图8：OpenRouter全球AI大模型排行榜（2026年6月2日-6月8日周数据）



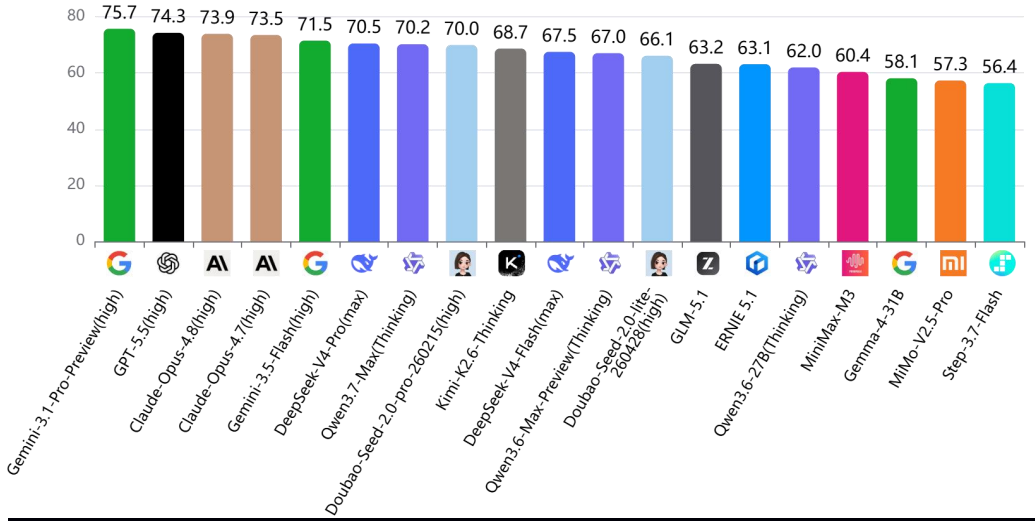
资料来源：OpenRouter，华龙证券研究所

国产token的性价比

高性价比模型+电价优势，看好国产token竞争力提升

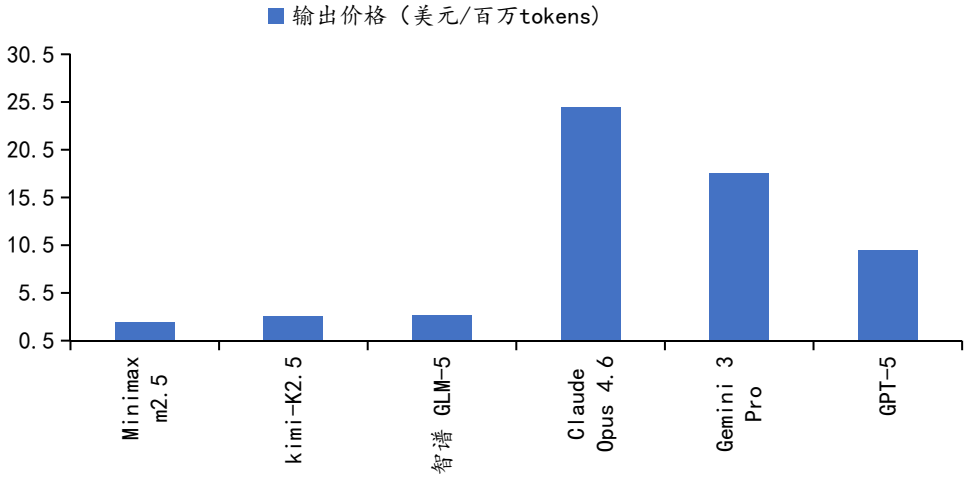
- 高性价比国产模型构成国产token的底层竞争力。根据中文通用大模型评测基准—SuperCLUE于2026年5月的测评结果，在综合评分排名前十中，国产模型占据五席，得分与海外模型接近。
- 中国模型在性能接近甚至部分超越海外主流产品的同时，定价具备压倒性优势。例如，Kimi K2.5 输入定价为每百万Token 4 人民币（缓存未命中）；而美国Claude Opus 4.6输入定价高达每百万Token 5美元。输出成本方面，M2.5 的成本是 Opus、Gemini 3 Pro 和 GPT-5 的十分之一到二十分之一。国产模型API定价具备显著性价比。

