

| 证券研究报告 |

在“卡瓦格博之心”持续瞭望顶峰—— 解析Token经济

2026.06.12

分析师：苏仪
执业证书编号：S0740520060001

分析师：刘一哲
执业证书编号：S0740525030001

注：“卡瓦格博”系梅里雪山主峰、藏区八大神山之首，至今无人登顶；“卡瓦格博之心”为环绕该主峰东南坡、可正面仰望峰顶的徒步环线名称。本报告以此为题，喻指对token的探讨是人工智能研究中最核心、最值得抵达却尚未被完全征服的腹地。

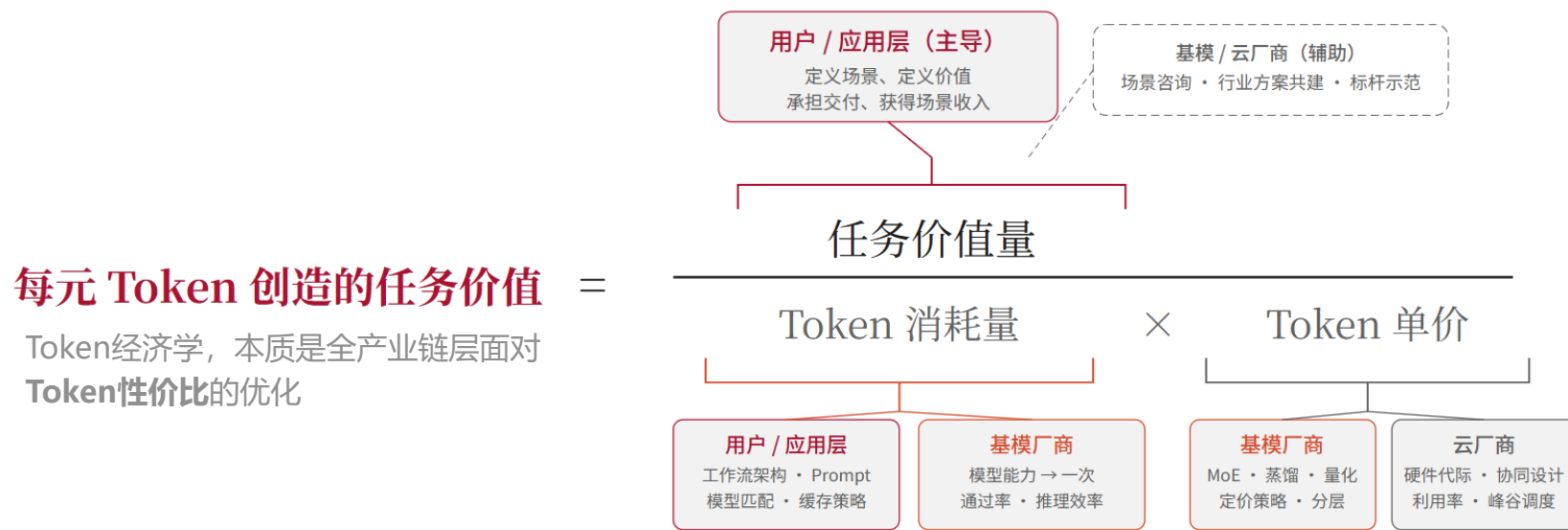
摘要

- **Token 经济本质上是一场关于性价比的全产业链优化运动。**这场优化可以用一个公式统一表达：**每元 Token 创造的任务价值 = 任务价值量 ÷ (Token 消耗量 × Token 单价)**。Token 作为智能时代的基本度量单位，其供需的非同质化属性催生运营层。Token 可类比电力时代的 kWh，但并非同质商品；其供给侧按 Input/Output/Cache 分层定价，需求侧按任务价值与复杂度匹配不同智能的模型，供给端“按量计价”与下游交付端“按结果/项目/订阅付费”之间存在错配。正是这一缺口，催生了以套餐化、路由聚合、效果打包等方式承接用量不确定性的 Token 运营层。
- **杰文斯悖论下 Token 需求爆发，新应用场景持续解锁。**单位 Token 成本的快速下降非但没有压缩总开支，反而随着模型能力提升解锁更多应用场景、把总消耗推向更高量级。下游需求放量沿场景梯次展开：**Coding**率先爆发，**Agent**因上下文重读与多轮工具调用产生大量“隐形词元”，子 Agent、Agent 集群进一步放大消耗，叠加多模态物理 AI 数字孪生、智能驾驶、具身机器人等全新场景，共同构成 Token 需求的长期增长曲线，应用层价值量占比有望逐步提升。
- **算力是核心受益环节，数据中心正变为“Token 工厂”。**当 Token 成为可计量、边际成本递减的标准化产出，数据中心便以类似制造业的逻辑运转为“Token 工厂”，竞争最终落到单位 Token 成本与算力约束上。随着互联网厂商资本开支持续攀升，训练转向推理利好国产芯片渗透，政策加码国产数据中心建设，Token 供给持续爆发；算力租赁则有望盘活存量资源，并向算力调度、一体化算力网与整体 AI 解决方案等高附加值运营服务升级。
- **投资建议：**建议关注基础模型层智谱、MiniMax、商汤、科大讯飞等；算力层浪潮信息、紫光股份、首都在线、东方国信等；应用层彩讯股份、万兴科技等。
- **风险提示：**存在 AI 等底层技术变革不及预期，下游客户 IT 支出意愿与力度不及预期，政策落地不及预期，行业竞争加剧，第三方数据失真，市场规模测算偏差，研报信息更新不及时等风险。

Token经济学可以理解为全产业链优化Token性价比的活动

- Token 经济本质上是一场关于性价比的全产业链优化运动。这场优化可以用一个公式统一表达：每元 Token 创造的任务价值 = 任务价值量 ÷ (Token 消耗量 × Token 单价)。
- 公式中的分子和分母分别由产业链不同的参与者进行优化。应用层是唯一同时触达分子和分母的角色，他们定义场景（决定分子大小），也选择 workflow 架构和模型档位（决定分母中的消耗量）；基模厂商主要关注分母中的 Token 单价（通过 MoE、蒸馏、推理加速等技术路径降本）和消耗量（通过提升模型能力减少用户重试）；云厂商主攻分母最底层的物理成本（硬件代际、协同设计、利用率管理）。三方各有侧重，但分母优化的红利往往会通过价格竞争流向应用层。

图表：Token经济学的优化目标与参与方



资料来源：中泰证券研究所



1

拆解Token产业：从价值创造与分配的角度理解token

Token：智能时代的基本度量单位，可类比电力时代的kWh

- **Token是AI时代对“智能”的基本度量单位。** 类比此前的技术革命浪潮，正如 kWh 衡量电能、Mbps 衡量带宽，Token 正在成为 AI 智能产出的通用计量标准。Token 与电力、带宽、云计算有类似的商品属性：短期供给弹性低（大型数据中心建设周期往往在1年以上，类似电厂扩容）、底层依赖重资产基础设施等。这些共性意味着Token经济的演进可能与电力、互联网与云计算的历史节奏相似。但同时，Token的定价由服务商单方设定，且需求弹性随使用场景有较大变化，这种异质需求结构使得Token 的价值分配远比kWh复杂。

图表：历史通用技术浪潮中可以与token形成对比的要素

	电力	带宽	云计算资源	AI Tokens
标准化程度	高（单位：千瓦时 kWh）	中等（单位：兆比特每秒 Mbps）	中等（单位：实例 / 小时）	高（单位：百万代币）
价格波动性	极高	中等	中 - 高	当前较低，预期未来较高
供给弹性（短期）	低	中等	中等	低
需求弹性	低	中等	中 - 高	随使用场景变化
定价机制	市场 + 监管调控	合同约定 + 拥塞定价	按需定价 + 现货市场	服务提供商设定
物理基础 / 底层支撑	发电厂	网络基础设施	数据中心	GPU 计算集群

资料来源：AI Token Futures Market: Commoditization of Compute and Derivatives Contract Design，中泰证券研究所

Token分类：供给侧，基础模型厂商为Input / Output / Cached给予不同定价

- 基模厂商往往按照Token类别计费。以OpenAI为代表，API层面将Token分为四类计量：输入Token（用户输入的提示）、输出Token（模型生成的回复）、缓存Token（复用的上下文，按低费率计费）、以及推理Token（推理CoT，不单独计费）。其中Output Token中很大部分来自推理过程。将其定义为“隐藏的思考Token，计入output但不返回给用户”。Cache命中率、Input/Output比、上下文长度等指标往往影响最终的Token成本。
- 厂商也可对批量处理（Batch Processing）提供折扣，定价策略丰富。批量处理即用户将大量请求打包提交，由平台在算力负载较低的时段异步执行，等待时间可能稍长。这种逻辑类似电力市场的峰谷电价机制，实时调用占用即时算力（峰时），Batch请求填充闲置产能（谷时），厂商通过平滑负载曲线、提升GPU利用率来覆盖折扣成本。

图表：各厂商旗舰模型定价情况（统计时间2026年6月5日）

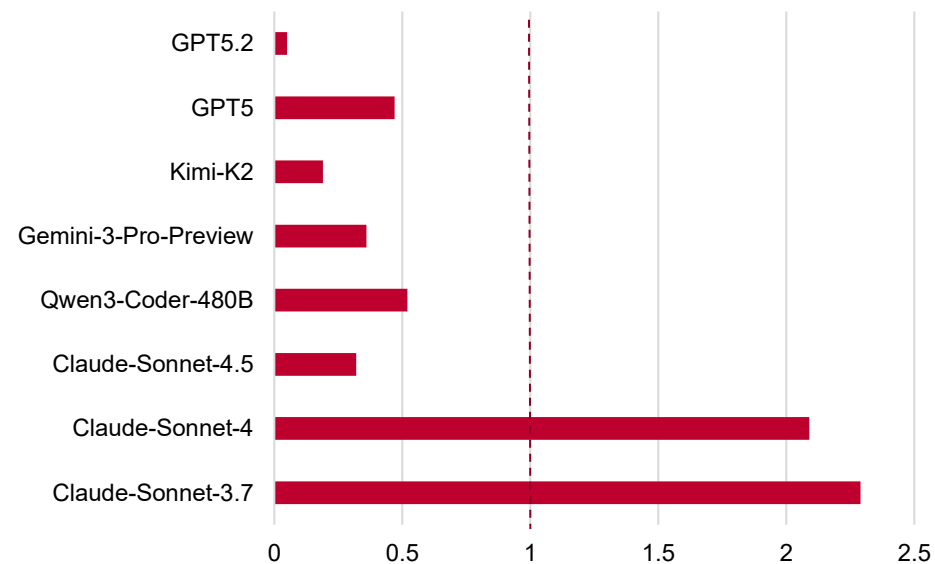
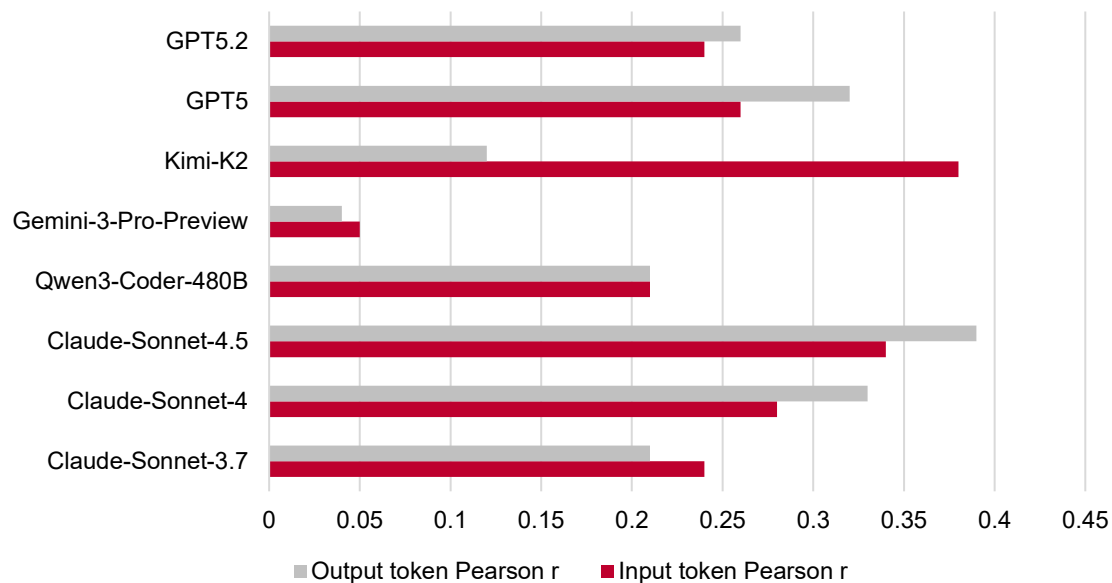
厂商	模型	Input 输入	Output 输出	Cached Input 缓存
Anthropic	Claude Opus 4.8	\$5.00/M	\$25.00/M	\$0.50/M (-90%)
OpenAI	GPT-5.5	\$5.00/M	\$30.00/M	\$0.50/M (-90%)
DeepSeek	V4 Pro	¥3元/M (原价¥12, 永久2.5折)	¥6元/M (原价¥24)	¥0.025元/M
智谱	GLM-5.1 (≤32K)	¥6元/M	¥24元/M	¥1.3元/M
MiniMax	M3 (标准≤512K)	¥2.10元/M (限时五折, 原价¥4.20)	¥8.40元/M (限时五折, 原价¥16.80)	¥0.42元/M
月之暗面	Kimi K2.6	¥6.50元/M	¥27.00元/M	¥1.10元/M
小米	MiMo V2.5 Pro	¥3.00元/M	¥6.00元/M	¥0.025元/M

