仅供内部参考，请勿外传

Token 经济学：Token 工厂价值锚重构，价格分层与需求跃迁共振

计算机

评级： 增持

姓名	电话	邮箱	登记编号
杨林(分析师)	021-23183969	yanglin2@gtht.com	S0880525040027
魏宗(分析师)	021-23180000	weizong@gtht.com	S0880525040058
钟明翰(研究助理)	021-38031383	zhongminghan@gtht.com	S0880124070047

本报告导读：

Token 正在成为 AI 产业统一计量基准，推动 AI 商业模式从模型能力展示走向成本、效率和价值的精细化核算。我们认为，随着 Token 价格长期下行、Agent 应用放大调用需求、Token 工厂推动算力资源工业化生产，AI 产业机会将从单一模型竞争延展至算力基础设施、模型服务和场景应用的系统性机会。

投资要点：

- **Token 成为 AI 产业价值锚，统一计量框架加速形成。**词元（Token）是大语言模型开展信息解析与内容生成的基础粒度单元。Token 计量体系与 AI 五层产业架构相辅相成，共同构筑 AI 产业价值拆分、盈利定位与效率核算的核心分析范式。相较于模型参数、算力峰值、硬件装机规模等传统指标，Token 具备技术适配性突出、商业通用性较强和产业协同属性显著等优势，有望成长为贯穿 AI 全产业链的标准化度量基准。
- **Token 定价体系逐步成熟，成本、能力与商业策略共同推动价格分层。**输入、缓存与输出构成 AI 调用成本的核心变量，差异化计费反映不同推理环节的成本结构和价值属性。横向比较来看，中美模型价格差距显著，中国高性价比模型正在加速全球 Token 调用渗透。展望后续，Token 成本长期下行确定性较强，但价格体系将进一步走向能力分层：定价锚有望从 GPU 逐步迁移至能源和知识密度，硬件迭代、推理优化与模型效率提升将推动单位 Token 成本持续下降；与此同时，低端 Token 趋向商品化，高价值 Token 则凭借能力稀缺性具备涨价基础，计费模式也将从单一按量收费演进为订阅、结果导向与算力金融化并行。
- **Token 调用量进入指数级增长阶段，需求驱动正在从人类交互转向机器流程。**国内 Token 调用量高增叠加中国模型份额提升，高性价比模型成为需求放量的重要驱动力。编程和复杂任务的 Token 消耗占比持续提升，反映大模型使用形态正在从轻量交互向重负载工作流迁移。更重要的是，Agent 将 AI 调用从单轮问答推向长链条任务执行，使 Token 消耗随任务拆解、模型调用、工具使用和上下文积累快速放大，显著提升单次任务的 Token 消耗强度。需要注意的是，Token 消耗量不等同于商业价值，未来判断 Token 经济质量不能只看调用规模，更应关注场景质量、业务结果和单位 Token 产出的真实经济意义。
- **Token 经济打开 AI 基础设施、模型服务和应用场景的系统性投资机会。**Token 需求放量正在重塑 AI 产业供需关系，算力、模型和应用均迎来价值重估。推理需求快速增长加剧算力供给约束，高可用基础设施成为产业瓶颈；Token 工厂将算力中心升级为智能生产线，生产效率成为 AI 竞争新焦点；Token 出海则有望推动低成本能源和模型能力转化为全球智能服务供给。
- 风险提示：AI 应用商业化落地不及预期，算力供给、能源资源及基础设施建设不及预期，模型价格竞争及技术路线变化。

相关报告

计算机《推理与 Agentic AI 浪潮下，CPU 重回 AI 基础设施核心中枢》2026.05.29

计算机《英伟达 GTC 前瞻：聚焦 Rubin 落地、Feynman 前瞻与基础设施重构》2026.03.11

计算机《君逸数码战略投资并签约银河通用机器人》2026.02.13

计算机《银河通用“干活”机器人登春晚，具身智能迎来主流时刻》2026.01.25

计算机《上海发布“模速智行”行动计划，自动驾驶产业驶入加速赛道》2026.01.18

仅供内部参考，请勿外传

目录

1. Token 成为 AI 产业价值锚 3
2. Token 定价进入分层阶段 4
2.1. Token 成本拆解：输入、缓存与输出共同决定 AI 调用成本 4
2.2. 中国高性价比模型加速全球调用渗透 5
2.3. Token 成本呈现长期下行趋势，价格体系按能力分层 7
2.3.1. Token 定价锚向能源和知识密度迁移，成本约束动态变化 7
2.3.2. 硬件迭代、推理优化与模型效率提升共同推动 Token 成本下降 8
2.3.3. 高价值 Token 凭借能力稀缺性具备涨价基础 10
2.3.4. Token 计费将从单一按量收费演进为多层级混合定价体系 12
3. Token 调用量进入指数级增长阶段 12
3.1. 高性价比驱动中国模型需求放量 12
3.2. Token 应用走向复杂工作流，编程消耗占比不断提升 14
3.3. Agent 将 AI 调用推向长链条任务，显著提高 Token 消耗 16
3.4. 场景质量和业务结果决定 Token 真实经济意义 19
4. Token 经济打开 AI 基础设施、模型服务和应用场景的机会窗口 20
4.1. Token 需求放量使算力、模型和应用迎来价值重估 20
4.1.1. 推理需求快速增长加剧算力供给约束 20
4.1.2. Token 工厂标志 AI 竞争从模型能力转向生产效率 22
4.1.3. Token 出海打开数字贸易空间 23
5. 风险提示 24

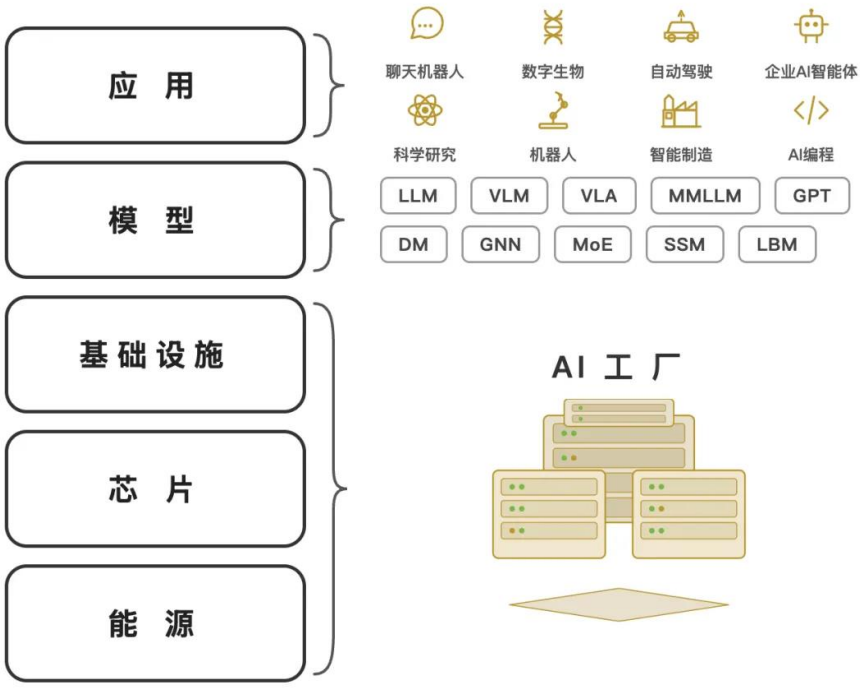
仅供内部参考，请勿外传

1. Token 成为 AI 产业价值锚

词元（Token）是大语言模型开展信息解析与内容生成的基础粒度单元。2026 年 3 月 23 日，国家数据局正式敲定 Token 官方中文标准译名为词元，并将其战略定位为智能时代的价值锚点。词元可作为串联技术供给端与商业需求端的标准化结算载体，为 AI 相关商业模式规模化落地构建了可量化、可计价的底层支撑体系。从产业维度看，词元可类比为驱动 AI 产业有序运转的数字电力，同时亦是 AI 商业化服务计费与清算的核心基准单元。若将大语言模型视作智能化算力终端设备，词元即为该设备运行过程中消耗的基础算力当量，也是行业对外商业化服务按量计费的核心依据。

Token 词元计量体系与 AI 五层产业架构相辅相成，共同构筑了 AI 产业价值拆分、盈利定位与效率核算的核心分析范式。2026 年 3 月，英伟达 CEO 黄仁勋于 GTC 大会前夕发表署名文章《AI Is a Five-Layer Cake》（《AI 是一个五层蛋糕》），系统性提出 AI 产业“五层架构”模型，将产业链自下而上拆解为能源、芯片、基础设施、模型、应用五大核心层级。其中，能源、芯片、基础设施三层构成“AI 工厂”，核心职能为制造智能；模型层作为智能能力的核心载体，将算力转化成 AI 的输出；应用层则面向终端场景实现智能价值的商业化落地，完成价值闭环。从产业流转逻辑看，“五层蛋糕”本质是一条 Token 的“生产—消费”全链路：能源转化为电力，电力驱动芯片运行，芯片支撑数据中心算力调度，数据中心运行模型，模型为应用提供智能服务，最终应用向用户输出 Token 并完成价值交付。产业链各环节主体（发电厂、芯片厂商、数据中心运营商等）的商业活动，均围绕 Token 的生产、流转与消耗展开。“五层蛋糕”模型明确了 AI 产业的价值分布与盈利节点，而 Token 则确立了产业效率、成本与收入的统一计量基准。二者互为补充，共同构建起覆盖 AI 产业技术、商业与价值维度的完整分析框架，为产业链投资与价值判断提供指引。

图 1：Token 是贯穿 AI 五层产业架构的统一计量单位



资料来源：英伟达官方博客，腾讯研究院

Token 具备三大核心属性，有望成长为贯穿 AI 全产业链的标准化度量基准。当前行业传统评价体系多以模型参数、算力峰值、硬件装机规模等维度为核心指标，虽在细分赛道具备一定计量参考价值，但仍存在统计口径不统一、