图9：中文通用大模型评测基准（SuperCLUE）下各模型排名



资料来源：SuperCLUE ，华龙证券研究所

图10：国内模型API定价具备优势

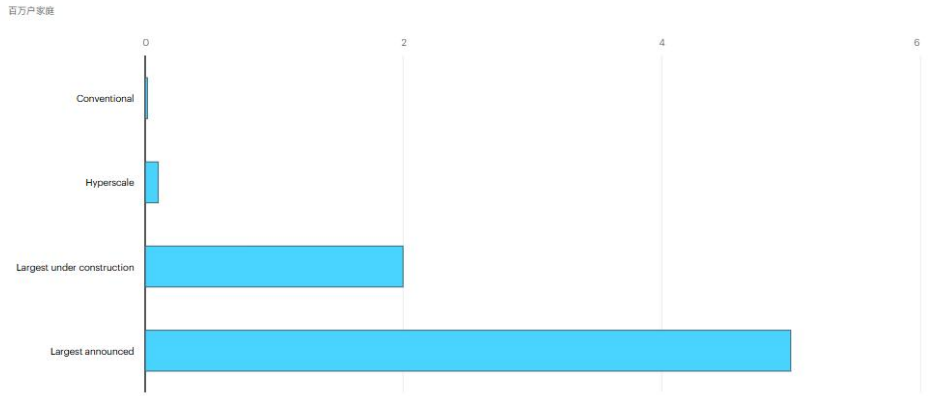


资料来源：Minimax，Claude，Kimi，智谱，Gemini，Cursor IDE，华龙证券研究所

高性价比模型+电价优势，看好国产token竞争力提升

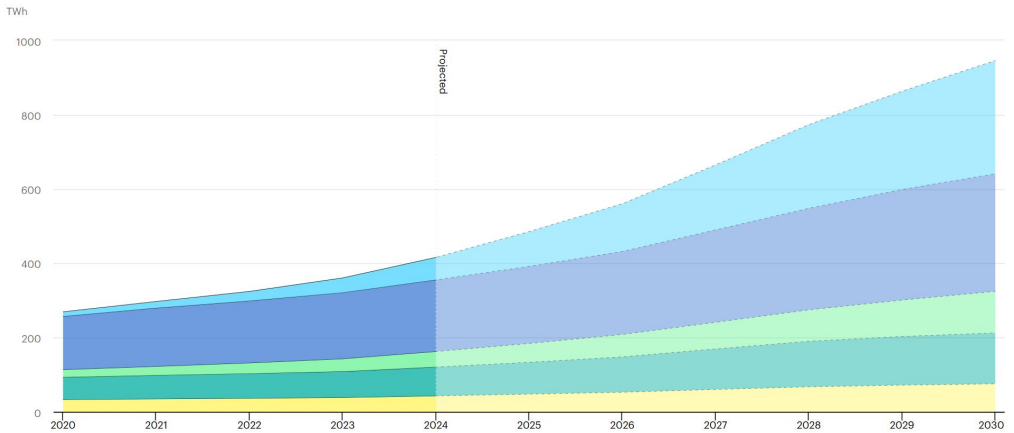
- 人工智能模型的训练和部署主要在数据中心进行。传统数据中心使用 10 兆瓦（MW）到 25 兆瓦（MW）的电力，而超大规模人工智能中心的需求可能超过 100 MW——相当于 10 万户家庭的年电力消耗量。已宣布的最大数据中心预计将消耗相当于 500 万户家庭的电力。2024 年，数据中心占全球电力需求的 1.5%。到 2030 年，根据国际能源署的基本情景预测，这一比例将上升到约 3%，全球数据中心的电力需求将超过两倍，达到约 945 太瓦时（TWh）。
- 国产绿电成本有优势，能够为tokens提供定价上的支持。与北美电力进行对比。由于AI训练和推理都是耗电巨兽，电力成本占运营总成本的60%—70%，一个国家的电网稳定性、电力成本（特别是绿色电力成本）决定了其Token的生产成本竞争力。在能源层面，中国东数西算工程与统一大电网建设使西部绿电价格可以低至0.2元/度，约合0.028美元/度，而欧美电力价格普遍在0.08-0.12美元/度区间。

图11：2024年数据中心耗电量（最后更新时间：2025年3月）



资料来源：IEA，华龙证券研究所

图12：2020-2030年全球数据中心耗电量预测（最后更新时间：2025年4月）



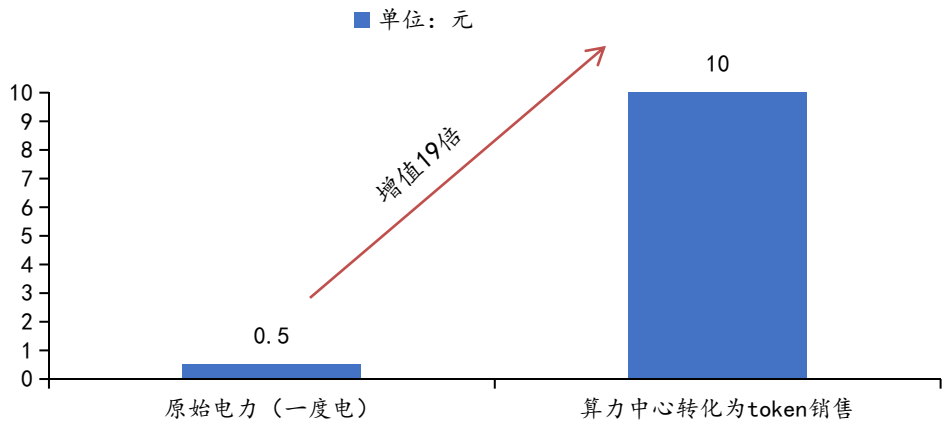
资料来源：IEA，华龙证券研究所

国产token的性价比

上游cpu、存储开启涨价，算力通胀向下传导

- Token的价值链涵盖硬件制造、基础设施建设、算力提供、平台运营、应用开发五大环节。在数据中心的运营成本中，电力及算力折旧成本合计占比可达70%左右，成为决定Token国际竞争力的底层要素。“每瓦特Token吞吐量”（Tokens per Watt）成为衡量AI企业竞争力的核心指标。这意味着在固定的电力预算下，谁能以更高的能源效率生产更多Token，谁就拥有最低的生产成本和最强的市场竞争力。
- 从电费角度看，中国工业用电0.4-0.6元/度，美国0.8-1.2元/度，叠加国内电力输送和响应速度优势，国产token具备显著的电价优势支撑。
- 未来，由于底层支持不同（电力、算力），按交互速度和使用场景划分，token经济将成为一套类似于电力的定价系统。根据黄仁勋在GTC 2026上给出的token定价框架，token分成了五个价格区间：免费层（高吞吐、低交互速度，靠广告变现）、中级层（每百万token 3美元）、高级层（每百万token 6美元）、高速层（每百万token 45美元）到超高速层（每百万token 150美元）。