资料来源：各厂商官网，中泰证券研究所

Token分类：需求侧，基于任务价值量和复杂度需要不同“价值”的token

- **Token并非同质化的商品。**一方面不同技术路径和工程化手段决定了token产出的不同成本，另一方面简单任务、中等任务、复杂任务对应不同智能程度的模型，单位Token的价值差距很大。因此为不同的任务分配合适的、更具性价比的模型也就非常重要。即使是模型自身，也很难精确预测任务的token消耗量。特别是，针对Agent场景中的多轮、复杂任务，反而出现“越贵的Token越省钱”的现象。

图表：模型自身很难准确预测任务的token消耗



资料来源：arXiv:2604.22750, Exponential View Analysis, 中泰证券研究所

下游决策时，任务ROI依旧是最终决策标准

- 我们认为下游企业的AI投入决策方式往往不同，但ROI才是最终标准。硬件采购时往往考虑GPU租赁价和Flops单价进行选型，完全不反映Token产出，多数企业仍采用这种方式进行决策；穿透到Token运营视角，需要逐步建立调用级可观测性、按团队的成本归因和AI预算管控能力；但最终，还是业务部门对投入ROI进行测算。三层穿透构成一条AI成本管理的成熟度阶梯，企业沿阶梯的攀升速度，将直接决定AI支出从成本中心转向利润中心的节奏。

图表：下游采购AI决策视角



资料来源：中泰证券研究所

供给“按量计价”与交付“按结果/项目/订阅付费” 错配催生Token运营需求

- 我们认为Token运营并非凭空出现的环节，而是由于成本端按量计价，交付端按结果/项目付费引发的错配所致。具体而言：1) B端：用户端AI应用的ROI难以测算，终端客户往往需要承担不确定性；特别是国内SaaS付费基础不深，因此往往采用项目制/场景交付，对于任务场景的token消耗难以预估和明确测算。2) C端：付费模式以订阅制为主（包月/包年）或免费增值，成本端依旧按token消耗结算，高价值用户订阅需求难以满足。

图表：理解Token运营存在的原因



资料来源：中泰证券研究所

Token运营的几类模式

- **Token运营本质是在转移不确定性：谁吸收 Token 用量波动的风险，谁就可以分配价值。**按量计价可能面临用户暴露问题，按结果付费可能面临服务商暴露问题。这种矛盾创造空白，吸引部分运营层入场。基模厂商的API定价从按量后付费到企业定制方法，运营层玩家通过套餐化、路由聚合、效果打包和行业封装，以各自的方式承接用量不确定性并将其转化为定价权。

图表：Token运营模式示意

底座层 —— API 定价模式光谱（风险由用户侧向平台侧转移）



运营层 —— 四种 Token 再封装模式



资料来源：PROFESSIONAL STRATEGY CONSULTING GROUP，中泰证券研究所

Token的含义逐渐丰富：成本、研发费用还是资产？

- 围绕 Token 消耗产生的支出，在会计上正在出现多重身份并存的现象。不同使用场景下，同样一笔 Token 支出可能进入完全不同的会计科目。用于推理交付可能是COGS，用于训练微调可能是R&D，用于客户获取可能是SGA。对基模厂商而言，API 按 Token 计费使其成为营业收入的计量基础，而训练、微调、评测消耗的 Token 则构成研发投入；对云厂商和运营商而言，Token 是算力服务收入的载体，对应 GPU、电力、网络等基础设施成本与折旧；对应用层和企业用户而言，调用产生的 Token 支出计入直接成本或获客费用，预购的 Token 额度类似原材料库存形成预付资产，而 KV Cache、知识库、Agent 记忆等沉淀甚至可确认为可摊销的无形资产。Token 正从单纯的成本项，演变为贯穿收入、费用与资产的核算单元。

图表：Token成本在不同玩家处呈现不同的角色

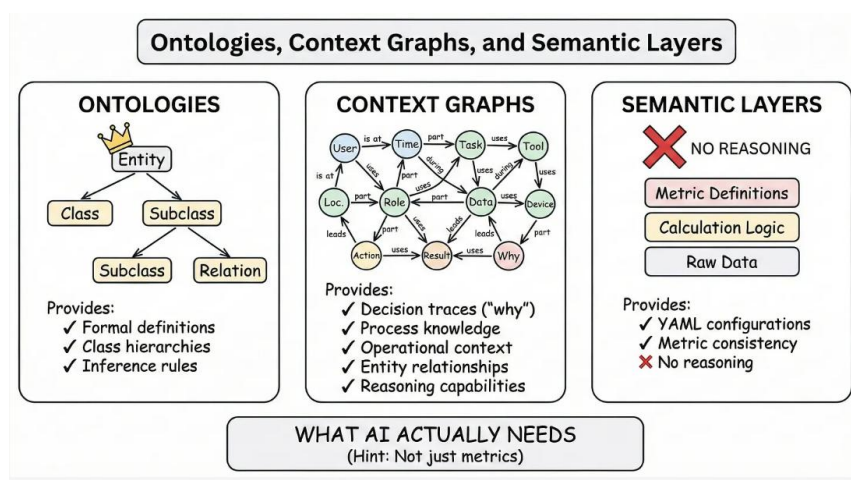


资料来源：中泰证券研究所

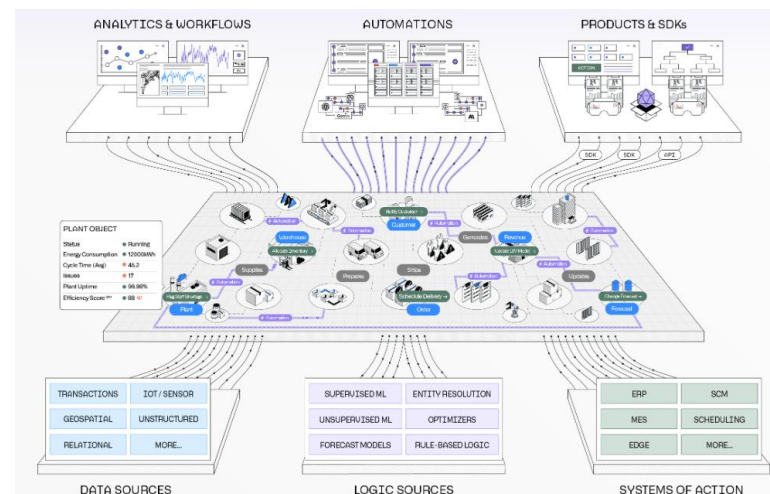
Token作为隐形资产：上下文图&决策痕迹可能是被低估的价值

- 我们认为，最值得注意的是，Token正在从纯流量概念向存量资产沉淀演进，即企业内的上下文图谱和决策痕迹（Decision Traces）。智能体每执行一次任务，都会在消耗Token的同时留下一条决策痕迹——调用了哪些系统、适用了哪条政策、由谁批准了哪项例外，这些痕迹随时间累积成一张可查询的组织决策图谱。
- Token消耗与存量资产由此形成飞轮。每一次Token消耗都在向图谱写入新的先例，可检索的先例又让后续每单位Token完成任务时更准确、所需的背景重述更少，对应公式中分子扩张与分母压缩的同步改善。模型能力趋于同质化，而组织独有的决策图谱无法事后重建、难以复制，其价值随使用时间复合增长。更关键的是这类隐形资产的归属权。如Palantir的本体层与处于编排路径的智能体公司，正在将下游客户消耗Token产生的决策痕迹沉淀为自身平台的复利资产。Token的支出方与存量资产的占有方出现分离，归属权之争将成为价值分配的核心战场。

图表：What AI Actually Needs



图表：Palantir Ontology系统





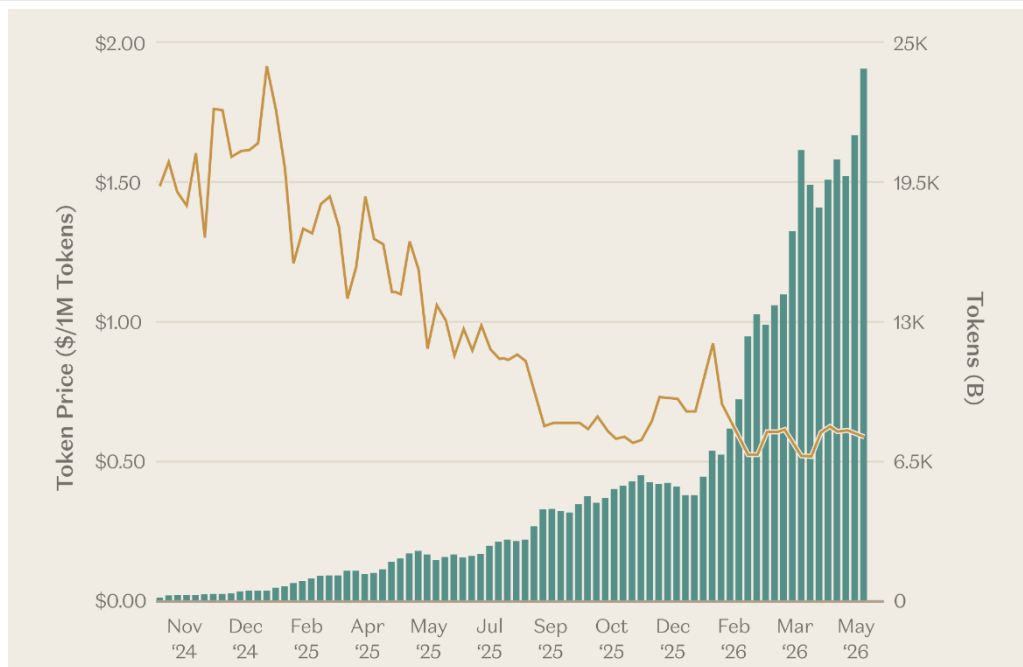
2

Token的杰文斯悖论：新应用场景解锁 提升token需求

Token经济学中的杰文斯悖论

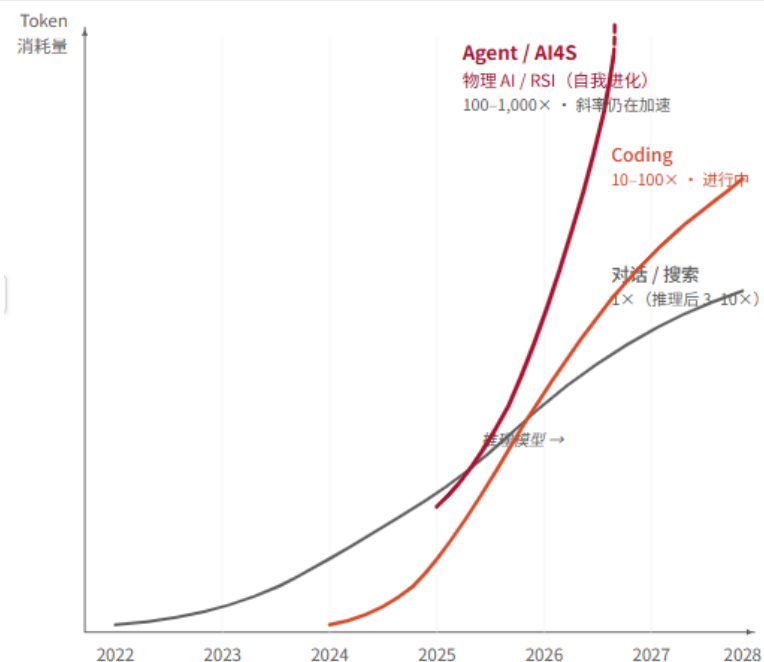
- 杰文斯悖论（Jevons Paradox）指效率提升降低了每单位服务的有效成本，从而刺激了更多需求、解锁了更多应用场景、带动了更大规模的经济活动；最终总消耗量不减反增，甚至远远超过效率提升所节省的部分。
- Token经济同样呈现了杰文斯悖论。随着技术优化Token单价正在持续降低，但消耗总量快速提升，总支出持续扩大。从定义公式的角度拆解，当分母中的Token单价下降（基模降价、硬件代际跃迁、MoE架构优化）；新的场景一旦被解锁，就会产生新的Token消耗。特别是，新解锁的场景更复杂，往往会比已有场景消耗更多Token（如Coding之于Chatbot）。