仅供内部参考，请勿外传

商业化计费模式僵化等痛点短板。相较传统评价维度，Token 具备三大核心优势：

1）技术适配性突出，可对 AI 应用场景、大模型、云服务、算力基础设施及能源等产业各层级实际资源耗用实现精准量化；

2）商业通用性较强，可充当产业统一价值标尺，抹平业态差异、降低产业链交易成本与协作壁垒；

3）产业协同属性显著，作为通用价值载体可在 AI 产业链上中下游各环节自由流转，助力行业资源高效优化配置，驱动整体产业结构迭代升级。

伴随人工智能产业进入规模化落地阶段，依赖资本补贴、流量粗放投放的扩张模式逐步走向收敛，算力调度、模型运算、服务交付等隐性运营成本日益凸显。Token 可实现产业链各环节成本的显性化、标准化表征，以其为核心的 AI 产业经济生态正加速构建与成型。

2. Token 定价进入分层阶段

2.1. Token 成本拆解：输入、缓存与输出共同决定 AI 调用成本

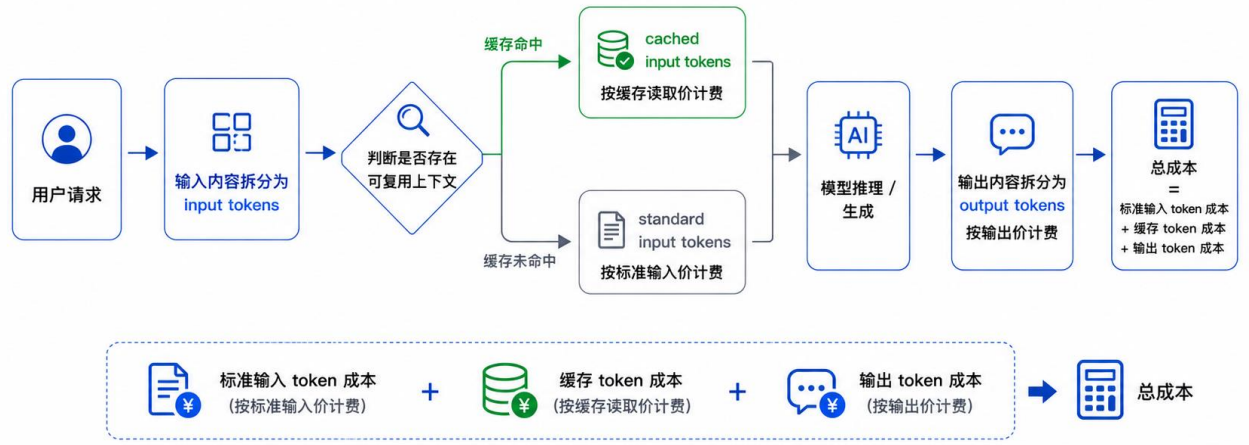
大模型标准推理可划分为输入 Token 与输出 Token 两大类型，二者在功能属性、业务场景及价值作用上存在明确差异。输入 Token 为用户向模型提交的原始交互内容，扩充输入体量可丰富上下文信息维度，助力模型生成更高匹配度的应答内容；在 RAG 检索增强生成场景下，外部检索知识库文本会并入 prompt 链路，进一步抬升输入 Token 规模。输出 Token 为模型自主推理生成的应答内容，更大的输出 Token 容量能够支撑模型输出更详实、完整的结果，同时适配思维链等多步复杂推理范式。此类链式推理会产生大量中间过程 Token，且多数中间 Token 对终端用户不可见，形成隐性 Token 消耗。

海内外大模型厂商普遍实行输入与输出 Token 差异化定价，价差背后由成本结构、价值属性与商业策略三重逻辑共同支撑。纵观 OpenAI、Anthropic、Google 等海外头部闭源模型定价体系，输出 Token 单价普遍为输入 Token 的 5–8 倍；国内厂商整体价差进一步收敛，多数模型维持 3–5 倍区间，DeepSeek 等高性价比模型更是收窄至约 2 倍水平。从成本端看，输出 Token 采用自回归逐词生成模式，运算特征为内存带宽密集型任务（memory-bound），GPU 大量时间闲置等待显存数据调度，实际有效计算占比仅 1%–5%；而输入 Token 为并行理解处理，属于计算密集型（compute-intensive）任务，硬件利用效率显著更高，天然形成成本分层。从价值端看，输入 Token 仅为基础原始素材，输出 Token 承载模型推理、解答、创作等核心价值产出，对应原材料与成品的价值溢价逻辑。从策略端看，低价输入可引导用户扩充上下文长度，避免过度精简提示词拉低生成质量，通过差异化定价实现用户行为引导与产业整体运行效率优化。

缓存命中计费价格显著低于未命中场景。大模型推理调用过程中，多轮对话、同一内容重复提问等场景普遍存在输入信息重叠。上下文缓存（Context Cache）技术可对推理请求的公共前缀进行存储缓存，留存历史任务指令、输入文本及模型生成结果。当用户发起相似推理请求并实现缓存命中时，模型无需重复运算与生成，仅产生极少量新增 Token 消耗；若缓存未命中，模型则需从头完成全流程推理，Token 消耗规模大幅增加。由于 Token 消耗量直接决定计费水平，缓存命中场景 Token 耗用远低于未命中场景，因此对应定价更具成本优势；同时该机制还可提升模型响应速度，在不降低回复质量的前提下，进一步减少用户综合使用成本。

仅供内部参考，请勿外传

图 2：输入、缓存与输出共同决定 AI 调用成本

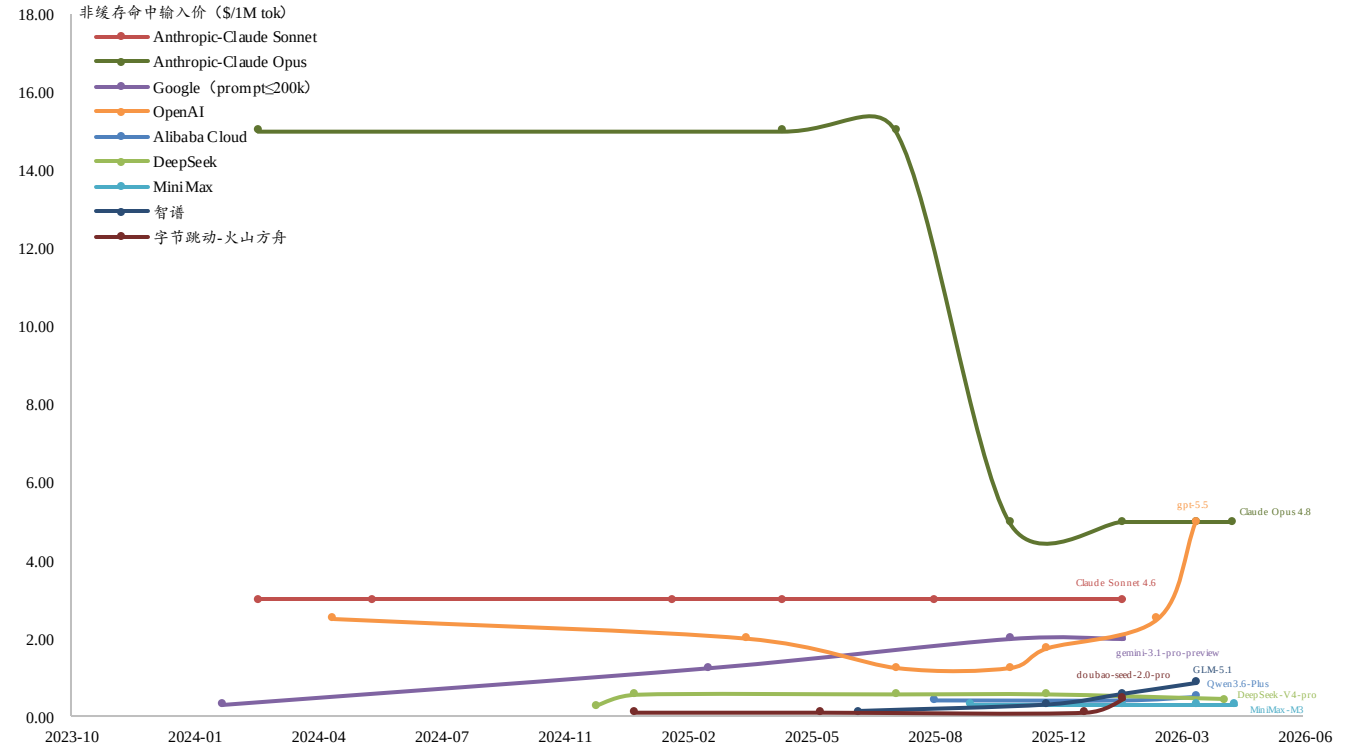


资料来源：国泰海通证券研究

2.2. 中国高性价比模型加速全球调用渗透

中美主流大模型的价格存在巨大差距。以非缓存命中输入价格为例，中国模型中，DeepSeek-V4-pro 为每百万 Token 1.74 美元，MiniMax M3 为 0.30 美元，doubao-seed-2.0-pro 为 0.47 美元（输入长度[0,32]k）。美国模型中，Anthropic Claude Opus 4.7 为 5.00 美元，Claude Sonnet 4.6 为 3.00 美元，OpenAI GPT-5.5 为 5.00 美元。美国主流模型的输入价格约为中国主流模型的 2 至 16 倍。输出价格差距更为明显。中国模型方面，DeepSeek-V4-pro 为每百万 Token 3.48 美元，MiniMax M3 为 1.20 美元，doubao-seed-2.0-pro 为 2.33 美元（输入长度[0,32]k）。美国模型方面，Anthropic Claude Opus 4.7 为 25.00 美元，Claude Sonnet 4.6 为 15.00 美元，OpenAI GPT-5.5 为 30.00 美元。中美主流模型输出价格差距约为 5 倍至 25 倍。