图13：电力通过算力中心转化为toekn的价值增值过程



资料来源：央视财经，华龙证券研究所

表2：黄仁勋于GTC2026上提出的token定价框架

层级	定价
免费层	提供高吞吐量但交互速度较低的服务，主要通过广告变现
中级层	每百万Token收费3美元
高级层	每百万Token收费6美元
高速层	每百万Token收费45美元
超高速层	每百万Token收费150美元

资料来源：英伟达，华龙证券研究所

目录

全球资本开支加速，带动AI基建加速扩张

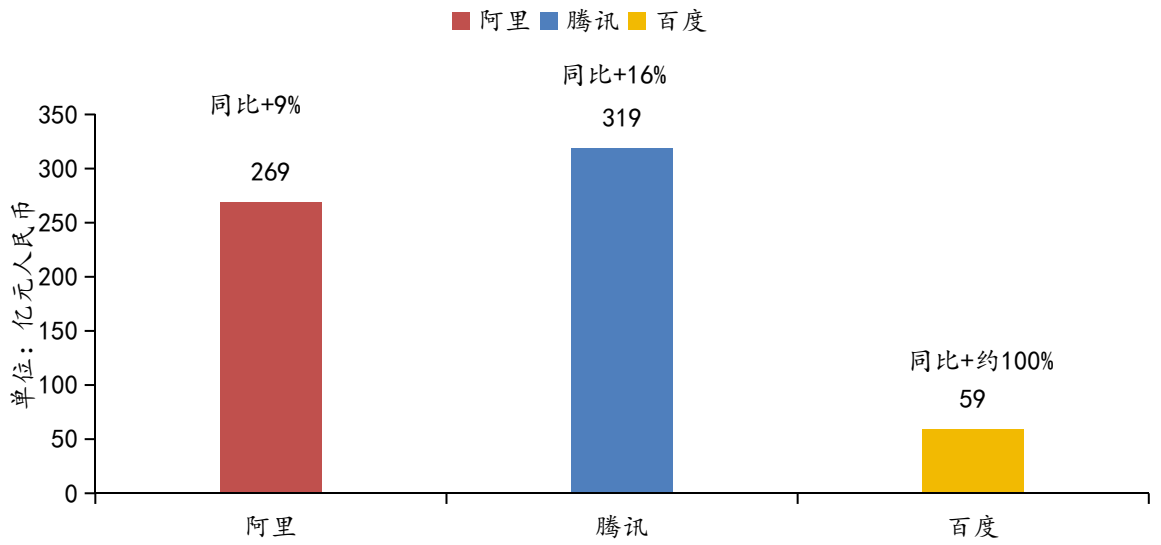
- 从数据来看，北美厂商在2026年第一季度的资本开支呈现出强劲的增长态势。微软、Meta、谷歌资本开支同比增幅均在45%以上，其中谷歌资本开支同比增幅高达107%。海外云厂投资主要集中于人工智能芯片、数据中心、服务器和网络基础设施，且预计2026年的投入规模将进一步显著扩大，显示出对AI等前沿技术持续扩张的坚定信心。
- 与此同时，中国主要的科技企业同样在加大资本投入。其中，截至一季度末，腾讯用于AI相关投入的资本开支付款为370亿元，计入当期的资本开支为319亿元，同比增长16%，环比大增63%，创近年来单季新高。此外，一季度阿里集团资本支出269亿元，同比增长9.24%，未来五年资本支出或将远超3800亿元；百度资本支出59亿元，同比增约100%，环比约达上季度的3倍。字节跳动称今年计划资本支出将超过2000亿元，较此前口径增加了25%。国内科技企业持续的投入态势与北美云厂商相呼应，共同指向全球科技行业在云计算和人工智能基础设施领域正在进行一场高强度的“军备竞赛”。
- 我们认为，当前全球科技巨头的资本开支正处于一个高强度、高增长的周期，这一轮针对AI等核心技术的投资热潮在未来几年仍将持续。

表3：北美云厂最新资本开支及指引

公司	Capex（自然年2026Q1）	用途	指引
微软	319亿美元（同比+49%）	约三分之二投向以GPU、CPU为核心的短周期算力资产	2026财年资本支出约为1900亿美元
Meta	198亿美元（同比+45%）	零部件涨价与数据中心建设成本上升驱动资本开支增加；并且加大资源从传统业务调配至AI基础设施的比例。	2026年资本支出上限从1350亿美元提高至1450亿美元
谷歌	357亿美元（同比+107%）	长期投入AI	2026年全年资本支出指引从1750亿至1850亿美元，上调至1800亿至1900亿美元；2027年资本支出预计将大幅增加