图表：Token价格与支出总量的对比



资料来源：a16z，中泰证券研究所

图表：核心场景Token消耗示意

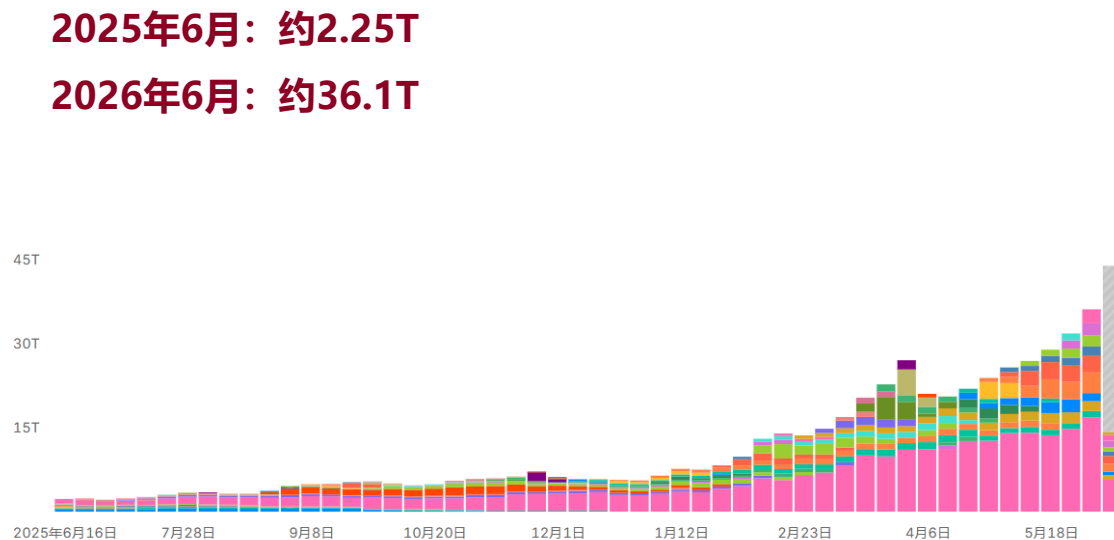


资料来源：中泰证券研究所

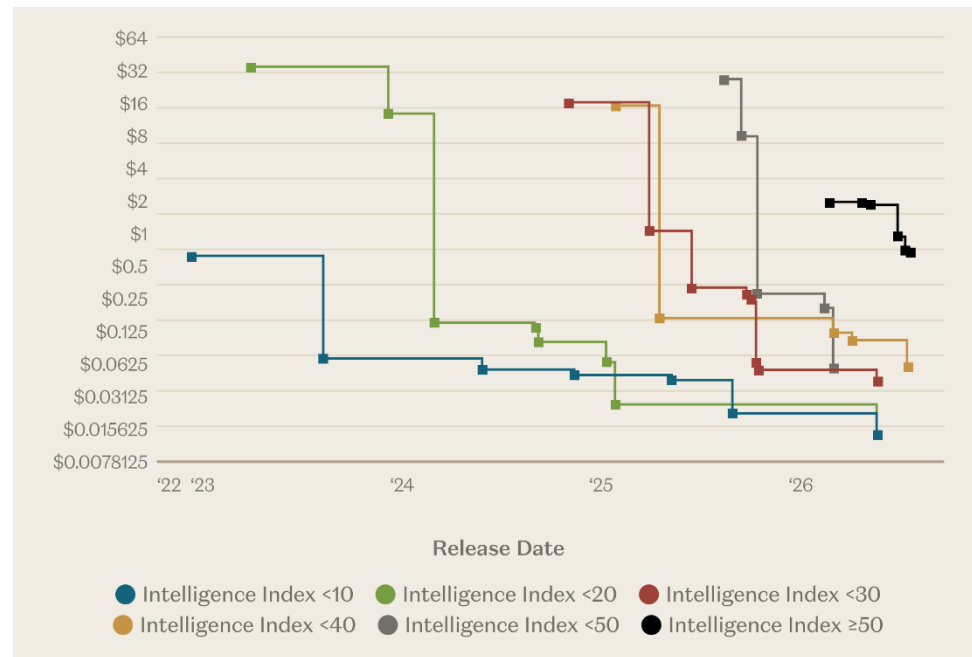
Token经济学中的杰文斯悖论

- 总量来看，OpenRouter平台的周度Token吞吐量从2025年6月的约2T快速提升至2026年6月约36T，一年内增长约16倍，且增速仍在加快。单价来看，Token单价的持续下降。a16z的数据将模型按智能指数分层追踪发布价格，每一条阶梯线都在向右下方移动：同等智能水平的模型，每隔12-18个月价格下降一个数量级。

图表：2025年6月-2026年6月OpenRouter平台token调用量大幅攀升



图表：智能模型发布价格下降



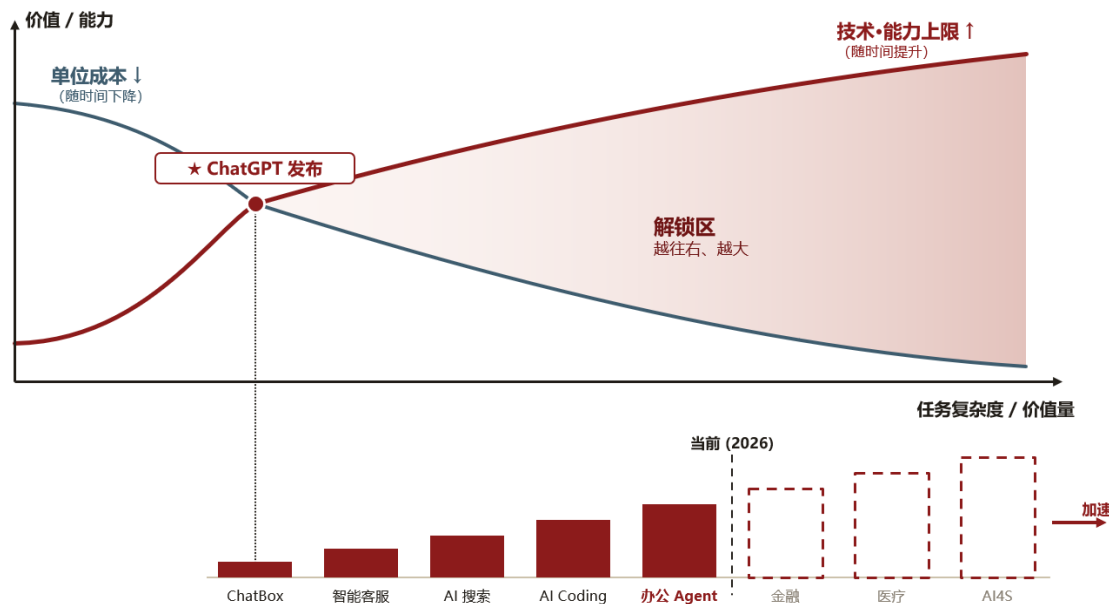
资料来源：OpenRouter，中泰证券研究所

资料来源：a16z，中泰证券研究所

Token量爆发，依旧是基模“能力提升+成本下降”解锁更多场景的结果

- 应用层的渗透依旧遵循我们此前观察到的规律：随着性能提升和成本下降，越来越多的应用场景正在被解锁。
- Token调用量的攀升还是来自下游需求场景的爆发。以Coding场景为例，在编程能力显著增强的Claude 3.5模型发布后，Coding应用如Cursor、Lovable的ARR开始有了显著增长。

图表：价值-成本曲线交叉解锁更多应用场景



资料来源：中泰证券研究所

图表：Claude 3.5发布后Cursor等Coding AI公司ARR快速上升

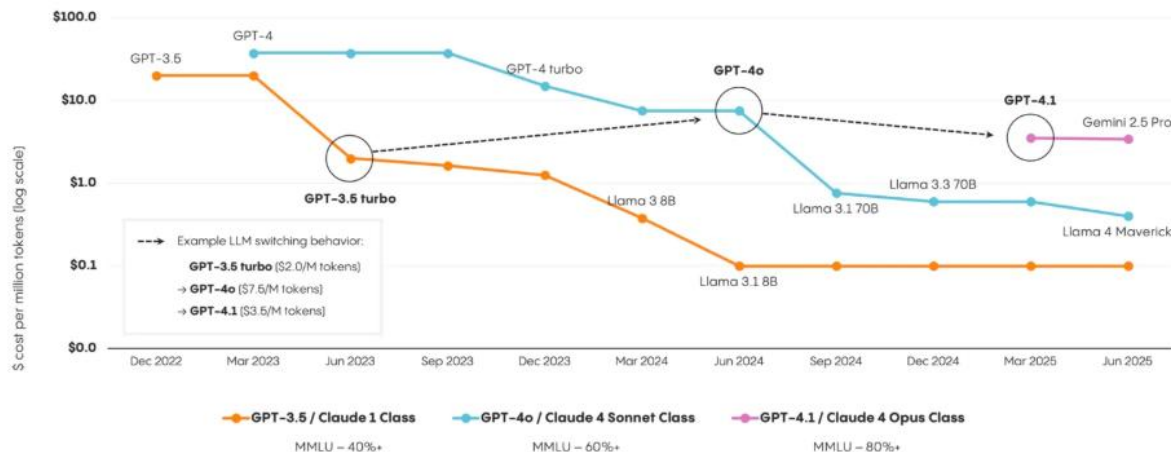


资料来源：Ethan Ding，中泰证券研究所

复盘Coding场景：早期开发者优先考虑性能，即使前沿模型更贵

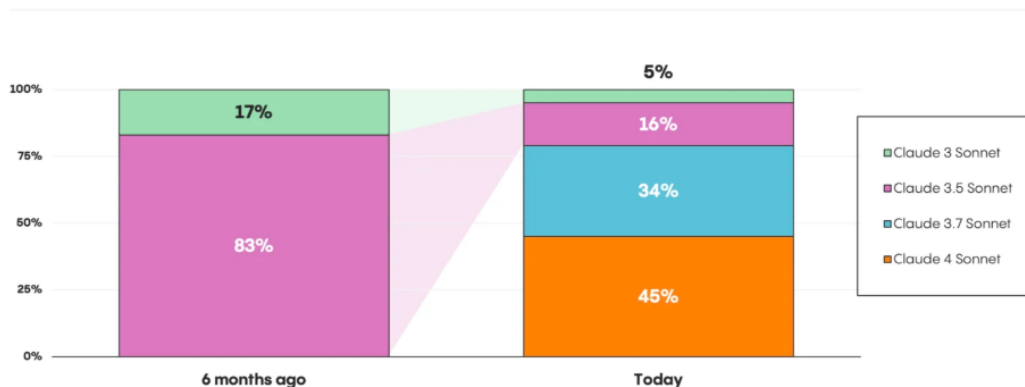
- 在Coding场景中，模型性能决定开发者的购买决策。开发者始终选择前沿模型，而非更便宜、更快速的替代方案。他们优先考虑性能并为性能付费。（始终聚焦开发前沿模型也是Anthropic一直坚持的策略）例如，在Claude 4发布后的一个月内，Claude 4 Sonnet 就吸引了45%的 Anthropic用户，而Sonnet 3.5的份额则从83%下降到16%。即使个别型号的价格下降了10 倍，用户也不会通过使用旧版本模型来节省成本，而是集体转向性能最好的新版本模型。