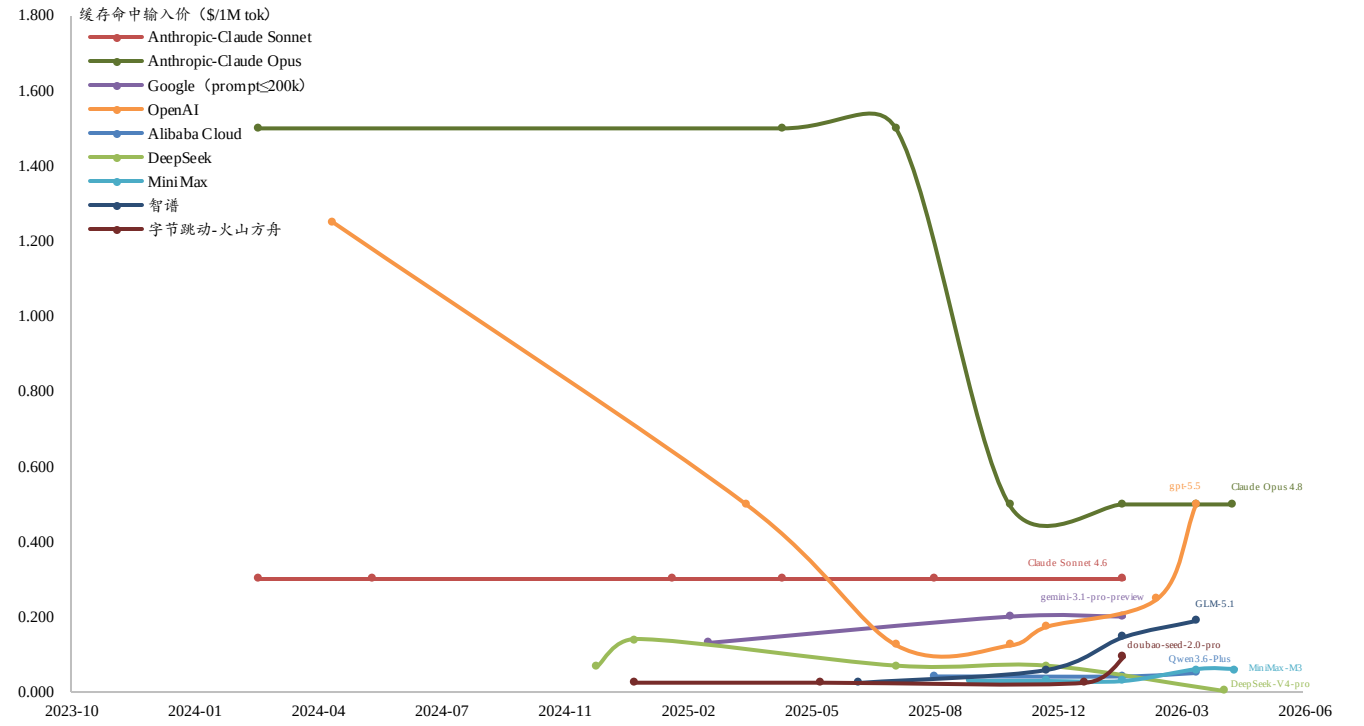
图 3：美国主流模型的非缓存命中输入价格约为中国主流模型的 2 至 16 倍



资料来源：Alibaba、Claude、DeepSeek、Google、MiniMax、OpenAI、智谱、火山方舟官网 API 定价页，国泰海通证券研究
注：Qwen、Gemini、GLM 和 doubao 系列 Token 采用阶梯报价，图表中数据为各模型第一阶梯报价。智谱和火山方舟的报价初始单位为人民币，按 USD/CNY=6.8722 统一换算为美元口径

仅供内部参考，请勿外传

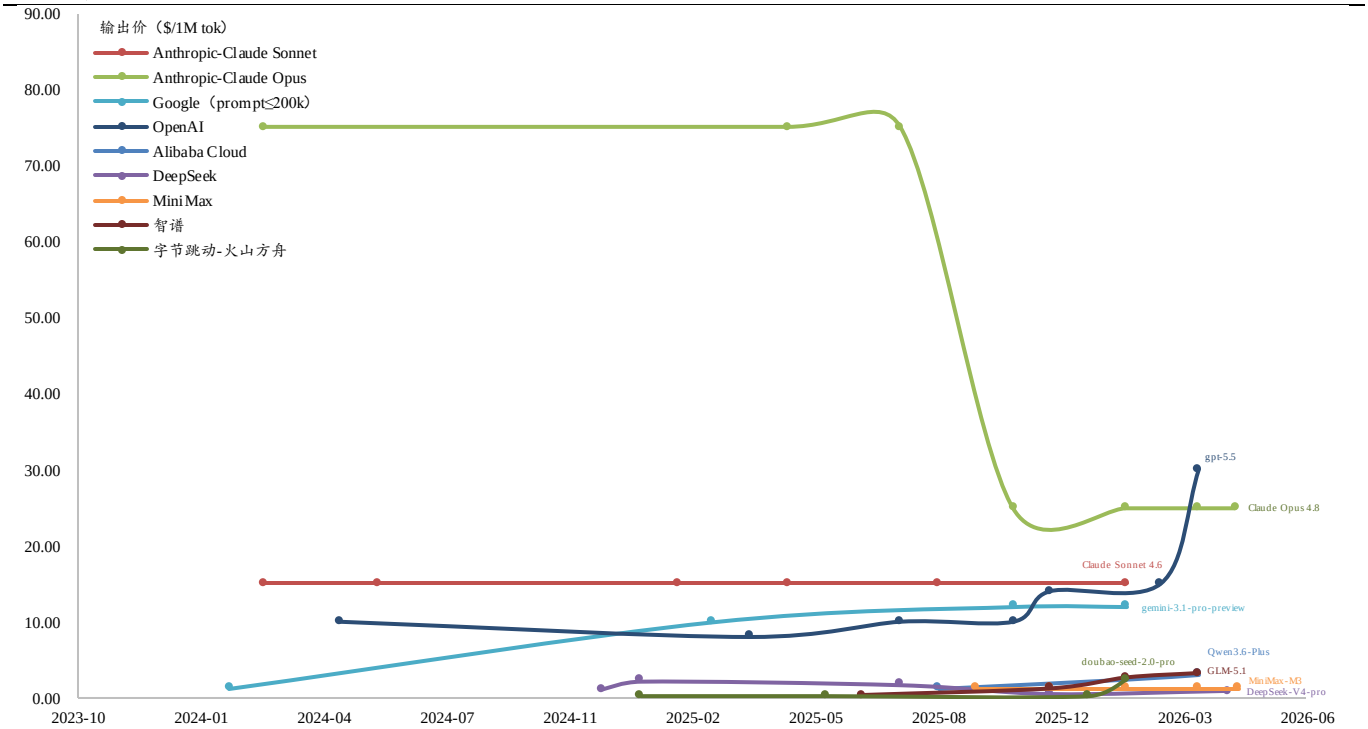
图 4：美国主流模型缓存命中输入价格显著高于中国主流模型



资料来源：阿里巴巴、Claude、DeepSeek、Google、MiniMax、OpenAI、智谱、火山方舟官网 API 定价页，国泰海通证券研究

注：Qwen、Gemini、GLM 和 doubao 系列 Token 采用阶梯报价，图表中数据为各模型第一阶梯报价。智谱和火山方舟的报价初始单位为人民币，按 USD/CNY=6.8722 统一换算为美元口径

图 5：中美主流模型输出价格差距约为 5 倍至 25 倍



资料来源：阿里巴巴、Claude、DeepSeek、Google、MiniMax、OpenAI、智谱、火山方舟官网 API 定价页，国泰海通证券研究

注：Qwen、Gemini、GLM 和 doubao 系列 Token 采用阶梯报价，图表中数据为各模型第一阶梯报价。智谱和火山方舟的报价初始单位为人民币，按 USD/CNY=6.8722 统一换算为美元口径

从模型层级看，Token 单价通常与模型能力呈正相关。同一厂商产品矩阵中，从 Nano 级轻量模型到标准版/旗舰模型，每百万 Token 价格差异可达十倍量级。其底层逻辑在于，参数规模更大、推理能力更强、可处理任务复杂度更高的模型，通常对应更高的算力消耗、工程投入和服务价值。但横向

仅供内部参考，请勿外传

比较不同厂商时，前沿模型的 Token 定价并非单纯由能力或成本决定，而是受战略定位、商业模式、客户结构、成本约束和生态诉求共同影响。以 OpenAI、Anthropic 和 Google 为例，三者均可概括为“价值定价为核心、成本约束为底线、竞品价格为参照”的复合定价体系，但定价锚点有所差异：OpenAI 偏向以竞争格局为锚，通过低价入门模型扩大开发者生态，并以高阶模型承接盈利；Anthropic 更强调安全性、可靠性和企业级服务溢价；Google 则更重视 Gemini 与 GCP 云生态的协同价值。

中美大模型 Token 定价核心差异更多来自训练与推理成本结构，而非低毛利。智谱 2024/2025 年全年综合毛利率为 56.3%/41.0%；2025 年的 MaaS API 平台毛利率同比提升近 5 倍至 18.9%。MiniMax 方面，2025 年全年整体毛利率提升至 25.4%，较 2023 年的-24.7%和 2024 年的 12.2%持续改善；分业务看，2025 年前三季度 AI 原生产品毛利率仅 4.7%，开放平台及其他基于 AI 的企业服务毛利率达到 69.4%，显示企业/API 类服务毛利率显著高于 C 端 AI 原生应用。相比之下，OpenAI 曾向投资者表示，2025 年推理成本同比增长四倍，导致调整后毛利率从 2024 年的 40%降至 33%。Anthropic 也曾将 2025 年毛利率预期下调至约 40%，主要原因是推理成本上升。因此，从已披露口径看，中国大模型并未表现出显著低于美国大模型的毛利率水平，低 Token 定价更多来自更低的单位训练和推理成本，而非单纯牺牲利润率。DeepSeek 论文显示，DeepSeek R1 训练成本仅 29.4 万美元，显著低于美国同类模型开发成本。由此看，中国模型的价格优势更接近“低成本供给能力”的外化：工程效率、稀疏激活架构、国产化算力适配和开源竞争共同压低单位 Token 成本，使中国模型能够在维持一定毛利空间的同时，以更低 API 价格扩大开发者生态和全球调用渗透。