数据来源：Meta，新华财经，财联社，华龙证券研究所

图14:阿里、腾讯、百度资本开支情况

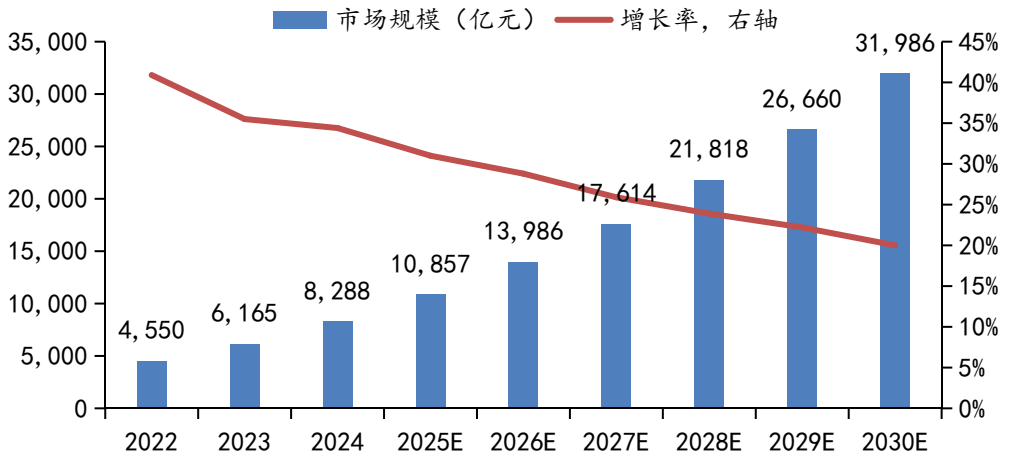


数据来源：各公司财报，华龙证券研究所

AI带来的智算需求推动云服务市场结构性增长

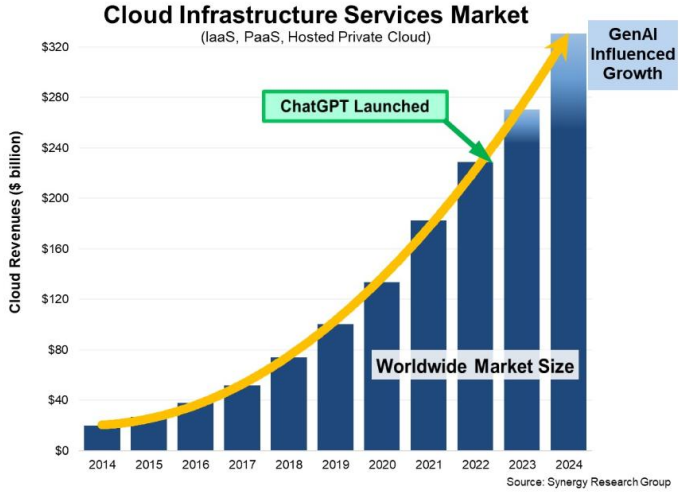
- **受人工智能等技术创新推动，我国云计算市场保持较高速增长。**据中国信通院统计数据，2024年我国云计算市场规模达到8288亿元（同比增长34.4%），延续2022-2023年较高速增长态势，预计2025年我国云计算市场规模达10857亿元人民币（预计同比增速达31%），有望首次突破万亿市场规模。人工智能等新兴技术将与云计算进一步融合，推动云计算市场拓展。到2030年，我国云计算市场规模有望突破3万亿元。
- **AI对云服务的拉动作用已得到海外先验。**根据Synergy Research Group的数据，2024年全球云基础设施服务市场增长了22%，达到3300亿美元。通过结合新的 GenAI 平台服务、GPU 即服务以及对各种其他云服务的增强，生成式AI至少贡献了云服务收入增长的一半。

图15:中国云计算市场规模



数据来源：中国信通院，华龙证券研究所

图16:生成式AI对云服务市场增长的拉动作用（统计日期：2025年2月）

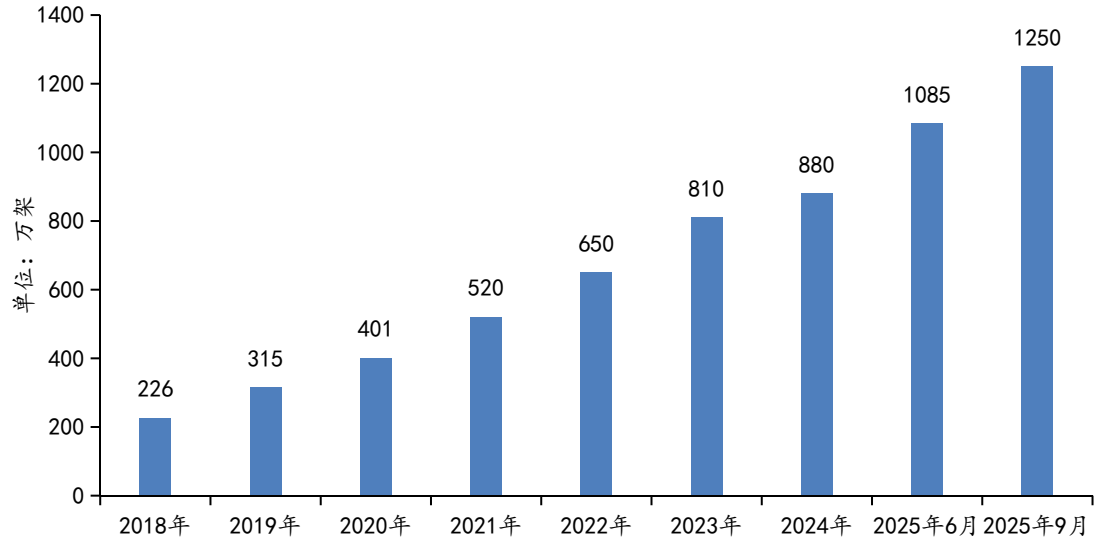


数据来源：Synergy Research Group，华龙证券研究所

AI推动智算规模增长，云服务需求提升

- **IDC向AIDC加速演进：**2025年行业呈现爆发式增长态势，仅第三季度机架规模便从1085万架跃升至1250万架，反映出AI算力需求正成为驱动行业增长的新引擎。与此同时，全国算力中心平均PUE已降至1.42，绿色化、集约化、高密化转型加速推进，IDC行业正从传统数据中心向智算中心（AIDC）演进，开启新一轮高质量发展周期。
- **算力租赁快速落地：**二季度以来，多家上市公司近期密集跨界布局算力赛道：东阳光两度公告合计约260亿–310亿元算力订单，晶科科技拟投245亿元建设宁夏中卫1GW算力中心，华策影视拟斥资不超33亿元采购服务器开展云算力服务，凸显算力租赁商业化落地正加速推进。