图表：开发者坚持采用最先进版本模型



图表：开发者倾向于选择前沿模型

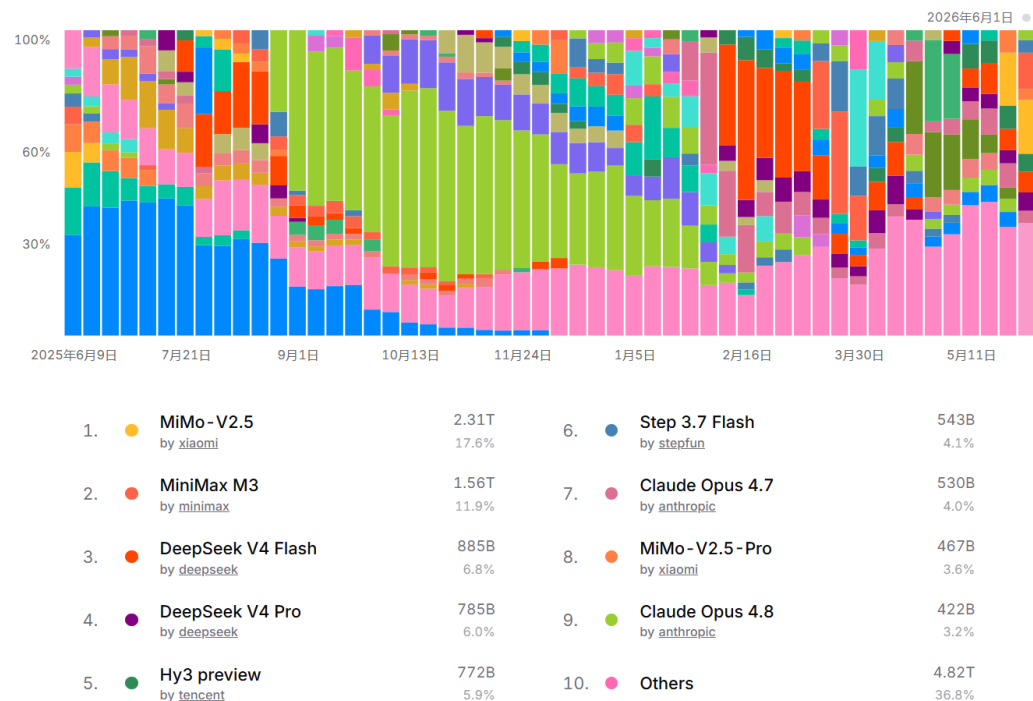
Builders Choose Frontier Models



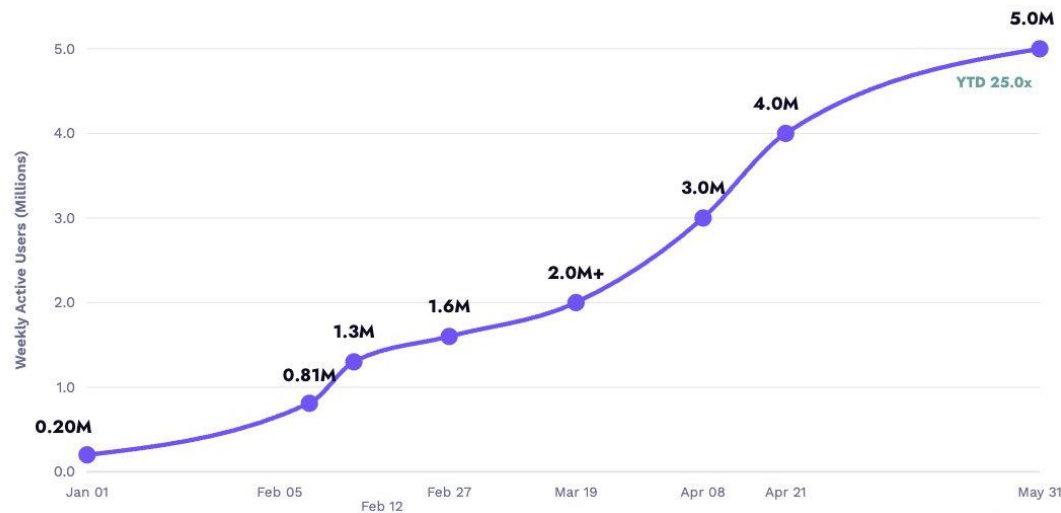
复盘Coding场景：中后期低价模型Token快速放量，性价比优势传导加速

- 近期OpenRouter编程场景的统计中，MiMo-V2.5、DeepSeek V4等性价比模型调用量占比提升。反映了随着模型能力的迭代，多数Coding场景对模型可能已经变得“简单”，此时用户会为了性价比因素，将简单任务分配给能力稍弱、但价格更低的模型。2026年5月，CodeX实现了500万活跃用户数量。

图表：编程场景OpenRouter调用量占比统计（2026年6月8日统计）



图表：CodeX用户数量快速增加



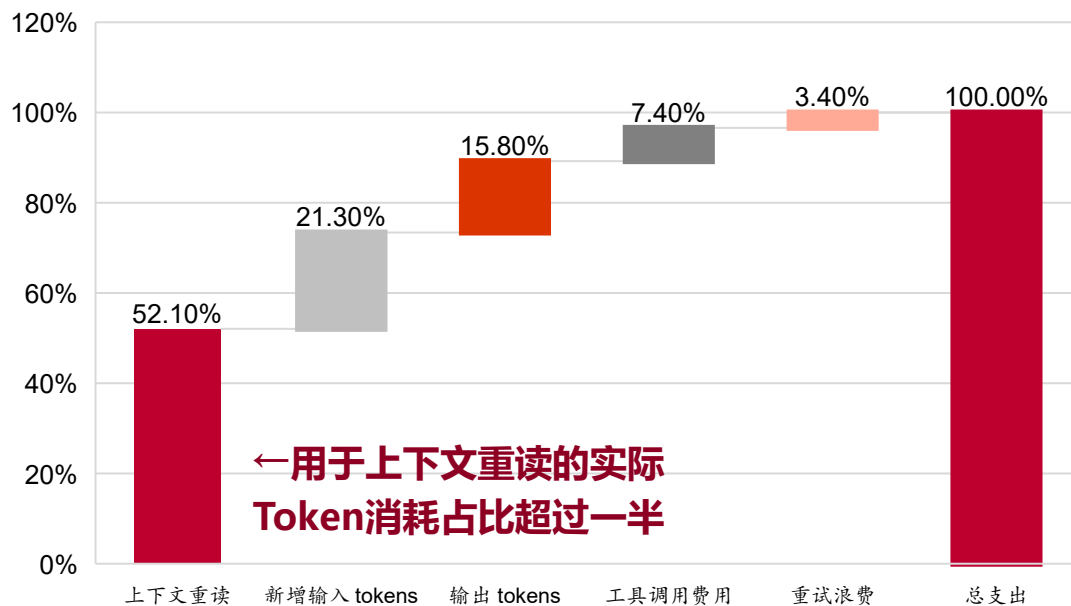
资料来源：OpenRouter，中泰证券研究所

资料来源：ARK Invest，中泰证券研究所

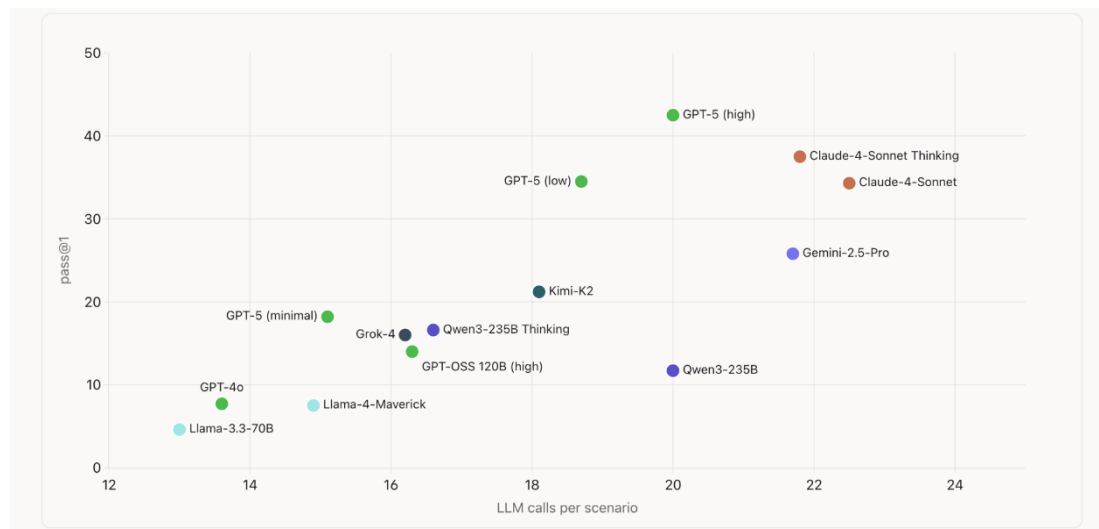
Agent: 上下文重复读取+工具调用推高Token消耗量

- Agent往往消耗很多隐形词元（ghost tokens）。第一个因素来自上下文重复读取。一个Coding Agent 在 10 轮迭代中，每一轮都需要重新读取完整的上下文窗口来维持连贯性，仅这一项重复读取，就可能产出比单轮查询高达 55 倍的 Token 消耗。第二个乘数来自工具调用的长尾效应。据测试，Agent 每完成一个任务平均执行 5-25 次工具调用。每次工具调用都会向上下文追加返回结果，进一步膨胀后续轮次的输入 Token 量，同时提高模型在长链路中出错并需要重试的概率。