从成本端看，Token 的“真实成本”并不容易被精确计量。训练投入、推理算力、研发人员、数据采购、安全对齐、模型评测和运维保障等成本项目，均难以严格分摊至单个 Token，因此传统“成本加成定价”在大模型 API 场景下可操作性有限。更准确地说，成本更多是 Token 定价的底线约束，而非精确标尺。展望后续，随着前沿模型市场竞争加剧，头部厂商定价策略或逐步收敛，价格本身的博弈空间将被压缩，差异化将更多转向服务稳定性、模型安全性、企业级交付能力和生态价值。更长期看，计费模式也可能由“按 Token 计费”演进为“按任务计费”或“按结果计费”。一旦定价单位发生变化，当前围绕 Token 单价展开的比较框架，其参考意义也将相应下降。

2.3. Token 成本呈现长期下行趋势，价格体系按能力分层

2.3.1. Token 定价锚向能源和知识密度迁移，成本约束动态变化

Token 的定价底线产业阶段变化而动态迁移。Token 的生产成本主要由芯片、电力、数据与人才四类要素构成，但其定价底线并不存在单一、静态的成本锚，而是随产业阶段变化而动态迁移。这里的“锚”指成本结构中对价格下限形成最强约束的关键变量。如果把这条迁移路径沿时间轴展开，可以清晰地看到三个阶段：

1）**短期（1-2 年）：**芯片仍是 Token 成本的核心锚点。以当前 AI 数据中心主流使用的 Nvidia H100 为例，云租赁价格约为每小时 2.5–3.5 美元，其中电费占比约 10%–20%，水冷、网络、运维等成本占比较低，芯片成本在推理总成本中占比可达六至七成。因此，GPU 的可获得性直接决定 Token 供给能力与价格中枢。其传导链条表现为：上游 GPU 产能受限、交付周期较长；中游 OpenAI、Anthropic 等模型厂商受制于算力集群扩张节奏；下游 API 可输出 Token 规模受限，进而支撑 Token 价格维持高位，最终成本由终端 AI 用户承担。

2）**中期（3-5 年）：**随着芯片供给改善、模型架构优化和推理效率提升，芯片成本约束有望逐步弱化，电力将成为更刚性的成本边界。一方面，AI

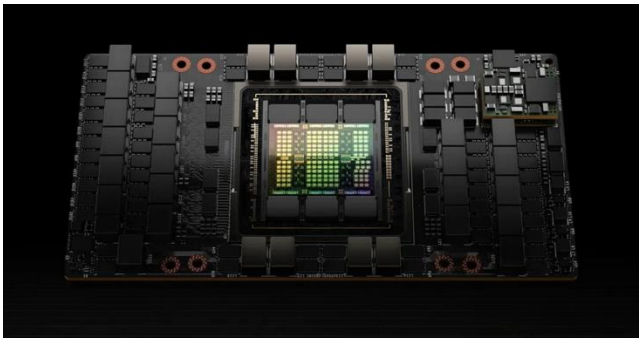
仅供内部参考，请勿外传

数据中心能耗快速增长，电力容量、并网资源和能耗指标将成为扩张约束；另一方面，电力成本具备更强的物理属性，压缩空间低于芯片与工程优化成本。需要强调的是，单位电量可产出的 Token 数并不由电力本身决定，而取决于所驱动的芯片类型、模型架构、负载状态和系统优化水平。同样一度电，用于高利用率输入处理与用于低利用率输出生成，其 Token 产出效率可能存在数量级差异。

3）**长期（5 年以上）：**人才与知识密度将主导定价。芯片和电力等物理成本终将被技术进步持续摊薄，而顶级模型训练所依赖的研究人才、工程经验和高质量数据积累更具稀缺性，将成为高端 Token 的核心定价锚。

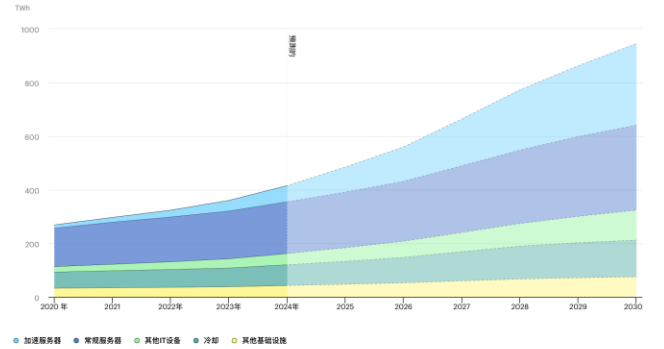
总体而言，Token 定价锚将沿“GPU→能源→人才/知识密度”的路径逐步迁移。

图 6：Nvidia H100 云租赁价格约为每小时 2.5–3.5 美元



资料来源：NVIDIA

图 7：全球数据中心耗电量快速增长



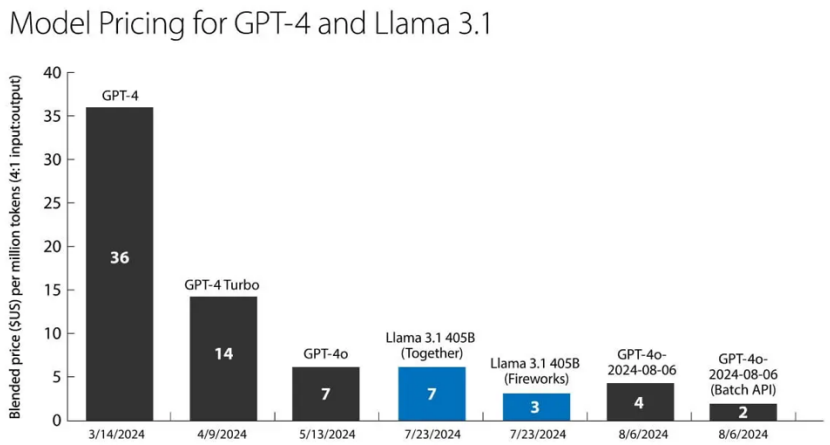
资料来源：IEA

2.3.2. 硬件迭代、推理优化与模型效率提升共同推动 Token 成本下降

当 GPU 供给瓶颈逐步缓解后，Token 价格中枢将进入长期下行通道。这一趋势可理解为摩尔定律在 AI 推理环节的延伸：单位算力成本下降、模型推理效率提升、服务端并发能力增强，共同推动单位 Token 成本持续压缩。2022 年，调用 GPT-3 级别模型需要 60 美元/百万 Token；GPT-4 在 2023 年发布初期 API 定价为每百万输入 Token 30 美元、输出 Token 60 美元；而到 2026 年初，同等或接近智能水平的模型价格已显著下探，2022-2026 年间部分模型降幅超过 99.9%。2026 年 6 月，为与 Anthropic 竞争市场份额，OpenAI 已开始评估大幅削减 Token 收费的方案。这意味着，AI 能力的普惠并不只依赖模型能力提升，更依赖 Token 成本快速下降。类比电力产业，低价电力推动了工业化和大众消费电气化；未来 Token 价格战也可能成为 AI 应用大规模普及的重要催化剂。

仅供内部参考，请勿外传

图 8：Token 价格下降趋势显著



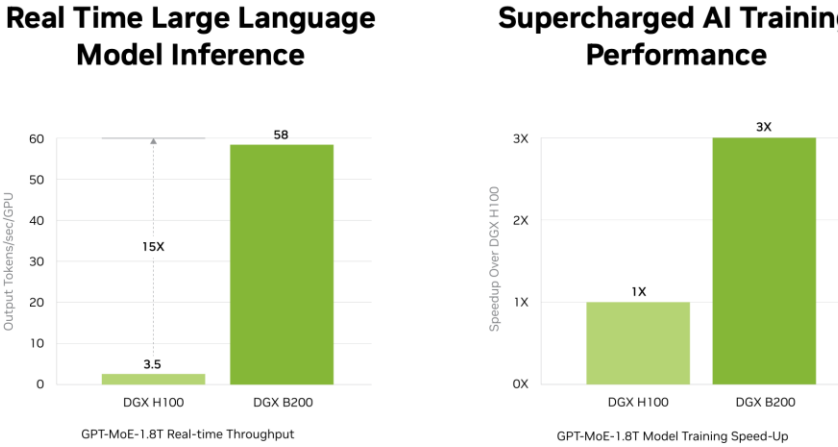
资料来源：DeepLearning.AI

Token 成本下降并非单点突破，而是硬件、软件、框架与模型架构共同作用的结果：

1) 硬件层面，GPU 性能持续提升且单位租赁成本回落。NVIDIA 官方数据显示，在 Megatron 530B chatbot inference 等特定大模型推理场景下，H100 相比 A100 集群最高可实现约 30X AI 推理性能提升；新一代 DGX B200 相比 DGX H100 可实现约 3X 训练性能和 15X 推理性能提升。
2) 软件层面，连续批处理、PagedAttention 等机制提升了 KV cache 利用率和请求并发能力；
3) 框架层面，vLLM、TensorRT-LLM 等推理优化工具进一步提升单卡吞吐；
4) 架构层面，MoE 模型通过稀疏激活降低了每个 Token 实际参与计算的参数量，例如 DeepSeek-V3 总参数量为 671B，但每个 Token 仅激活 37B 参数，从而在保持能力的同时降低推理计算开销。

在几者相乘效应的综合作用下，Token 成本每年下降 5-10 倍。同样规模的模型，从上一代 GPU 换到新一代，每个 Token 的能耗就能下降约 10 倍。

图 9：B200 相比 H100 可实现约 3X 训练性能和 15X 推理性能提升



资料来源：NVIDIA

模型本身在用更少的参数做到更多的事。过去，模型能力提升主要依赖参数规模扩张；但近两年，更高质量的数据、更成熟的训练方法和更优的架构设计，使较小参数规模模型也能逼近甚至在部分任务上超过上一代大模型。这意味着，同等能力不再需要同等规模的计算资源支撑。模型参数量下降不仅可以减少显存占用，更会降低推理过程中的数据搬运、显存访问和计算开销，使同一张 GPU 在相同时间内能够处理更多请求、生成更多 Token。换