图17:2018-2025年中国IDC行业机架规模



资料来源：工信部，信息通信发展司，前瞻产业研究院，华龙证券研究所

表4:2026年二季度部分A股上市公司算力租赁相关布局公告

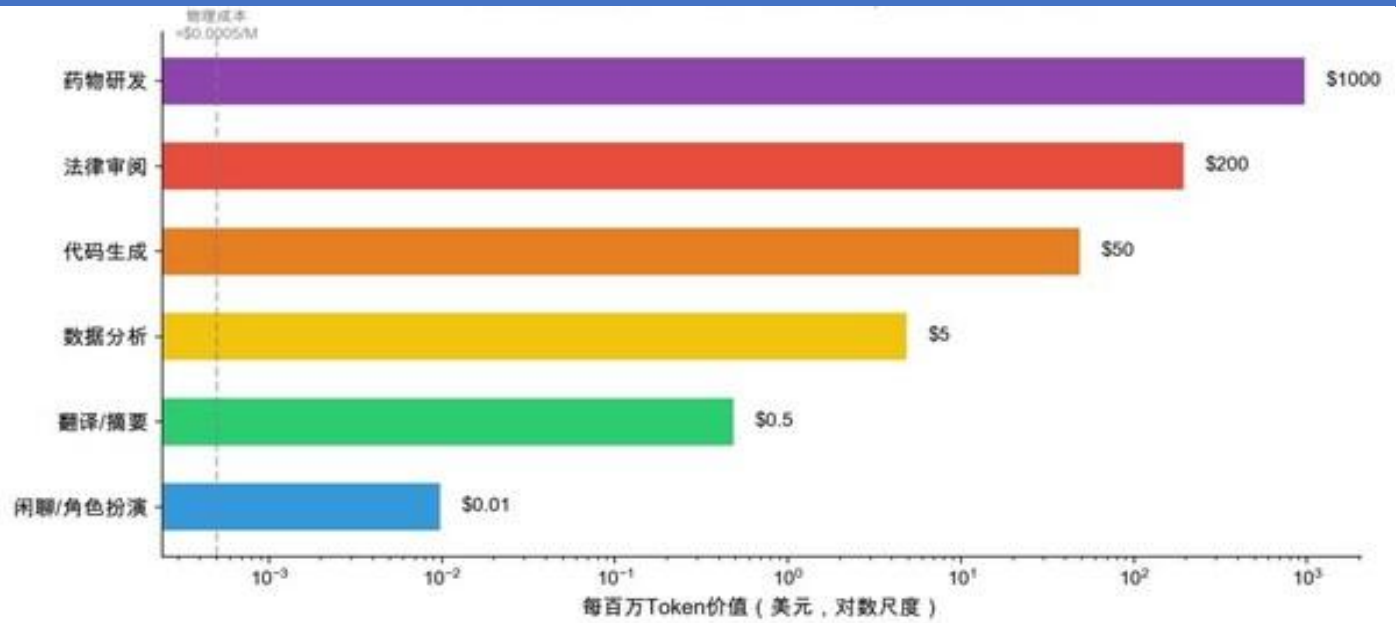
公司	公告时间	公告内容
晶科科技	2026年4月	与中卫市人民政府签署投资协议，计划总投资约245亿元建设宁夏中卫1GW算力中心项目，部署机柜约5万架。
华策影视	2026年5月	披露资产采购计划，拟投资不超过33亿元采购服务器提供云算力服务，对应算力规模2.9万P，服务期5年。
东阳光	2026年5月	控股子公司东阳光云智算签署了一份总金额区间预计为160亿元至190亿元的算力服务框架合同，期限为五年。
	2026年6月	控股子公司东阳光云智算签署合同，预计总金额100亿至120亿元，期限5年，负责采购并部署高性能算力服务器以租赁方式交付算力资源。

资料来源：上海证券报，Wind，华龙证券研究所

新业态：Token工厂

- “Token工厂”可以理解为将电力、GPU、冷却、网络和推理优化技术进行打包，规模化、工业化地生产大模型可消费 Token 的新型智算基础设施。其核心功能是将电力资源转化为大模型可直接消费的Token输出，区别于传统数据中心仅提供物理空间与硬件托管，这类设施聚焦于智能产出的效率与稳定性。商业模式上，Token工厂突破了传统算力租赁按机柜或时长收费的资源售卖逻辑，转向以Token产量为计量单位的服务收费模式，收入来源包括基础算力服务费、按Token调用量计费、API接口服务费及基于应用效果的收益分成，运营核心在于通过提升单卡日均有效Token产能、缓存命中率及推理框架效率，最大化每度电的Token产出比，从而降低单位成本并提升利润空间。
- 产业层面上，弘信电子已在无锡落地基于昇腾架构的超节点Token工厂，直接将算力集群与推理优化打包交付并按Token计费。