图表：实际Token消耗占比情况



图表：Gaia2分数与LLM工具调用次数情况



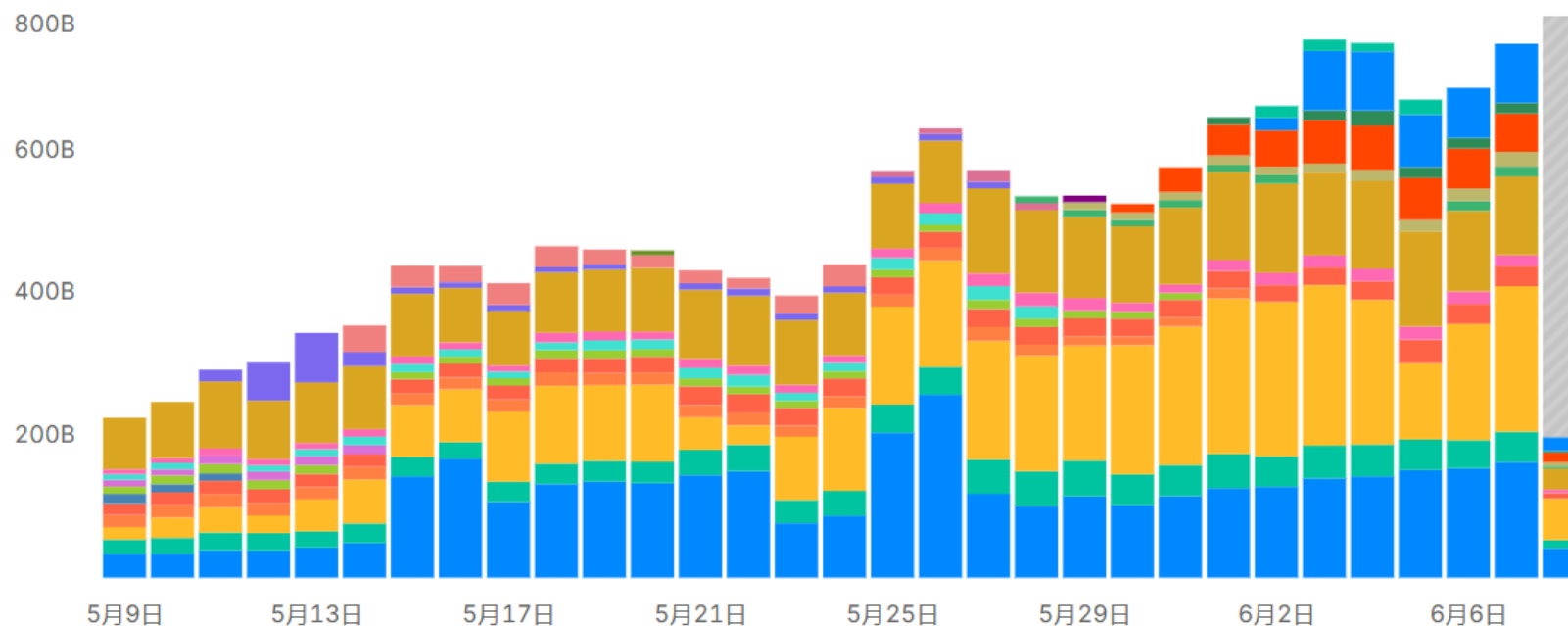
资料来源：Exponential View Analysis，中泰证券研究所

资料来源：Exponential View Analysis，中泰证券研究所

OpenClaw、Hermes等智能体进一步推升token消耗

- 2026年以来，以OpenClaw和Hermes Agent为代表的支持持久运行任务的自主Agent迅速崛起，成为OpenRouter平台上最大的Token消耗来源。这类Agent以后台常驻进程运行，具备跨会话记忆累积、自主任务调度和自我改进循环能力，并可以将 workflows 提炼为可复用技能，下次遇到相似任务时调用并迭代优化。2025年6月，Hermes在openrouter上的token消耗量已达约800B数量级，较5月初约翻4倍。

图表：Hermes在Openrouter上的token消耗情况（统计时间：2026年6月8日）

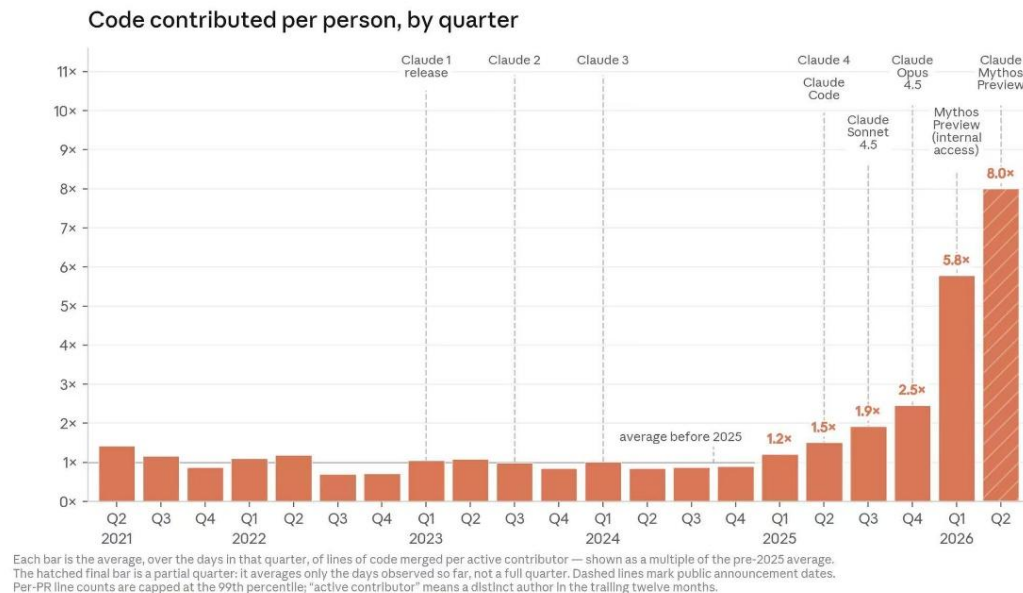


资料来源：Openrouter，中泰证券研究所

Agent或将打破软件工程“人月神话”规律

- 几十年来约束软件行业的“人月神话”正在 **Token** 时代失效。Fred Brooks 在 1975 年提出的 Brooks 定律指出：向一个已经延期的软件项目增加人手，只会让它更晚交付：新成员加入后需要一段时间才能有所产出，人与人之间的沟通成本快速增长。这使得软件开发的规模化路径与制造业截然不同：工厂加十名工人可以多造十件产品，但软件公司加十倍的程序员和十倍的预算，产出无法达到十倍。
- 如今现代 **AI** 开发的核心资源投入从“人”变成了“算力”，而算力的扩展几乎是线性的：双倍 **GPU** 带来双倍 **Token** 产出。更少或者不变的工程师 + 更多的**Token**投入 = 更多更高质量的代码产出。

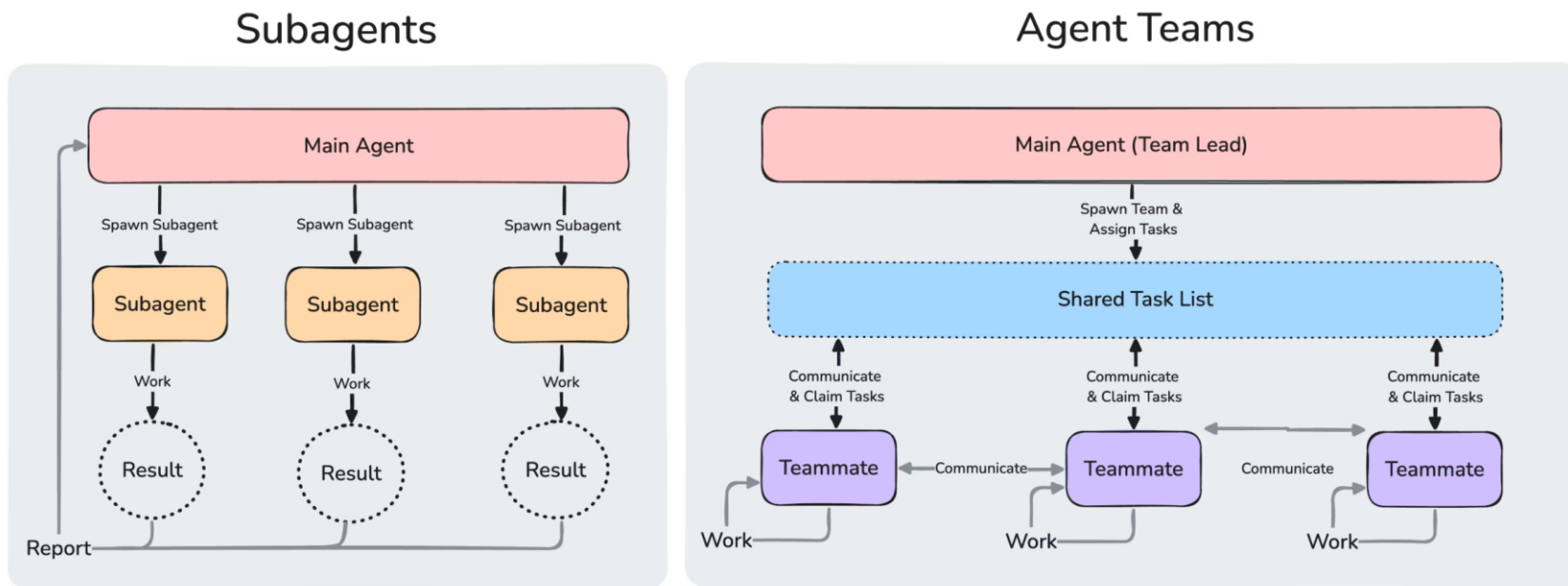
图表：单人提交的代码量快速增长



并行Agent：子代理与代理集群大势所趋

- **Agent并行化**正在成为复杂任务执行的主流范式，当前分化为子代理与代理集群两种架构路线。子代理（Subagents）由主代理统一调度，各自完成聚焦任务后将结果摘要回传主上下文，Token消耗相对可控。代理集群（Agent Teams）中每个成员拥有独立上下文窗口与完整模型实例，通过共享任务列表自主认领工作并直接通信，在并行调研、多模块开发、竞争性假设调试、跨前后端协同等场景中展现出显著优势。

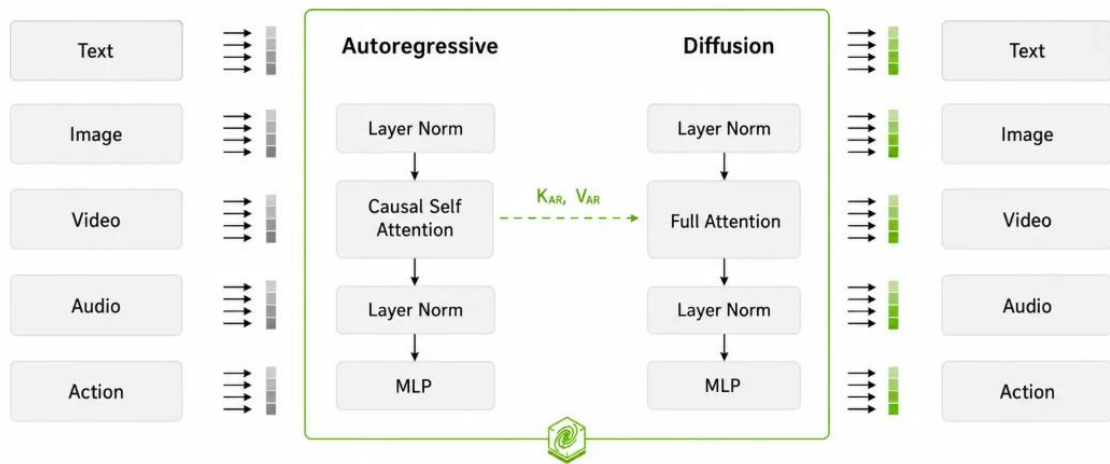
图表：SubAgents与Agent Teams架构对比



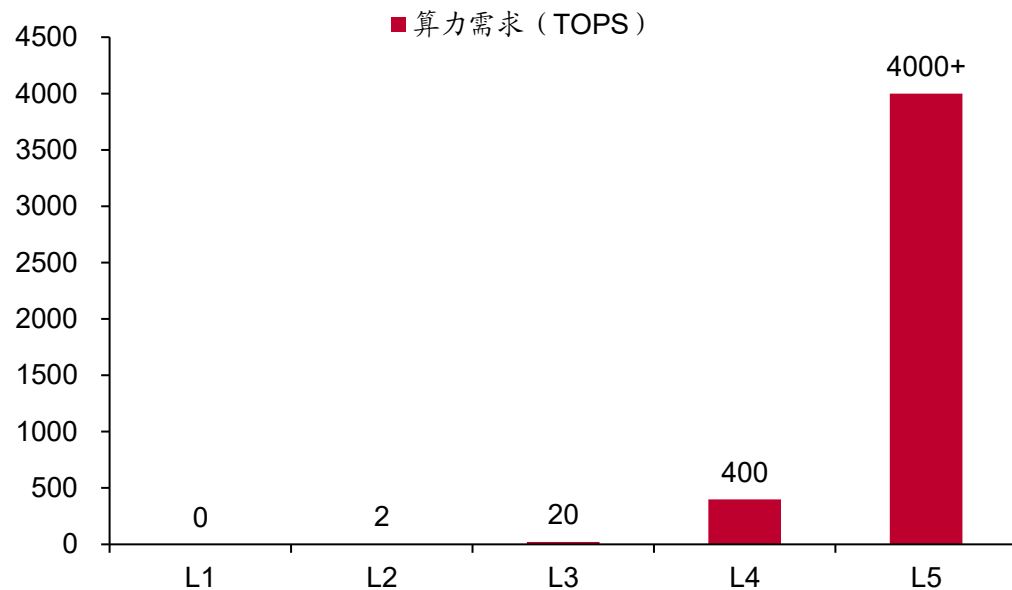
其他场景：多模态与物理AI有望推升Token需求

- 以多模态、物理AI为代表的场景将带来新一轮的token需求爆发。多模态场景的Token消耗量远超纯文本：以Gemini的计量规则为参照，视频理解每秒消耗约300个Token，Seedance 2.0生成15秒视频需约31万Token，单次视频生成消耗Token远超文本交互场景。NVIDIA Cosmos为代表的多模态世界模型也正在快速迭代，训练规模达9000万亿Token。
- 以物理AI代表性的智能驾驶场景为例，L3级自动驾驶对算力需求为20-30TOPS，L4级需要200TOPS以上，L5级则超过2000TOPS，背后同样代表了不同数量级的Token流量。