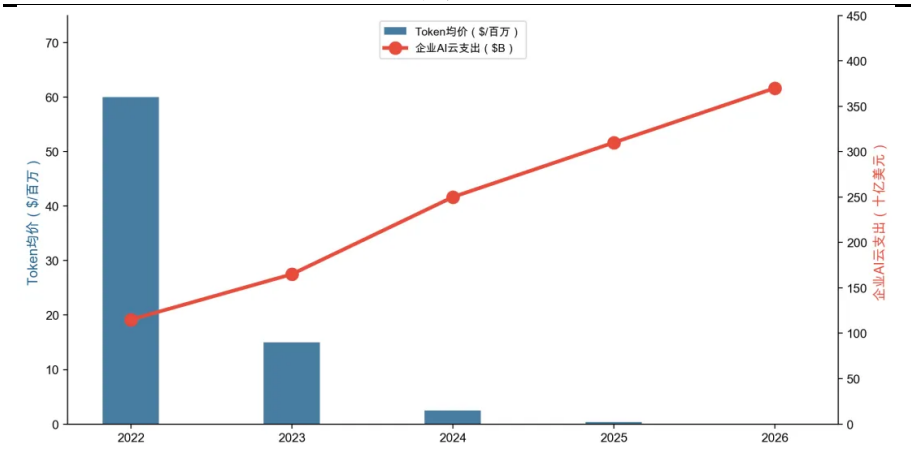
仅供内部参考，请勿外传

言之，Token 成本下降并不只来自 GPU 价格回落，也来自模型效率提升：在接近的任务效果下，模型所需调用的参数规模和推理计算量持续减少。

基础 Token 价格可能趋近于零，但 AI 的稀缺资源将转向更高层次。随着基础模型推理成本持续下降，通用 Token 在未来五年内存在实现经济意义上的趋零的可能，即多数用户在基础使用场景中不再显著感知 Token 成本。但真正的变化不在于价格本身，而在于商业模式切换：行业可能从单一“按 Token 计费”，逐步走向“免费基础能力+订阅增值+企业解决方案+云生态闭环”等多元模式。Google 推出更低成本的 Gemini Flash/Flash-Lite 系列，正是低成本高吞吐模型推动大规模调用的典型案例。需要强调的是，Token 成本趋零并不等于 AI 使用成本归零；当 Token 不再稀缺后，高质量数据、人类处理 AI 输出的注意力、AI 输出的可信度验证，将成为新的稀缺资源。

杰文斯悖论正在 Token 经济中重演：单位价格越低，总消耗和总支出反而可能越高。表面上看，Token 价格快速下降应当降低企业 AI 成本，但现实是，降价释放了大量此前被成本约束压制的潜在需求。2022-2026 年间 Token 平均单价降了 99.9%，但全球企业生成式 AI 支出从 115 亿美元增至 370 亿美元，增长约 3.2 倍；CreditSights 预计，前五大超大规模云厂商资本开支将从 2024 年约 2560 亿美元增至 2026 年约 6020 亿美元，其中约 75%投向 AI 基础设施；Google 的 token 处理量也在 18 个月内增长超过 130 倍。这背后的逻辑在于，当 Token 价格处于高位时，AI 推理主要服务于金融分析、药物研发、复杂代码生成等高价值场景；而当价格大幅下探后，代码审查、实时客服、个性化教育、智能办公、内容生成乃至 AI 角色扮演等大量场景开始具备经济可行性，每一个新场景都对应新的 Token 消耗入口。进一步看，用户购买的并不是 Token 本身，而是答案、决策建议或自动化任务结果。随着模型能力提升，单个高质量答案往往需要更长上下文、更复杂推理链、更大模型调用和更多轮交互，因此即便单个 Token 的成本持续下降，单次任务的总 Token 消耗和电力消耗仍可能上升。换言之，Token 降价并不是简单压缩市场规模，而是在扩大 AI 可用场景边界的同时，推动全社会 AI 调用量进入更高量级。

图 10：杰文斯悖论正在 Token 经济中重演



资料来源：Epoch AI，NavyaAI，Tom's Hardware，罗德岛大学 AI 实验室，各厂商 API 定价页，腾讯研究院

2.3.3. 高价值 Token 凭借能力稀缺性具备涨价基础

Token 价格体系正从单一降价逻辑转向分层定价。如果从单个 Token 单价扩展到整体供给侧视角来看，已持续数年的算力军备竞赛，未来将会遵循“短期紧缺—中期宽松/过剩—价格竞争—行业整合”的产业周期运行。当前全球科技巨头仍在持续加码 AI 数据中心建设：2025 年，Google、Microsoft、Meta 和 Amazon 合计 Capex 已达约 4100 亿美元，据《金融时报》汇总的第

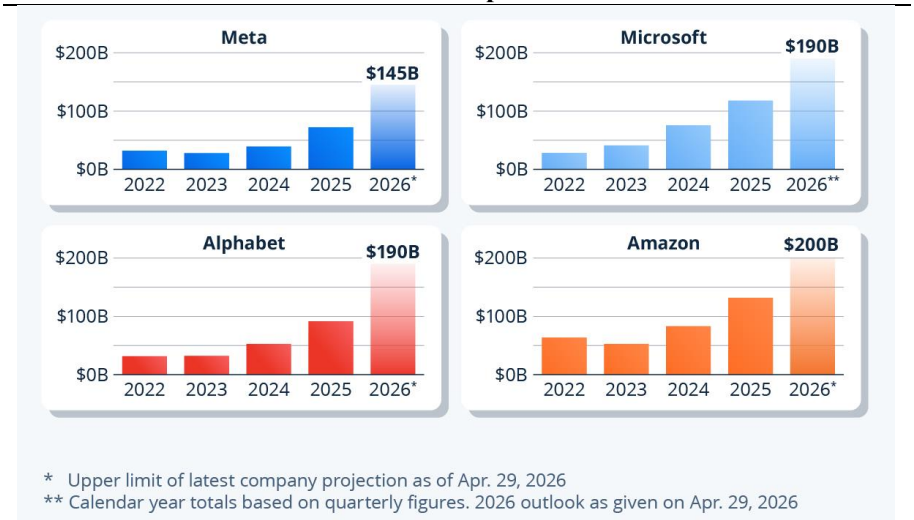
仅供内部参考，请勿外传

一季度财报显示，2026 年合计 Capex 将进一步升至约 7250 亿美元。考虑到 GPU 集群和数据中心从投资到落地通常存在两到三年建设周期，新增产能将在中期集中释放；而需求侧虽然受 Agent、自动化工作流等拉动快速增长，但企业采购与应用落地节奏相对平滑，容易与阶跃式释放的产能形成错配。中期若供需趋于宽松，Token 市场可能出现更清晰的分层：

1) **低端 Token 市场：**基础对话、简单生成和标准化任务对应的低端 Token 更接近大宗算力商品，竞争焦点在价格和可获得性，利润率可能被持续压缩，类似早期云计算价格战；
2) **高端 Token 市场：**复杂推理、专业知识和高可靠场景对应的高端 Token，更强调模型能力、稳定性、安全性和交付确定性，具备更强的价值定价基础。

其中，开源模型如 Llama、Mistral 等将成为重要的价格挑战者，它们未必会取代闭源巨头，但会为低端 Token 建立价格天花板，并将供给主体从少数大厂扩展至大量独立部署者，从而重塑 Token 市场结构。

图 11：四大科技巨头的 2026 年合计 Capex 将进一步升至约 7250 亿美元



资料来源：Meta、Microsoft、Alphabet、Amazon 2022-2025 年年报，Statista

近期大模型和云厂商涨价说明，高价值 Token 已具备价格上行动力。与“Token 长期降价”的直觉不同，2026 年以来部分模型和 AI 算力产品已出现阶段性涨价，背后反映的是高能力模型、Agent 场景和企业级算力资源的供需紧张。智谱 CEO 张鹏在 2025 年业绩说明会上表示，2026 年一季度智谱 API 调用定价提升 83%，但调用量仍增长 400%，市场呈现供不应求；智谱将 GLM Coding Plan 订阅价格上调至 60%，API 调用价格提升 67%-100%，以应对编码模型用户需求快速增长。腾讯云也在 2026 年 3 月调整智能体开发平台部分模型计费策略，输入与输出价格普遍涨幅超过 4 倍。火山引擎总裁谭待同样认为不能抛开模型的价值谈 token 的价格：随着模型能力不断增强，单 Token 经济价值会更高，理论上价格也会上涨。

表 1：AI 需求与硬件成本上升推动部分云厂商上调价格

时间	厂商	涉及产品/服务	涨价/调整内容
2026 年 2-3 月	智谱	GLM 系列 API/代码订阅套餐	2 月 GLM 系列 API 上涨约 30%，取消首购优惠；3 月整体 API 价格较 2025 年底提升约 83%
2026 年 3 月 13 日起	腾讯云	混元 Tencent HY2.0 Instruct/Think	混元系列模型输入/输出价格普遍上涨 4 倍以上，其中 Tencent HY 2.0 Instruct 输入价格从 0.0008 元/千 tokens 调至 0.004505 元/千 tokens，涨幅约 463%
2026 年一季度	智谱	智谱 API	2026 年一季度 API 调用定价提升 83%，但调用量仍增长 400%，市场供不应求

仅供内部参考，请勿外传

2026 年 4 月	智谱	GLM 系列 API	GLM 系列 API 价格再度上调约 10%
2026 年 4 月 28 日公告，5 月 15 日起执行	腾讯云	AI 编程助手 CodeBuddy / WorkBuddy	企业旗舰版由 78 元/人/月上调至 198 元/人/月，涨幅约 154%；企业专享版由 158 元/人/月上调至 316 元/人/月，涨幅约 100%