图18:各垂类Token的价值光谱



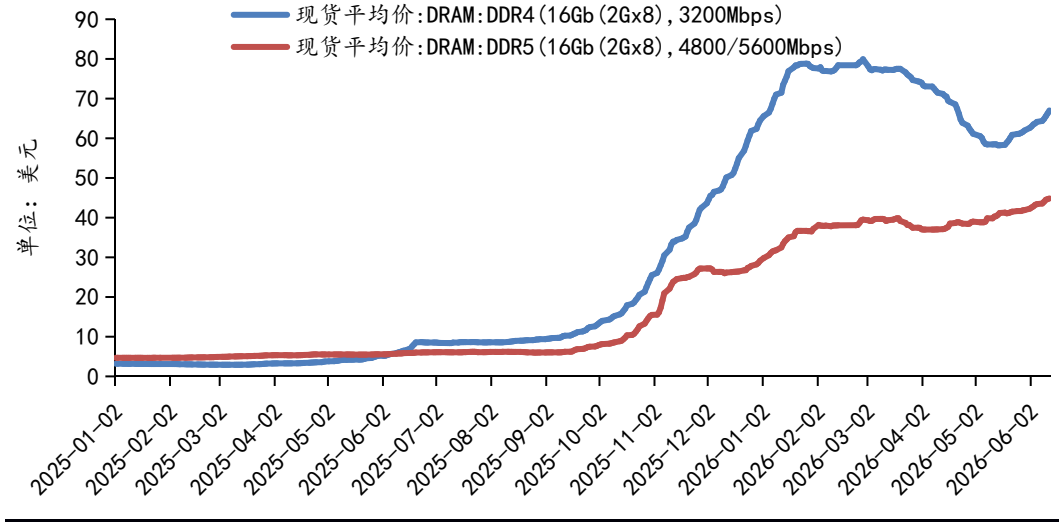
资料来源：CXO UNION，Litowitz et al. (2026)；Bergemann et al. (2025)，华龙证券研究所

目录

上游开启涨价潮

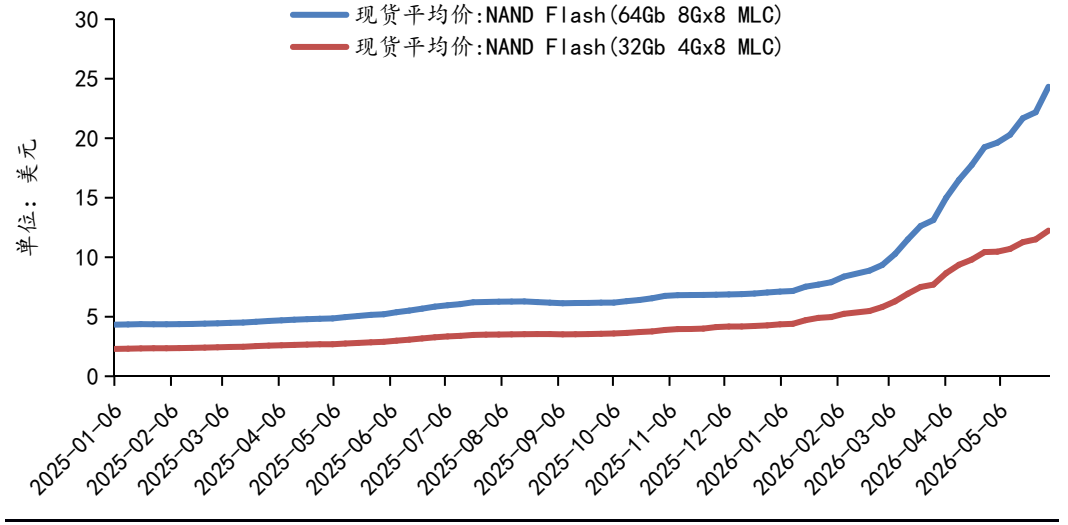
- 上游算力硬件开启涨价潮，本轮涨价的核心驱动因素是AI算力需求的爆发。全球AI大模型和数据中心的加速扩张，导致供需失衡。
- **CPU：**英特尔和AMD 已分别于2026年3月底和4月起，通知客户将上调全系列CPU产品价格。此次涨价平均幅度在10%至15% 之间，部分产品涨幅更高。与此同时，CPU的交货周期也从过去的一至两周，大幅拉长至平均八至十二周，个别情况下甚至需要等待长达六个月。
- **芯片代工：**德州仪器宣布将于2026年4月1日启动近一年内的第三轮调价，涨价幅度最高达85%。与此同时，中芯国际、世界先进、华虹、力积电、晶合集成等芯片代工巨头纷纷确认或计划上调代工价格，涨幅普遍在5%至20%之间。
- **存储：**三星电子、SK海力士等内存大厂已于2025年开始上调存储芯片价格。自2025年3月至2026年5月，消费级DRAM 16GB DDR4价格从约200元暴涨至2000元，涨幅高达约900%，16GB DDR5涨幅达300%；NAND闪存方面，256GB和512GB产品价格普遍上涨了200%至250%。此外，服务器级256GB内存单条价格达到约4万元，而手机存储（12GB+512GB）成本也从约400元涨至近2000元，涨幅约为400%。