图表：Cosmos 3的多模态架构



图表：不同等级自动驾驶的算力需求



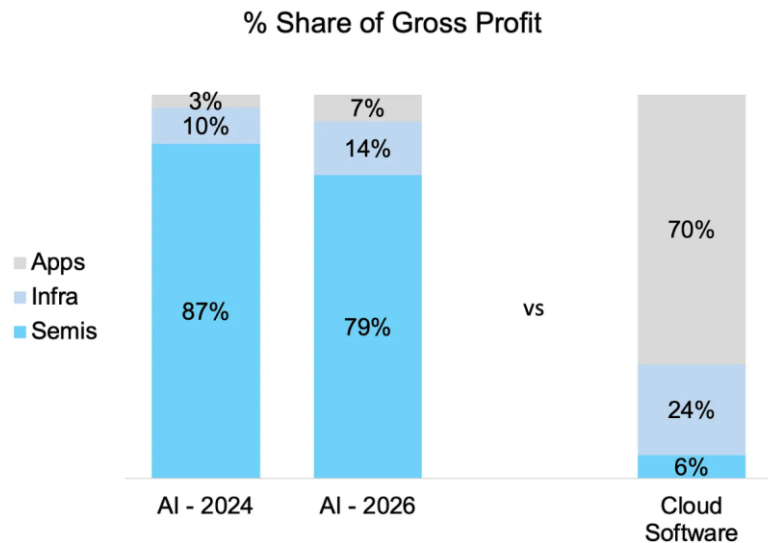
资料来源：NVIDIA，中泰证券研究所

资料来源：，中泰证券研究所

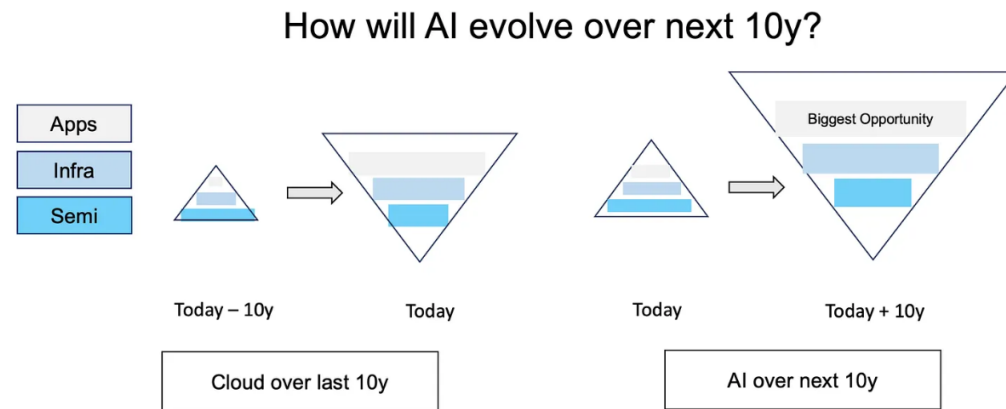
伴随流程改造，未来应用层在价值链占比将提升

- 当下产业正处于转型的早期阶段，半导体（如NVIDIA）占据了产业链上大部分价值，高价值应用场景还未被完全挖掘，ROI难以测算，应用层价值量并未完全凸显。我们认为，更先进的AI直接塞进陈旧的工作流程中很难显著提升效率，只有同时进行工作流和组织架构的“原生AI化”改造，AI才能创造更大的应用价值，这也意味着可能需要更长的时间。但我们依旧相信随着基础模型能力提升，更多应用场景终将不断跑通终将创造更大的价值；通用人工智能目前的倒金字塔形收入结构不会永远保持下去，应用层在价值链上的占比将逐渐提升。

图表：各层AI与云产业毛利占比



图表：未来10年AI Stack将如何发展？



Token出海：以Token计费完成价值跨境交付

- Token出海本质是中国AI以Token为核心计价与流通单位，向全球输出推理算力与智能服务的新型数字贸易模式。其本质是将国内电力、算力和算法转化为可标准化计量、可全球交付的数字服务，即电力留在境内电网，但电力的价值通过Token完成了跨境交付。Token的出现，首次解决了数据资产化的难题。数据经标准化转化为Token后，变得可计量、可收费，完成了从“资源”到“资产”的历史性跃迁。随着未来跨境数据流通监管的逐渐完善，Token出海模式有望更加规范，为产业带来全新增量机遇。

图表：OpenRouter当月调用量排行榜（统计时间2026年6月9日）

 LLM Leaderboard

Compare the most popular models on OpenRouter 

1.



DeepSeek V4 Flash

by [deepseek](#)

13.2T tokens

↑ 634%

2.



Hy3 preview

by [tencent](#)

12.8T tokens

↑ >999%

3.



Claude Opus 4.7

by [anthropic](#)

7.67T tokens

↑ 121%

4.



Claude Sonnet 4.6

by [anthropic](#)

7.61T tokens

↑ 30%

5.



Owl Alpha

by [openrouter](#)

6.09T tokens

↑ >999%

6.



DeepSeek V4 Pro

by [deepseek](#)

5.39T tokens

↑ 370%

7.



Gemini 3 Flash Preview

by [google](#)

4.54T tokens

↑ 0%

8.



DeepSeek V3.2

by [deepseek](#)

4.52T tokens

↓ 5%

9.



MiMo - V2.5

by [xiaomi](#)

4.34T tokens

↑ >999%

10.



MiniMax M3

by [minimax](#) new

2.94T tokens

11.



Kimi K2.6

by [moonshotai](#)

2.82T tokens

↓ 42%

12.



Nemotron 3 Super (free)

by [nvidia](#)

2.57T tokens

↑ 1%

13.



Gemini 2.5 Flash Lite

by [google](#)

2.53T tokens

↓ 4%

14.



Gemini 2.5 Flash

by [google](#)

2.48T tokens

↓ 1%

15.



MiMo - V2.5 - Pro

by [xiaomi](#)

2.47T tokens

↑ 661%

16.



MiniMax M2.7

by [minimax](#)

2.32T tokens

↓ 35%

17.



Claude Opus 4.6

by [anthropic](#)

2.14T tokens

↓ 37%

18.



GPT- 5.5

by [openai](#)

1.98T tokens

↑ 218%

19.



gpt-oss- 120b

by [openai](#)

1.75T tokens

↑ 1%

20.



Laguna M.1 (free)

by [poolside](#)

1.72T tokens

↑ 590%



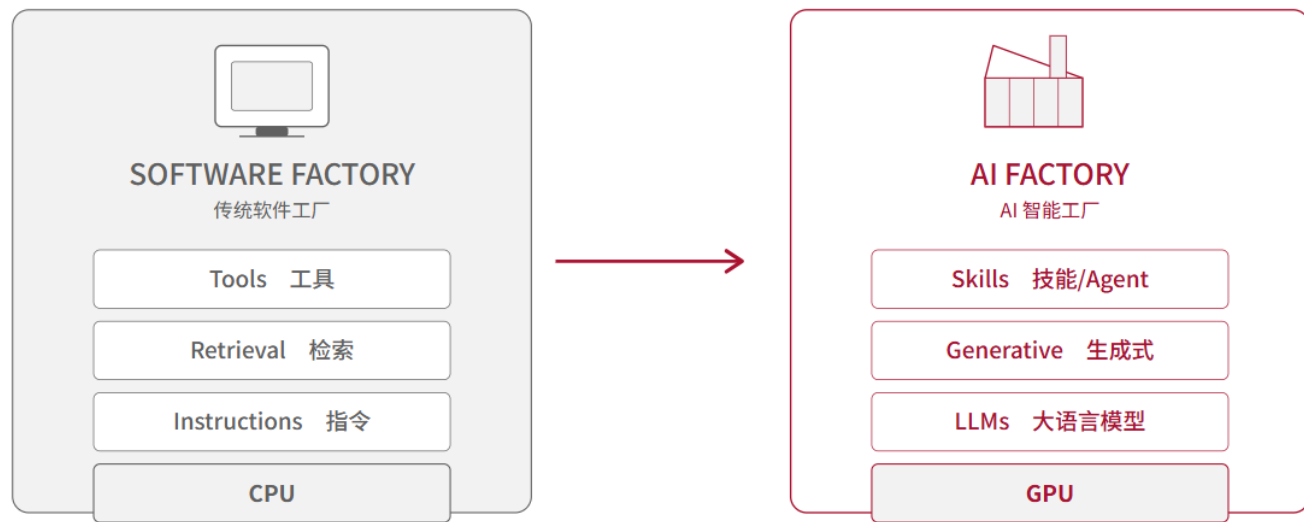
3

算力依旧是核心受益环节

数据中心正在转变为“token工厂”，制造业逻辑下产业面临算力约束

- 我们认为AI产业一定程度上有类似制造业的逻辑。资本开支（算力投入）往往是前瞻性投入，是根据未来算力需求预估的。Jensen Huang 在 GTC 2026 上将数据中心重新定义为“AI工厂”。一家前沿 AI 公司运营的不是传统意义上的软件业务，而是一座智能工厂。Token的生产同时具有制造业的产能约束、软件业的迭代速度、云计算的按量服务模式、以及前沿技术快速迭代的不确定性。
- 这同样解释了前沿模型厂商面临算力瓶颈的核心原因。一方面，即使是前沿模型和云厂商，也难以准确预测一年后所需的产能（算力基础设施规模）；另一方面，产能就是收入上限，不投产能就是放弃增长。这也形成了对算力需求的拉动作用。