资料来源：36 氪，南方财经，科创板日报，钛媒体，华尔街见闻，国泰海通证券研究

2.3.4. Token 计费将从单一按量收费演进为多层级混合定价体系

Token 定价将从单一按量收费，演进为多层级混合定价体系。从 3-5 年维度看，行业可能经历“三浪叠加”的结构性迁移：

1) **按量计费与包月订阅并行（已在发生）：**API 按量计费与 C 端包月订阅并行，前者服务开发者和企业调用，强调弹性、可控和可计量，后者面向消费者产品，如 ChatGPT Plus，强调体验和使用权；
2) **结果导向定价（2-3 年内）：**随着 Agent 和自动化工作流成熟，定价单位将从“消耗多少 Token”转向“完成什么任务”，例如按合同生成、客服接待等结果计费，用户无需理解底层 Token 消耗，厂商则可通过模型效率提升获得更高利润率；
3) **金融化和衍生品市场（3-5 年内可能出现雏形）：**当 Token 或算力成为企业生产要素后，市场可能出现价格对冲需求，推动算力现货、算力互换甚至衍生品雏形出现。2026 年全国两会期间，全国人大代表、东方财富董事长提出，算力及相关要素价格波动风险正显著提升，建议积极开发算力期货等创新工具，探索成立算力交易所。算力现货方面，我国早有探索，2023 年 12 月份，“上海算力交易平台 2.0”正式发布；2025 年 5 月 21 日，中国算力平台（上海）暨上海算力交易平台正式启动。海外方面，CME 与 Silicon Data 计划于 2026 年下半年推出全球首个标准化算力期货产品，预计将为 AI 产业链提供有效的套期保值工具，规避价格涨跌带来的风险。

目前来看，完整的 Token 交易所仍面临标准化和监管两大约束，不同模型 Token 的智能水平、稳定性和可替代性差异显著，较难像原油一样直接同质化交易。因此更现实的路径，可能是先形成“标准化算力单位”交易，如 GPU 小时或等效算力包，再在其基础上叠加模型能力、服务稳定性和智能溢价，最终形成“基础层按量计费、应用层按结果/订阅计费、金融层提供对冲工具”的多层价格生态。

Token 或将重构跨境贸易模式。若将算力视作由电力、芯片和模型工程共同加工出的生产能力，Token 就是这一能力的标准化交付单位。以 H100 推理场景测算，1 度电理论上可产出约 923 万 Token，考虑 PUE 等散热与系统损耗后，实际产出约 550 万 Token；若按国内模型约 2 元/百万 Token 定价，则 1 度电通过 Token 形式可实现约 11 元收入，显著高于传统电力直接出口销售价格。电力出口与 Token 出口的核心差异在于，传统电力出口依赖跨境电网，存在输电损耗、基础设施投入和地缘政治约束；而 Token 出口依托通信网络即可完成全球交付，电力仍留在本国电网体系内，但其经济价值通过 AI 服务被全球用户消费。在这一框架下，低成本绿电、数据中心建设、模型效率和 API 分发能力共同决定一国 AI 服务的全球竞争力。例如中国西部绿电成本约 0.2 元/度，而欧美电价约 0.8-1.2 元/度，能源成本差异会进一步传导至 Token 成本和模型服务价格。更长期看，随着 AI 服务成为全球数字经济的重要基础设施，Token 有可能成为“能源-算力智能服务”链条中的核心计价单位，货币或金融体系未必真正锚定 Token，但 Token 对电力价值的重估和跨境变现能力，可能成为新一轮数字贸易的重要变量。

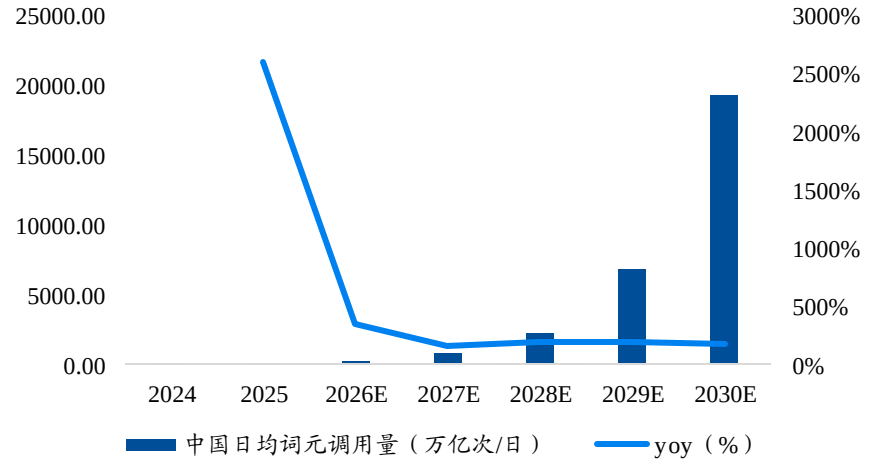
3. Token 调用量进入指数级增长阶段

3.1. 高性价比驱动中国模型需求放量

仅供内部参考，请勿外传

从行业景气度数据来看，国内词元日均调用量呈现指数级增长态势。2024 年初国内日均 Token 调用规模约 1000 亿级；2025 年底日均调用量大幅攀升至 100 万亿级；2026 年 3 月已突破 140 万亿级，两年时间实现千倍以上增长。数据高增背后，是以词元计费体系为核心的 AI 新型商业变现逻辑正加速成熟迭代。据艾媒咨询测算，2026 年我国词元日均调用量有望提升至 306.3 万亿次，2030 年将进一步增长至 19306.0 万亿次，2025-2030 年 CAGR 高达 210%，行业长期成长空间广阔。

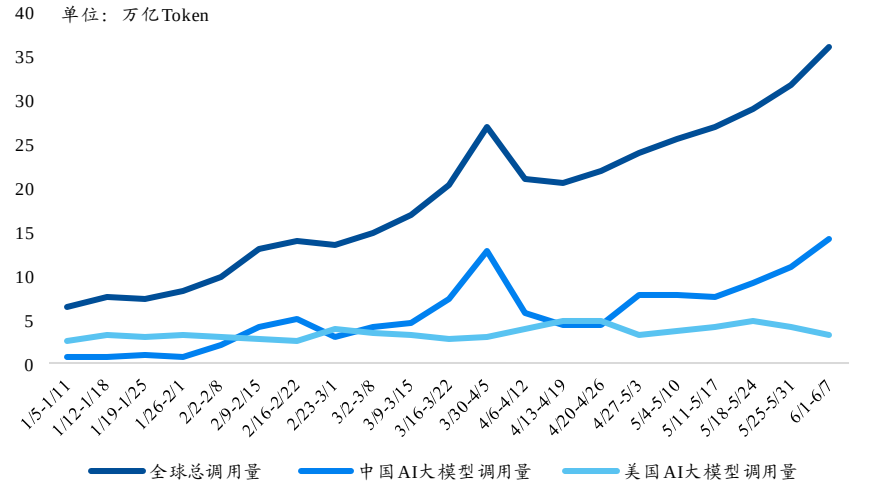
图 12：中国词元日均调用量呈现指数级增长态势



数据来源：艾媒咨询，国泰海通证券研究

全球大模型调用量延续高增，中国模型份额持续提升。根据 OpenRouter 周度数据，2026 年 6 月 1 日至 7 日，全球 AI 大模型总调用量达到 36.1 万亿 Token，环比增长 13.5%，已连续七周保持上行，反映大模型真实调用需求仍在加速释放。从地区结构看，上榜模型中，中国 AI 大模型周调用量达到 14.19 万亿 Token，环比增长 27.49%，连续三周实现增长；同期美国 AI 大模型周调用量为 3.2 万亿 Token，环比下滑 24.53%。中国大模型周调用量已连续六周超过美国，并稳居全球首位，说明中国模型在全球开发者生态和实际应用场景中的渗透率持续提升。整体来看，全球 Token 调用需求增长与中国模型份额提升同步发生，背后既有模型能力提升的驱动，也反映出高性价比模型对开发者调用行为的吸引力增强。

图 13：全球 Token 调用需求与中国模型份额同步提升



数据来源：OpenRouter，每日经济新闻，国泰海通证券研究

高性价比模型需求放大，推动中国头部模型调用量快速提升。从模型结构看，DeepSeek、Kimi、MiniMax 等中国模型在开发者生态中的使用量持续

仅供内部参考，请勿外传

提升，并呈现头部模型集中放量特征。其中，DeepSeek-V4-Flash 已连续三周位居榜首，周调用量达到 3.69 万亿 Token，环比增长 19%；腾讯 Hy3 preview 连续三周排名第二，周调用量达到 2.94 万亿 Token，环比小幅下滑 3%；6 月 1 日刚推出的 MiniMax M3 进入前三，周调用量达到 2.5 万亿 Token。据 MiniMax 官方介绍，MiniMax M3 是国内首个同时具备前沿 Coding 能力、1M 超长上下文和原生多模态能力的大模型。价格因素同样是重要催化：5 月 22 日，DeepSeek 官方宣布 V4-Pro 模型 API 价格将在 5 月 31 日结束 2.5 折优惠后，正式调整为原定价的四分之一。更低调用成本叠加强 Coding、Agent 和长上下文能力，进一步放大开发者使用意愿，也表明高性价比模型正在成为全球大模型调用增长的重要驱动力。

表2：国产高性价比大模型需求增长

大模型/指标	5 月 25 日–31 日调用量	6 月 1 日–7 日调用量	环比变化
DeepSeek V4 Flash	3.11T	3.69T	18.60%
Hy3 preview	3.03T	2.94T	-3.00%
DeepSeek V4 Pro	1.25T	1.70T	36.00%
DeepSeek V3.2	1.04T	1.17T	12.50%
Owl Alpha	1.66T	1.95T	17.50%
Claude Sonnet 4.6	1.92T	1.76T	-8.30%
Claude Opus 4.7	2.32T	1.44T	-37.90%
总调用量	**31.8T**	**36.1T**	**13.50%**