图19：DRAM现货涨价情况（部分）



资料来源：Wind，华龙证券研究所

图20：NAND现货涨价情况（部分）



资料来源：Wind，华龙证券研究所

算力通胀继续传导

- **云服务：**2026年一季度已经进入涨价实证期。其中，海外云先涨、国内中小云跟涨、随后头部云厂正式调价，模型厂商同步提价。本轮涨价涉及GPU云服务、存储、数据传输、模型调用以及安全类产品，主要是因为AI硬件成本上行叠加需求快速增加。
- **算力租赁：**大模型从训练走向推理，算力需求将进入7x24小时持续消耗。高端机型供需仍处于不平衡状态，预计未来1-2年仍有较强价格支撑。

表5：云服务/模型涨价情况梳理

服务类型	涨价区间	涨价厂商和幅度
AI算力/GPU云服务器	普遍上调5%—30%，部分紧缺算力涨幅更高	金山云GPU云服务器（GPU高效计算型P6V）上调15-50%；阿里云平头哥真武810E等算力卡产品价格上调5%至34%；百度智能云AI算力相关产品服务价格上调5%至30%；腾讯云GPU云服务器价格上调5%；AWS EC2机器学习容量块价格上调15%；
大模型API/套餐订阅	部分模型API涨价	腾讯云混元HY2.0 Instruct输入价由0.0008元/千Tokens涨至0.004505元/千Tokens，涨幅约463%；智谱GLM Coding Plan订阅套餐价格整体上调30%起
存储/高性能文件系统	AI相关存储产品上调约30%	金山云KPFS文件存储部分上调30%-50%；阿里云CPFS智算版上调30%；百度智能云并行文件存储等上调30%；网宿科技宣布对象存储产品上调40%
数据传输/CDN	部分地区翻倍	谷歌云北美数据传输由0.04美元/GiB涨至0.08美元/GiB，涨幅100%；欧洲涨60%、亚洲涨约42%；网宿科技对部分CDN产品上调35%—40%
DDoS/安全网络产品	安全防御类产品同步上调	阿里云DDoS高防弹性95价格由100元/Mbps/月上调至150元/Mbps/月

资料来源：金山云，网宿科技，财联社，中国工信网，证券日报网，澎湃新闻，华龙证券研究所

目录

- **Agent与多模态支撑token攀升。**随着AI Agent处理任务日趋复杂，其推理深度与调用链路不断延伸，将驱动底层Token消耗呈数量级跃升。IDC预计，活跃Agent的数量将从2025年的约2860万，快速攀升至2030年的22.16亿。这意味着五年后，能够帮助企业执行任务的数字劳动力数量将是今天的近80倍，年复合增长率139%。Agent真正干活的频率增长得更快，年执行任务数将从2025年的440亿次暴涨至2030年的415万亿次，年复合增长率高达524%。数据显示，年度Token消耗量预计将从2025年的0.0005 Peta Tokens激增至2030年的152667 Peta Tokens，年复合增长率高达3418%。此外，多模态成为新的竞技场。字节于2026年2月发布第三代AI视频生成模型Seedance2.0，支持最长15秒视频时长，新增多模态输入；多镜头叙事、音画同步、角色一致性等核心能力领先于全球主流竞品。快手、Minimax等亦快速跟进新一代多模态模型，有望进一步支撑token消耗量增长。
- **国产token的性价比凸显。**根据OpenRouter的数据，在最新一周（2026年6月2日-6月8日）全球大模型调用量排名前五的模型中，中国占据四席，合计贡献了前五名总调用量的86.47%。我们认为，国产头部AI模型在tokens消耗量上已稳居全球第一梯队，部分国产模型凭借高性价比使得增长速率更加陡峭，更预示着持续的扩张潜力。“每瓦特Token吞吐量”（Tokens per Watt）成为衡量AI企业竞争力的核心指标。这意味着在固定的电力预算下，谁能以更高的能源效率生产更多Token，谁就拥有最低的生产成本和最强的市场竞争力。从电费角度看，中国工业用电0.4-0.6元/度，美国0.8-1.2元/度，叠加国内电力输送和响应速度优势，国产token具备显著的电价优势支撑。
- **Token经济的涨价传导。**上游算力硬件开启涨价潮，本轮涨价的核心驱动因素是AI算力需求的爆发。全球AI大模型和数据中心的加速扩张，导致供需失衡。CPU：英特尔和AMD 已分别于2026年3月底和4月起，通知客户将上调全系列CPU产品价格。此次涨价平均幅度在10%至15% 之间，部分产品涨幅更高。芯片代工：德州仪器宣布将于2026年4月1日启动近一年内的第三轮调价，涨价幅度最高达85%。与此同时，中芯国际、世界先进、华虹、力积电、晶合集成等芯片代工巨头纷纷确认或计划上调代工价格，涨幅普遍在5%至20%之间。存储：三星电子、SK海力士等内存大厂已于2025年开始上调存储芯片价格。自2025年3月至2026年5月，消费级DRAM 16GB DDR4价格从约200元暴涨至2000元，涨幅高达约900%，16GB DDR5涨幅达300%；NAND闪存方面，256GB和512GB产品价格普遍上涨了200%至250%。云服务：2026年一季度已经进入涨价实证期。其中，海外云先涨、国内中小云跟涨、随后头部云厂正式调价，模型厂商同步提价。算力租赁：大模型从训练走向推理，算力需求将进入7x24小时持续消耗。高端机型供需仍处于不平衡状态，预计未来1-2年仍有较强价格支撑。
- **投资建议：**AI Agent与多模态爆发正驱动Token消耗呈指数级激增，国产模型凭借极高性价比与电价优势占据全球流量主导，催生巨大算力需求；高端算力供需失衡推动上游算力硬件至下游云服务开启涨价潮，Token经济正值风起之时。维持计算机行业**“推荐”**评级，建议关注：（1）国产芯片：寒武纪（688256.SH）、海光信息（688041.SH）、中国长城（000066.SZ）；（2）云厂商及IDC：润泽科技（300442.SZ）、润建股份（002929.SZ）、优刻得-W（688158.SH）、首都在线（300846.SZ）、大位科技（600589.SH）、网宿科技（300017.SZ）；（3）算力租赁：协创数据（300857.SZ）、宏景科技（301396.SZ）。