图表：传统软件工厂与AI智能工厂

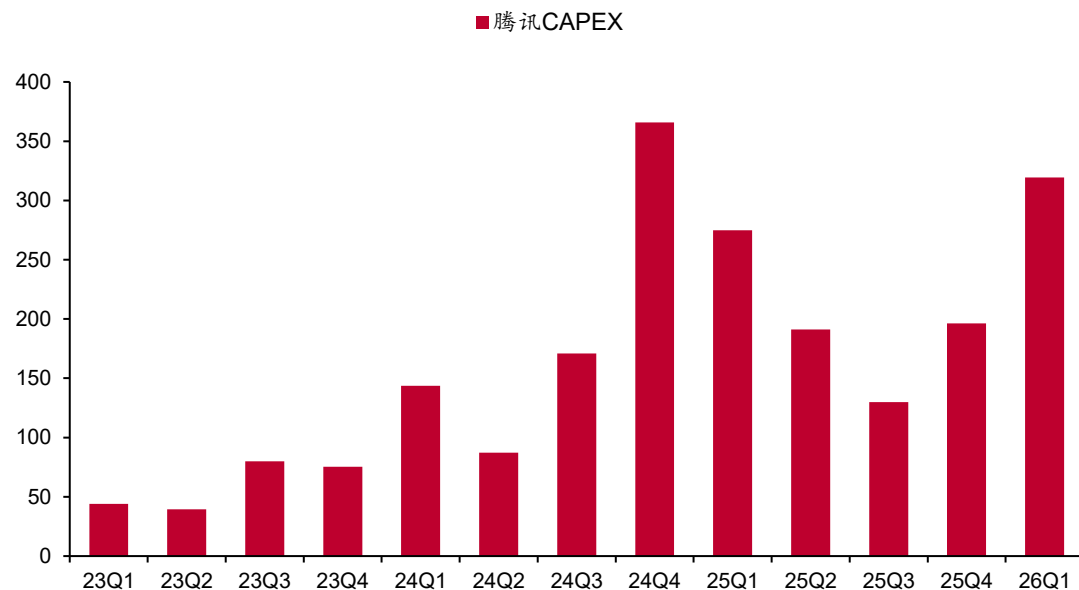


资料来源：NVIDIA，中泰证券研究所

主要互联网厂商资本开支持续攀升，算力需求加速

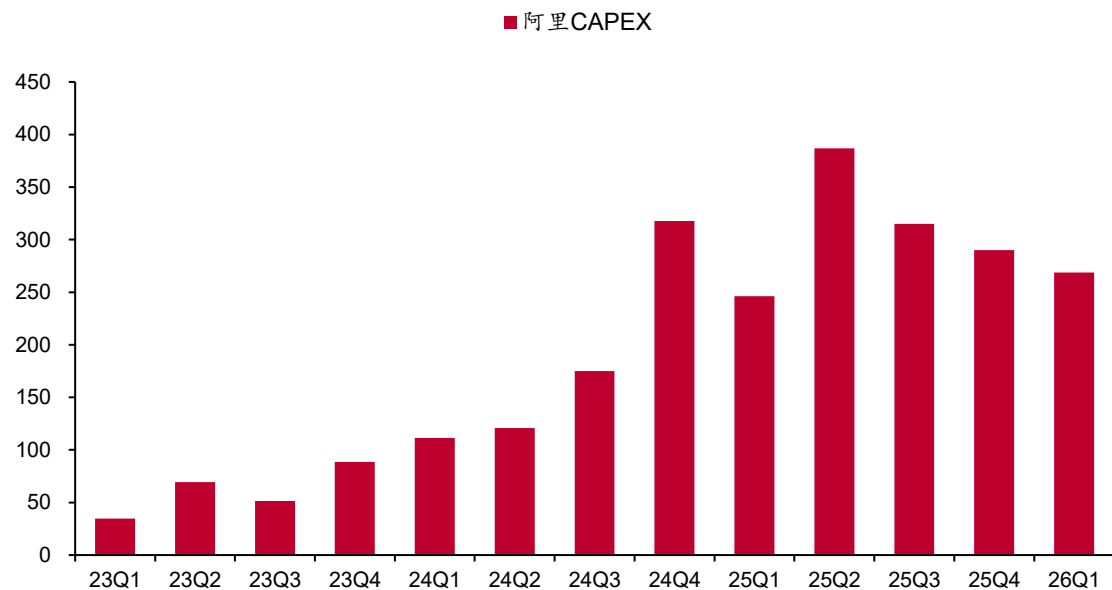
- 阿里巴巴：2025 年 9 月 24 日至 26 日，阿里巴巴集团董事兼首席执行官吴泳铭在云栖大会上表示，阿里巴巴正积极投入资源，计划在未来进行 3800 亿的 AI 基础设施建设，并可能追加更大的投资。
- 腾讯：2026 年第一季度公司资本开支达 319.36 亿元，资本支出快速加码。

图表：23Q1-26Q1 腾讯资本性支出（单位：亿元）



资料来源： ，中泰证券研究所

图表：CY23Q1-CY26Q1 阿里巴巴资本性支出（单位：亿元）

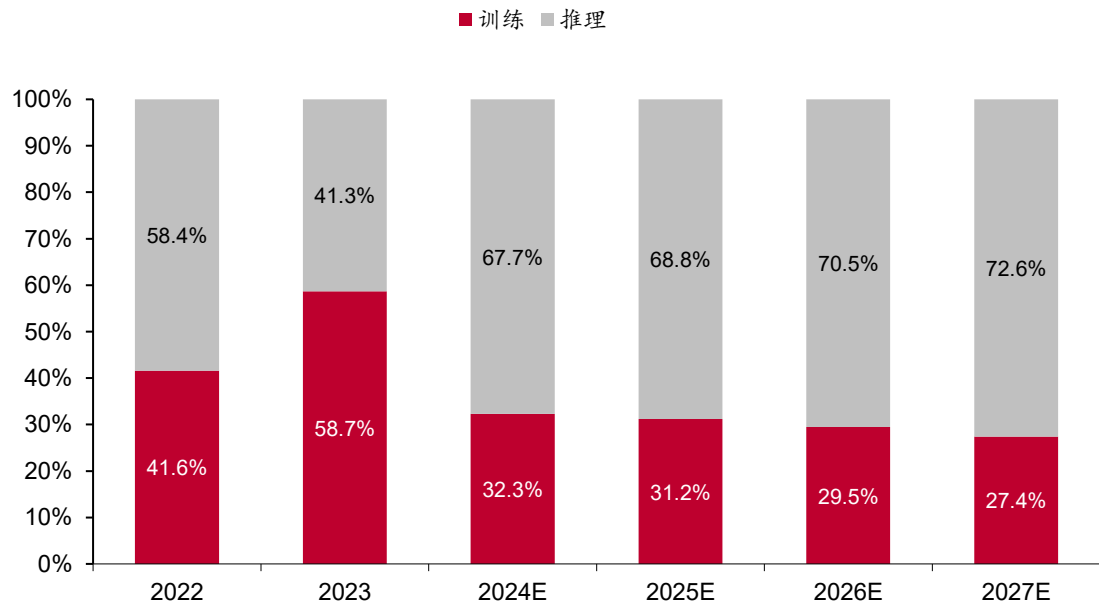


资料来源： ，中泰证券研究所（注：此处已调整为自然年季度表述）

从训练进入推理，利好国产芯片加速渗透

- 我国 **Token 消耗量、推理 Token 消耗量**快速增长。据国家数据局局长刘烈宏介绍，2024 年初，我国日均 Token 的消耗量为 1 千亿，截至 2025 年 6 月底，日均 Token 消耗量已经突破 30 万亿，一年半时间增长了 300 多倍，反映出我国人工智能应用规模的快速增长。华为常务董事汪涛也表示，AI 推理迎来爆发式的增长，推理的快速发展拓宽了人类智能化的广度。
- 随着训练模型的完善与成熟，模型和应用产品逐步进入投产模式，处理推理工作负载的人工智能服务器占比将随之攀升。据 IDC 报告（2024 年），推理市场占全球 AI 芯片支出的比例突破 65%，远超训练的 35%。IDC 预计，到 2027 年，中国用于推理的工作负载占比将达到 72.6%。
- 在 **AI 算力重心由训练端转向推理端**之际，**ASIC 芯片**的高度定制化和能效优势趋势得以逐渐清晰。行业数据显示，2025 年，全球 ASIC 市场规模预计达 220 亿美元，其中 AI 相关占比 15%，到 2030 年有望突破 400 亿美元。
- **AI 算力基础设施走向异构融合的道路上，国产 ASIC 芯片厂商**迎来前所未有的机遇。国产厂商在高速 SerDes 互连、高能效供电、先进存储控制以及信号链与接口等关键配套芯片领域已取得显著突破，并形成体系化能力。这些正是构建高性能、高集成度、低功耗 ASIC 的核心基石，也为国产厂商在 ASIC 道路的进一步升级提供了坚实保障。

图表：2022-2027 年中国人工智能服务器工作负载占比及预测

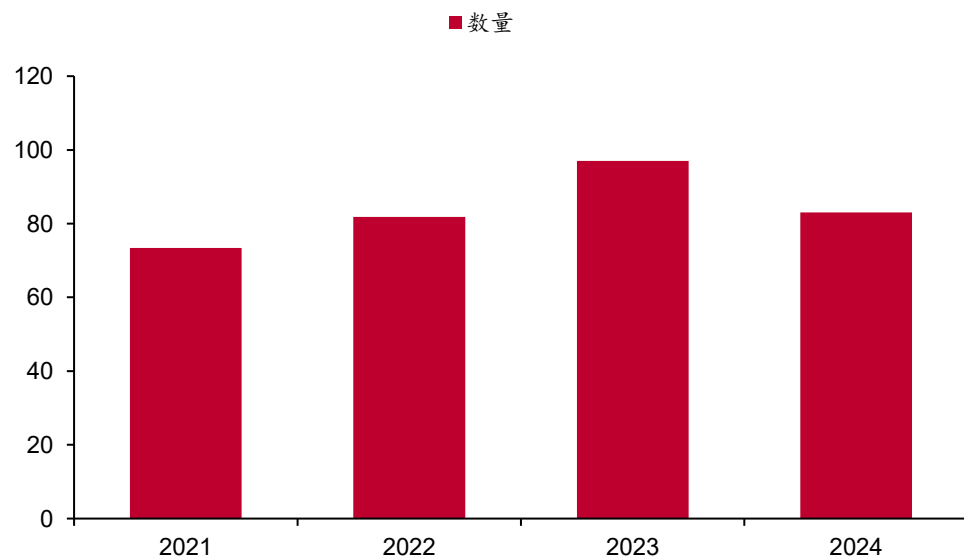


资料来源：IDC，中泰证券研究所

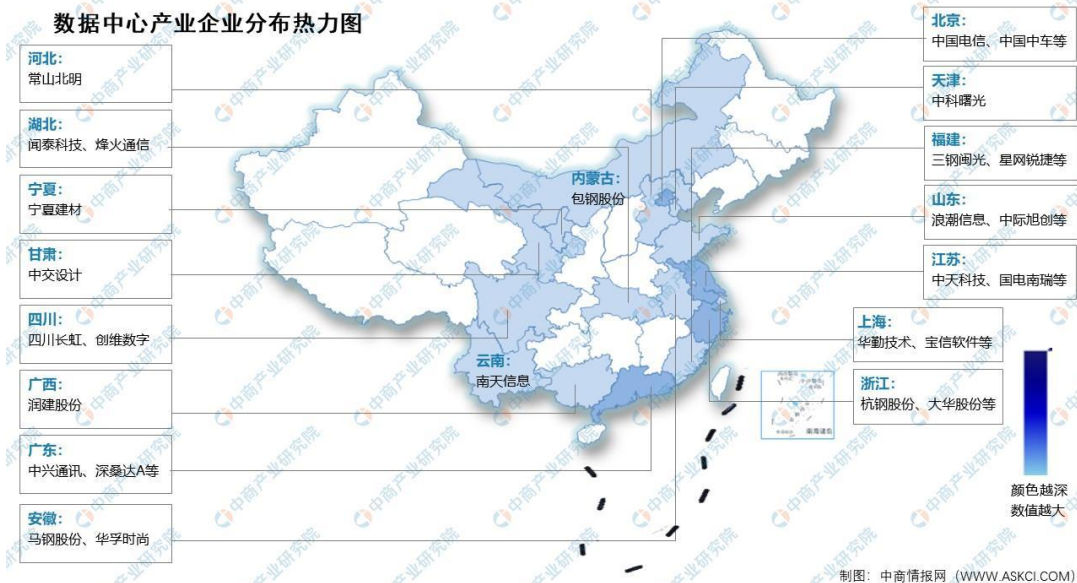
国内数据中心加速建设

- 数据中心建设协调推进。三家基础电信企业持续优化算力基础设施布局，截至 2024 年底，向公众提供服务的互联网数据中心机架数量 83 万个，推动提升算网协同和调度能力，提供更加多元化算力服务。
- 以三大运营商为代表，国产服务器的在集采项目中中标比例持续攀升。
 - 以2025年4月公布的中国电信服务器（2024-2025 年）集中采购结果为例，本次服务器集采中国产化系列数量达 10.53 万台，占比达 67.5%。在架构方面，ARM 服务器如鲲鹏、飞腾等占比达 45%，C86 架构（海光 X86 方案）占比8.7%。
 - 2025 年 9 月 27 日，在中国联通 2025 年通用服务器集中采购项目中，标包一由紫光华山、浪潮信息、超聚变三家公司中标，标包二由新华三、武汉长江计算、联想开天、宝德计算机、软通计算机、浪潮计算机等八家公司中标。从出货量角度来看，国产服务器已经超过 90%；从采购金额来看，国产服务器占比达 88.8%。