资料来源：OpenRouter，国泰海通证券研究

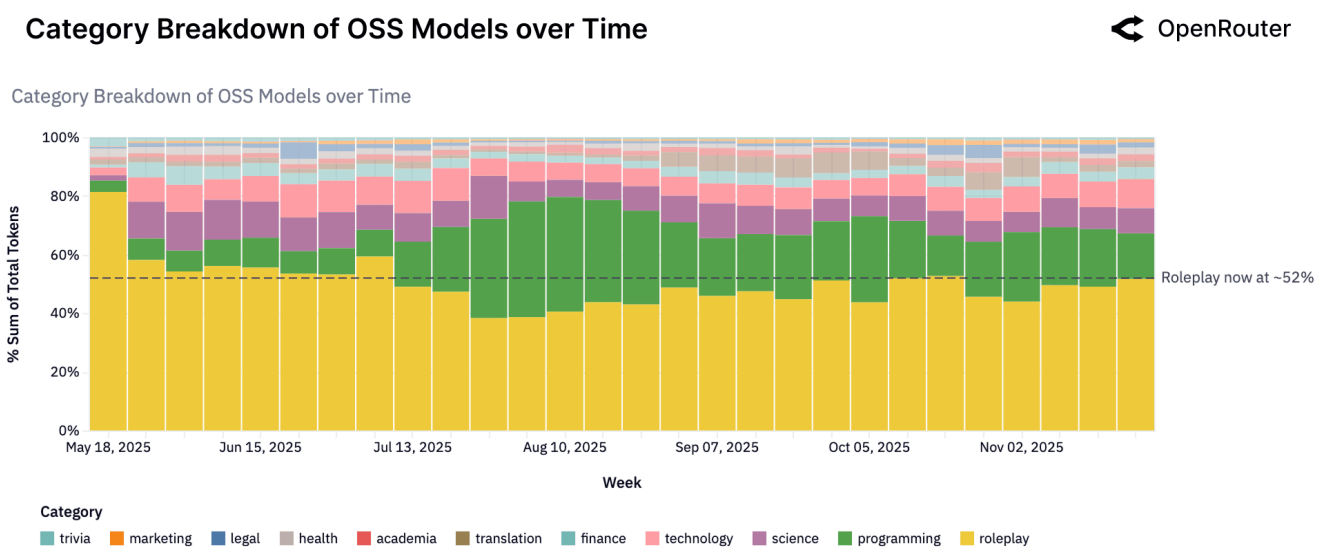
更多 Token 调用发生在私有云、本地和端侧场景中。当前公开可观测的 Token 数据主要来自公有云 MaaS/API 平台，但大量真实调用隐藏在企业私有化部署和端侧闭环系统之中。例如，金融机构出于数据安全和合规要求，往往会下载开源模型或采购私有化平台，在本地完成票据识别、风控审核、客服助手等任务；车端智能座舱为了保护用户隐私，部分对话与交互在车内闭环完成；机器人、工业视觉等场景中的 CV 小模型需要微秒至毫秒级响应，也很难依赖云端实时推理。对于传统政企客户而言，数据不出域、系统可控、审计合规仍是核心诉求，因此私有化部署通常是优先选项，并由此带动了大模型平台私有化厂商的发展。进一步看，企业基于开源模型进行后训练、微调和专属 Agent 构建，也会产生大量内部 Token 消耗，但这些调用并不会进入公有云统计口径。虽然当前主流 Agent 产品仍以云端交付为主，但在强敏感数据、强合规要求和高频调用场景下，客户会同时面临安全可控与综合成本压力，端侧和私有化方案的重要性因此提升。

3.2. Token 应用走向复杂工作流，编程消耗占比不断提升

角色扮演是 2025 年开源模型 Token 消耗占比最高的下游应用场景。OpenRouter 和 a16z 联合发布的《State of AI 2025》显示，开源模型的应用场景已覆盖创意、技术和信息类任务，其中角色扮演和编程是最主要的两个方向，二者合计贡献了开源模型 Token 使用量的大部分；其中，角色扮演占比约 52%，翻译、通用知识问答及其他应用占比相对较小。角色扮演在开源模型中占据主导地位，主要原因在于开源模型更适合承载创意表达，同时通常较少受到商业安全层和内容审核层约束，因而更容易应用于幻想娱乐、角色驱动体验、同人创作、互动游戏和模拟环境等场景。从产业视角看，角色扮演的高占比说明大模型消费侧需求并不完全由效率提升驱动，陪伴、叙事、沉浸式互动和个性化内容体验同样能够形成持续的 Token 消耗。因此，判断大模型下游应用空间时，除了办公、编程、搜索等显性生产力场景，娱乐互动、角色模拟和创意陪伴也有望成为高频、长时长、高粘性的 Token 消耗场景。

仅供内部参考，请勿外传

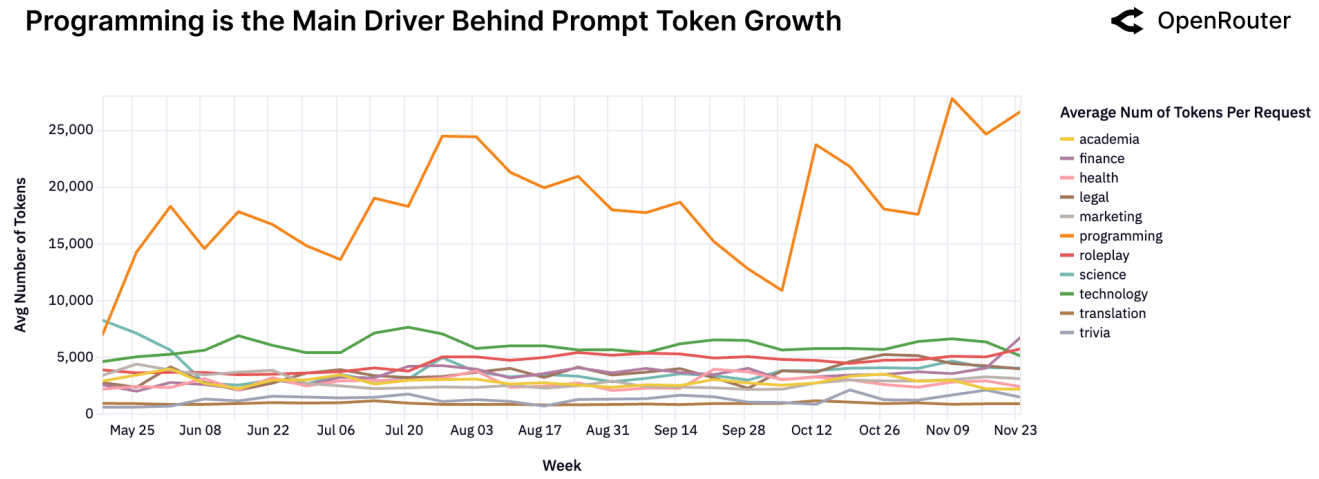
图 14：角色扮演（约占 52%）和编程占据开源模型 Token 消耗的主导地位



资料来源：OpenRouter

编程和复杂任务的 Token 消耗占比持续提升，反映大模型使用形态正在从轻量交互向重负载工作流迁移。《State of AI 2025》显示，2025 年模型工作负载结构出现明显变化，输入 Token 和输出 Token 均显著增长，平均 prompt 长度由约 1.5K 提升至 6K 以上，completion 长度也由约 150 提升至约 400。该变化表明，模型应用正在从简单开放式生成，转向上下文更长、任务链条更复杂的工作负载，模型能力定位也在从创意生成器逐步扩展为分析引擎。从分类数据看，编程任务是 prompt Token 增长的主要驱动力，代码理解、调试和代码生成类请求的输入长度经常超过 20K Token，而其他类别整体保持相对平稳且量级较低。报告还显示，过去 20 个月平均序列长度增长超过 3 倍，从 2023 年末不足 2,000 Token 提升至 2025 年末超过 5,400 Token，体现出长上下文、更深任务历史和复杂输出需求的结构性提升。编程相关 prompt 的平均长度约为通用 prompt 的 3—4 倍，说明软件开发工作流是长交互的重要来源；长序列的增加并非简单源于用户输入变长，而是更复杂、更嵌入式 Agent 工作流逐步形成的结果。