表6：重点关注公司及盈利预测

股票代码	股票简称	2026/06/12	EPS（元）				PE				投资评级
		股价（元）	2025A	2026E	2027E	2028E	2025A	2026E	2027E	2028E	
000066.SZ	中国长城	15.51	-0.02	0.05	0.10	0.27	/	343.9	150.6	57.3	未评级
002929.SZ	润建股份	58.63	0.14	1.32	2.20	3.55	284.4	44.6	26.7	16.5	未评级
300017.SZ	网宿科技	14.31	0.33	0.36	0.45	0.58	37.4	39.8	31.8	24.7	增持
300442.SZ	润泽科技	73.53	3.00	2.03	2.55	3.17	17.6	36.3	28.8	23.2	未评级
300846.SZ	首都在线	21.27	-0.34	-0.004	0.27	0.11	/	/	78.8	195.0	增持
300857.SZ	协创数据	223.70	3.38	4.95	7.42	10.99	49.9	45.2	30.2	20.4	未评级
301396.SZ	宏景科技	172.56	0.17	2.01	4.13	9.95	385.5	85.9	41.8	17.3	未评级
600589.SH	大位科技	9.35	-0.01	0.07	0.09	0.13	/	132.4	101.5	70.4	未评级
688041.SH	海光信息	280.00	1.10	2.00	2.74	3.88	204.0	140.0	102.2	72.1	增持
688158.SH	优刻得-W	35.00	-0.16	0.11	0.31	0.53	/	318.2	112.9	66.0	增持
688256.SH	寒武纪	1240.00	4.93	11.00	17.19	27.01	275.0	112.7	72.1	45.9	增持

数据来源：Wind，华龙证券研究所，注：网宿科技（300017.SZ）、优刻得-W（688158.SH）盈利预测来源于华龙证券研究所；首都在线（300846.SZ）、海光信息（688041.SH）、寒武纪（688256.SH）2026-2027年盈利预测来源于华龙证券研究所，2028年盈利预测来源于Wind一致预期；其余所有公司盈利预测来源于Wind一致预期。

目录

（1）所引用数据资料的误差风险。本报告数据资料来源于公开数据，将可能对分析结果造成影响。

（2）AI投资力度不及预期。相关技术突破与投资力度关系紧密。

（3）AI产品竞争加剧。竞争加剧可导致价格战。

（4）重点关注公司业绩不达预期。重点关注公司业绩会受到各种因素影响，如果业绩不达预期，会使得公司股价受到影响。

（5）政策标准出台速度不及预期。AI持续发展需政策引导。

免责及评级说明部分

分析师声明：

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉尽责的职业态度，独立、客观、公正地出具本报告。不受本公司相关业务部门、证券发行人士、上市公司、基金管理公司、资产管理公司等利益相关者的干涉和影响。本报告清晰准确地反映了本人的研究观点。本人在预测证券品种的走势或对投资证券的可行性提出建议时，已按要求进行相应的信息披露，在自己所知情的范围内本公司、本人以及财产上的利害关系人与所评价或推荐的证券不存在利害关系。本人不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。据此入市，风险自担。

投资评级说明：

投资建议的评级标准	类别	评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后的6-12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅。其中：A股市场以沪深300指数为基准。	股票评级	买入	股票价格变动相对沪深 300 指数涨幅在 10%以上
		增持	股票价格变动相对沪深 300 指数涨幅在 5%至 10%之间
		中性	股票价格变动相对沪深 300 指数涨跌幅在-5%至 5%之间
		减持	股票价格变动相对沪深 300 指数跌幅在-10%至-5%之间
		卖出	股票价格变动相对沪深 300 指数跌幅在-10%以上
	行业评级	推荐	基本面向好，行业指数领先沪深 300 指数
		中性	基本面稳定，行业指数跟随沪深 300 指数
		回避	基本面向淡，行业指数落后沪深 300 指数

免责声明：

华龙证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。

本报告的风险等级评定为R4，仅供符合本公司投资者适当性管理要求的客户（C4及以上风险等级）参考使用。本公司不会因为任何机构或个人接收到报告而视其为当然客户。

本报告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。本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，本公司可能会发出与本报告所载意见、评估及预测不一致的研究报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。

本报告仅为参考之用，并不构成对具体证券或金融工具在具体价位、具体时点、具体市场表现的投资建议，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。据此投资所造成的任何一切后果或损失，本公司及相关研究人员均不承担任何形式的法律责任。

在法律许可的情况下，本公司及所属关联机构可能会持有报告中提及的公司所发行证券的头寸并进行证券交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

版权声明：

华龙证券研究所

北京
地址：北京市东城区安定门外大街189号天鸿宝景大厦西配楼F4层
邮编：100033

兰州
地址：兰州市城关区东岗西路638号文化大厦21楼
邮编：730030
电话：0931-4635761

上海
地址：上海市浦东新区浦东大道720号11楼
邮编：200000

深圳
地址：深圳市福田区民田路178号华融大厦辅楼2层
邮编：518046