图表：2021-2024 年数据中心机架数量发展情况（单位：万个）



图表：数据中心产业企业分布热力图



算力租赁有望挖掘存量资源潜力，以高性价比缓解当下算力短缺困局

- 同时我们也看到，在当前市场环境下，适用于 AI 计算的高性能 GPU 供不应求，许多 AI 公司无法购买到足够的计算卡来搭建自己的算力集群，而算力租赁可以高性价比缓解当下算力短缺的困局。
- 算力租赁是当前中小企业解决算力需求的最优解之一。目前，除了少数大型互联网企业自身资金实力充沛，可购买较多的 GPU，算力储备较充足之外，剩下中小企业普遍面临算力紧缺，算力租赁需求突出。
 - 对于规模较小的公司，购买 GPU 搭建算力集群的投入成本过高，租赁外部算力相对自建算力更合算、灵活。
 - 自建集群规模固定、可扩展性较弱、可靠性较差，中小企业的算力需求往往难以通过自建算力设施解决，而算力租赁可有效降低除硬件成本外的维护、升级等长期投入，使其直接享受高性能算力的便利，以快速响应市场变化，把握发展机遇。

图表：算力自建与租赁成本对比

成本

自建算力

以175B参数，300B tokens训练集大小作为参照；以DGX A100(5petaFLOPS)，45%的训练有效性；训练30天作为中型企业的训练标准，需要55台A100,每台15万美元，共825万美元(近6000万人民币)，此外还有机房建设、运维、减值等费用。

算力租赁

$4.5 \times 24 \times 30 \times 8 \times 55 = 142.56$ 万美元(近1000万人民币)，没有其他费用，且中小型企业一次训练+更新后一般不用再训练，只有推理需求，且明后年降本50%以上(有降价趋势)。对于租赁运营商而言，有测算单台A100服务器生命周期内IRR为35%(需验证)。

资料来源：观研天下，中泰证券研究所

算力租赁市场未来或向提供更高附加值的运营服务方向转型

- 算力租赁市场现有商业模式：基于硬件资源的按需租赁和按量付费模式。面对激烈算力租赁竞争，多元算力融合成为关键，算力市场将更加重视辅助运营服务，从提供硬件资源逐步转变为提供算力服务。
- 我们认为，未来，国内算力租赁市场或有望向提供更高附加值的运营服务方向转型。
 - 算力调度：通过智能分配策略实现算力的灵活流动，进一步解决算力需求与资源分布不均的矛盾，快速满足上层应用多样化的算力需求，助推数字经济进入普惠共享的新阶段。
 - 提供整体 AI 解决方案：以 GPU 云为例，其除了提供算力外，还包括了如 AI 软件开发相关的增值服务，是未来算力租赁的进阶方向，增值潜力高。

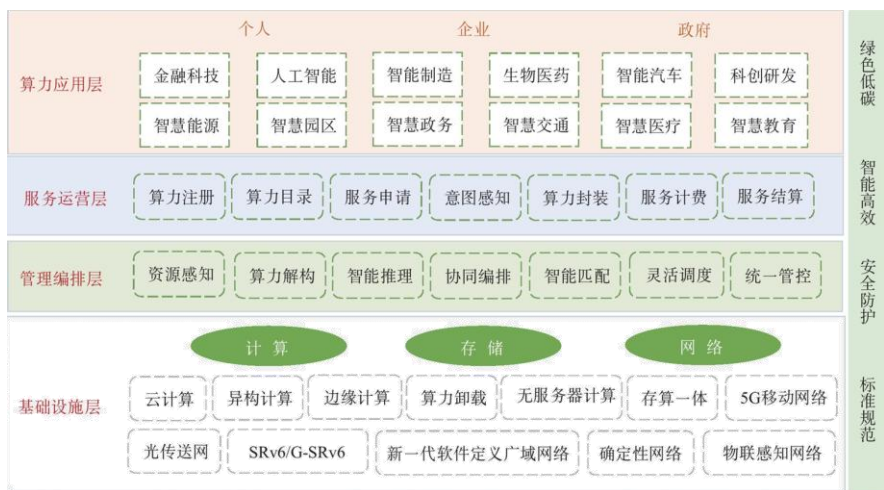
图表：算力市场向多元化服务发展



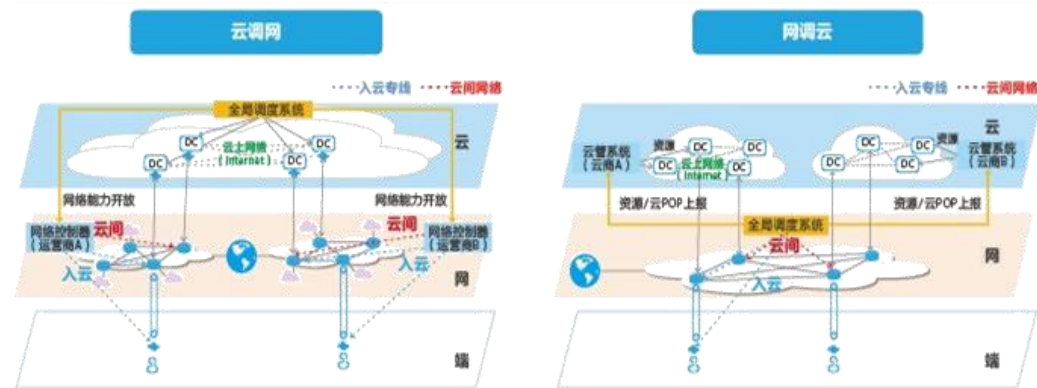
全国一体化算力网，算力调度运营进程加速

- **算力调度**是根据算力资源提供方的供给能力和应用需求方的动态资源需求，整合区域内算力基础设施底层的计算、存储、网络等多维资源，基于算力调度平台对算力资源进行一致性管理、一体化编排和统一调度，可以整合多张算力卡来应对外围禁售、优化算力资源配置，是解决算力供需矛盾、算力网络传输问题、算力资源普惠问题的新型能力体系。
- **“以网强算”**是发展算力网络的根本路径。算力网络可以实现云、边、端算力的高效调度，吸纳全社会算力资源，组成泛在、立体的算力网络，实现智能调度和全局优化。
- **全国一体化算力算网调度平台**综合集成网络情况+综合算力+算力调度“三位一体”推动我国算力算网调度发展。平台汇聚通用算力、智能算力、高性能算力、边缘算力等多元算力资源，针对通用、智算、超算等不同客户需求，设计异构资源池调度引擎，实现不同厂商的异构资源池的算力动态感知与作业智能分发调度。在AI训练作业调度流程中，作业可在智算资源池上进行训练推理，在通用算力资源池部署，从而实现跨资源池/跨架构/跨厂商的异构算力资源调度。

图表：算力调度涉及到的关键环节



图表：算网调度的两种路径

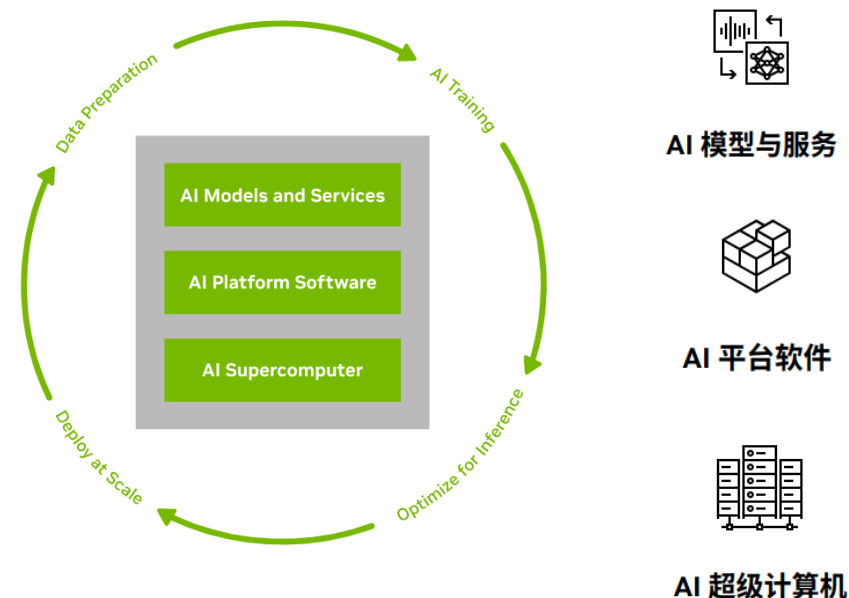


整体AI解决方案激发极致算力

- 整体 AI 解决方案重视全栈客户体验：从设备到算力，基于客户全场景需求，围绕算力咨询、建设和运营等全周期，提供端到端全栈专业服务，全程护航算力集群建设、人工智能创新、产业聚合发展。
- AI 解决方案的四个关键功能：
 - 自动化流程：通过收集和解释输入其中的大量数据，可以利用人工智能解决方案来确定流程中的下一步并无缝执行。
 - 数据分析与解释：创建结构化和非结构化数据的知识库、分析和解释数据，根据其发现做出预测和建议。
 - 用户个性化和参与度：使企业能够为客户提供个性化的服务，并实时预测和解决担忧。
 - 业务效能：支持新服务和功能。

图表：NVIDIA AI 解决方案

图表：NVIDIA AI 平台在计算、软件和 AI 模型方面进行全栈创新



结语

Token产业就像横断山脉深处的卡瓦格博——基模、算力与应用各踞一面山坡，所见皆为同一座山的某个侧影，无人能够一次窥其全貌。

这座山至今无人登顶，却让所有人心向往之。每一次研究都只是又一次接近；所以我们仍会带着新的数据与认知，一次次回到这里。山不语，瞭望即是抵达。



人会反复回到一个地方，
或许因为那里藏着他尚未完全理解的自己。

——《横断浪途》

风险提示

- **AI 等底层技术变革不及预期。**计算机产业发展的底层驱动力是技术的变革，若AI等技术变革不及预期将影响计算机产业业绩释放周期。
- **下游客户IT支出意愿与力度不及预期。**宏观经济当前仍处于修复阶段，下游行业整体景气度仍承压，存在下游客户IT支出意愿与能力不及预期风险。
- **政策落地不及预期。**政策对计算机产业供需两端均有重要促进作用，若以自主可控为代表的相关国产化政策落地时间和扶持力度不及预期，则可能会对计算机产业的业绩产生不利影响。
- **行业竞争加剧。**目前国产软硬件尚未呈现出清晰的格局，很多领域仍处于高度竞争状态，若后续行业竞争加剧，可能会影响计算机公司的毛利率与净利率水平。
- **第三方数据失真风险。**本报告部分数据引自第三方机构与公开渠道，若相关数据的统计口径、采集方法或披露内容存在偏差或失真，可能影响本报告相关分析结论的准确性。
- **市场规模测算偏差风险。**本报告中相关市场规模测算基于一定的前提假设，若假设条件发生变化或与实际情况存在差异，测算结果可能与真实市场规模存在偏差，仅供参考。
- **研报信息更新不及时的风险。**

重要声明

- 中泰证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员会许可的证券投资咨询业务资格。本公司不会因接收人收到本报告而视其为客户。
- 本报告基于本公司及其研究人员认为可信的公开资料或实地调研资料，反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。本公司力求但不保证这些信息的准确性和完整性，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，可能会随时调整。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。
- 市场有风险，投资需谨慎。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。
- 投资者应注意，在法律允许的情况下，本公司及其本公司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投资银行、财务顾问和金融产品等各种金融服务。本公司及其本公司的关联机构或个人可能在本报告公开发布之前已经使用或了解其中的信息。
- 本报告版权归“中泰证券股份有限公司”所有。事先未经本公司书面授权，任何机构和个人，不得对本报告进行任何形式的翻版、发布、复制、转载、刊登、篡改，且不得对本报告进行有悖原意的删节或修改。