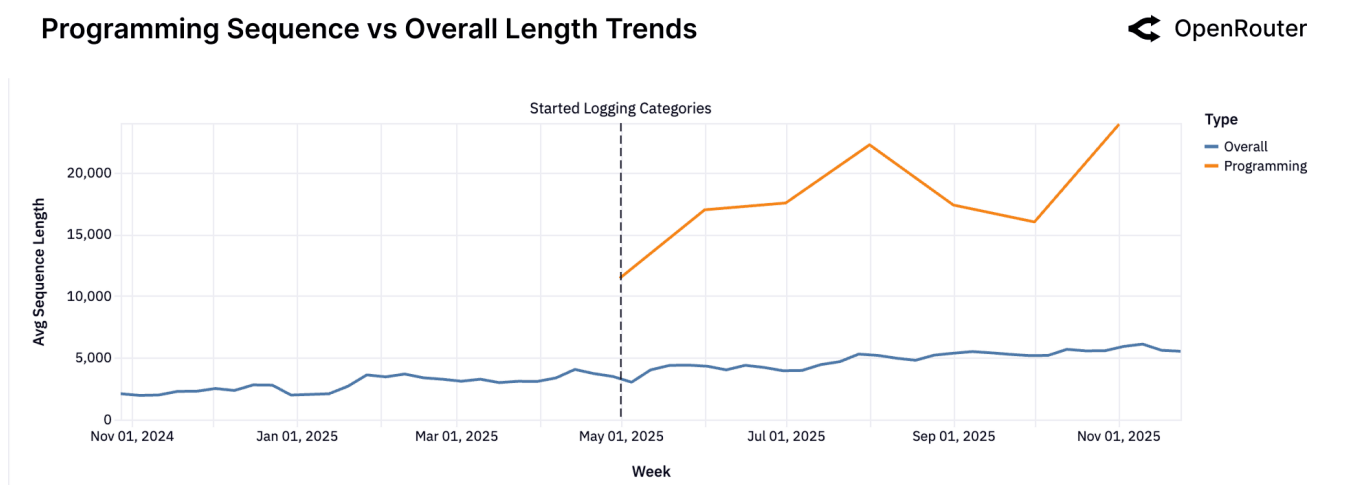
图 15：编程是 prompt Token 增长的主要驱动力



资料来源：OpenRouter

仅供内部参考，请勿外传

图 16：编程的序列长度增速远高于总体序列长度增速

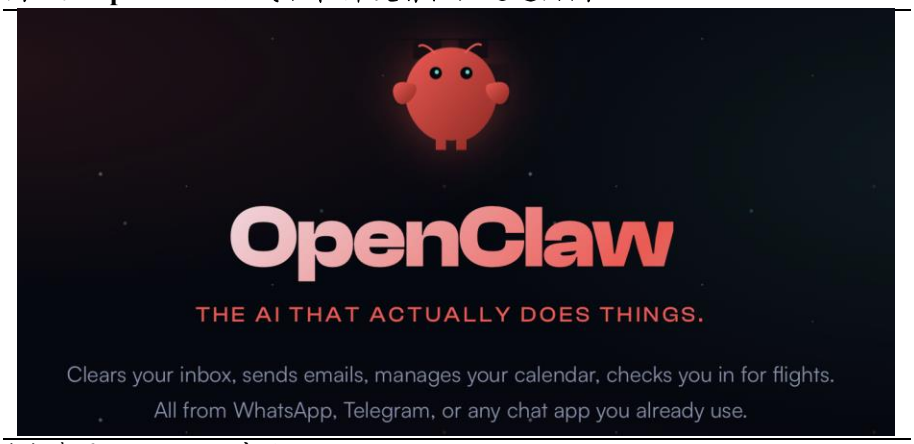


资料来源：OpenRouter

3.3. Agent 将 AI 调用推向长链条任务，显著提高 Token 消耗

Agent 应用开始成为 Token 消耗的重要增量来源。2026 年初，开源工具 OpenClaw 在开发者社区迅速出圈，并因名称谐音被称为“龙虾”。其上线后很快登顶 OpenRouter 应用排行榜第一，单周 Token 消耗超过 6000 亿。与传统“你问我答”式聊天不同，OpenClaw 属于典型 Agent 应用，可在电脑端自主完成编程、测试、文件管理等任务，用户无需逐步发出操作指令。也正因为其工作模式从“单轮问答”转向“多步骤执行”，Token 消耗不再停留在普通聊天的数百或数千级别，而是随任务拆解、模型调用、工具使用和上下文积累快速放大。Agent 带来的变化，本质上是大模型从内容生成工具升级为可执行任务的数字劳动力，进而推动 Token 需求从人类交互驱动转向机器流程驱动。

图 17：OpenClaw 上线后在开发者社区迅速出圈



数据来源：OpenClaw 官网

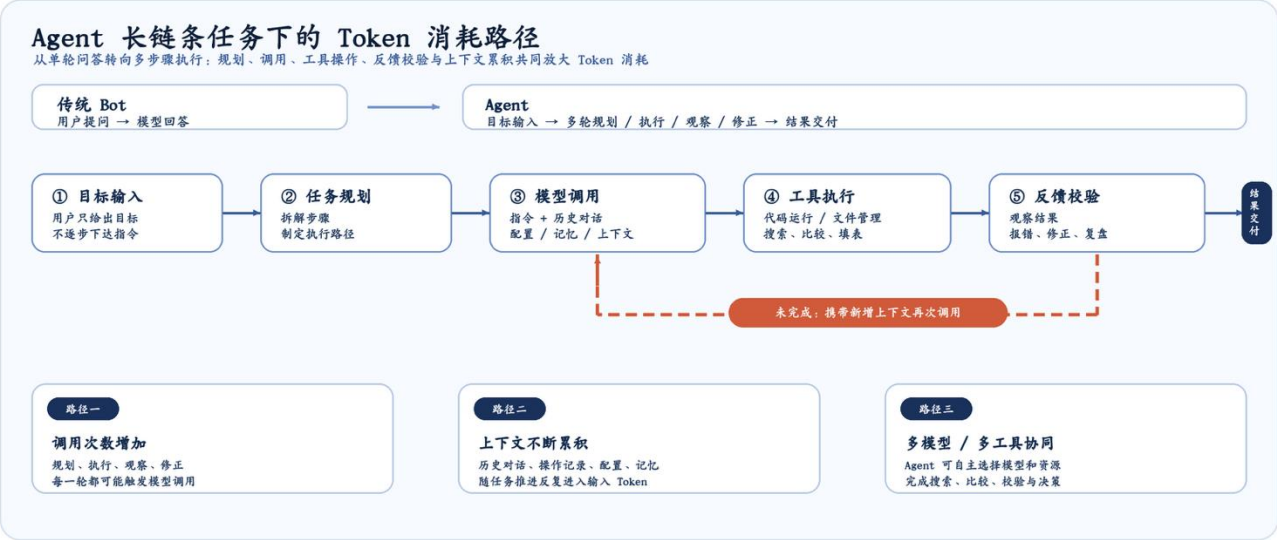
三重因素叠加，推动 Agent 成为 Token 调用量爆发的重要催化剂：

第一，Agent 将单次任务拆解为长链条流程，带动算力消耗呈数量级放大。在 Agent 出现之前，大模型更多承担简单问答、文案润色、基础咨询等轻量任务，单次交互短、上下文少，对算力需求相对有限。而 OpenClaw 等 Agent 普及后，大模型开始具备自主规划、自主执行和自主复盘能力，复杂任务往往需要多步骤协同完成。以编程任务为例，Agent 可能需要经历几十轮“写代码—运行—报错—修改—再运行”的循环，每一轮都会触发新的模型调用；同时，为了维持任务连续性，每次调用还要携带历史对话、此前操作记录和上下文信息，导致单次任务 Token 消耗从普通聊天的数百级别跃升至十万

仅供内部参考，请勿外传

甚至百万级别。更进一步，Agent 还可能在执行任务中自主决定调用哪个模型、消耗多少资源，例如机票比价场景中，几秒内调用多个模型完成搜索、比较、校验和决策，Token 市场由此出现一种新的买方：非人类买家。

图 18：Agent 长链条流程带动算力消耗呈数量级放大



资料来源：国泰海通证券研究

第二，Agent 将交互节奏从人类节奏压缩至机器节奏，显著提高调用频率。传统人机对话由用户逐轮输入驱动，每次交互之间往往存在较长停顿，分钟级间隔是可接受的，因此数百毫秒的响应延迟对整体体验影响有限。但 Agent 的运行方式完全不同，它可以在后台持续执行任务，智能体之间还可能高频联动、实时决策和不间断迭代，交互节奏被压缩至毫秒级。换言之，过去 Token 消耗受人类输入频率约束，而 Agent 场景下，调用频率开始由机器流程决定。只要任务尚未完成，Agent 就会持续进行规划、调用、观察和修正，Token 消耗从“用户每问一次产生一次调用”，变成“系统为完成目标持续触发调用”。这使得单位用户背后的真实 Token 消耗被显著放大，也使低延迟、高并发和高稳定性的推理基础设施成为 Agent 规模化落地的前置条件。

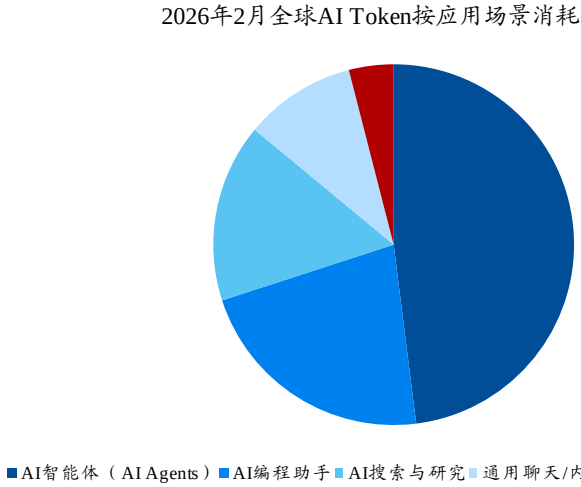
第三，Agent 对长时间连续运行的稳定性要求更高，推动高端推理资源需求上升。传统大模型交互通常是短时、单次、可中断的，偶尔出错或重新生成并不会显著影响整体体验；但 Agent 面向的是长程连续工作，一旦中途失败，可能导致整个任务链条重跑，甚至引发业务结果错误。业内主流 GLM5.1 模型在专业基建支撑下可实现单次持续工作 8 小时，这类长时间、高负荷、低中断的运行模式，对算力调度精度、系统稳定性和故障容错能力提出更高要求。由此看，Agent 并不只是“让 Token 用得更多”，还会改变 Token 供给质量要求：低端对话场景看重价格和可用性，高端 Agent 场景则同时看重推理速度、上下文承载能力、连续运行稳定性和系统级调度能力。

Agent 应用正在将 Token 消耗从线性增长推向数量级放大。以 OpenClaw 为代表的智能体应用快速普及后，OpenRouter 平台 Token 调用量出现明显抬升：2025 年 3 月 3 日至 9 日当周，OpenRouter 前十大模型周调用量合计约 1.24 万亿 Token；到 2026 年 2 月 16 日至 22 日当周，前十大模型周调用量已超过 8.7 万亿 Token，较前述阶段增长近 7 倍。其背后的核心原因在于，Agent 并非一次性问答，而是围绕目标持续进行任务拆解、模型调用、工具执行和结果校验。行业测算显示，在完成同一业务目标的情况下，Agent 模式消耗的 Token 约为 Bot 模式的 50—200 倍；智能体工作流还可能使单个任务调用模型 50—500 次，使用户可见输出背后的 Token 消耗增加 1—2 个

仅供内部参考，请勿外传

数量级。由此看，Agent 不仅提升了单用户 Token 消耗强度，也将企业 AI 成本从“按人类交互计量”推向“按机器流程计量”，进一步放大企业级 AI 总账单。

图 19：OpenClaw 上线后，Agents 成为 Token 调用量占比最大的下游应用



数据来源：OpenRouter，IDC，Gartner，弗若斯特沙利文，每日经济新闻，国泰海通证券研究

Token 需求曲线将呈现阶段性跃迁。随着 Agent 和自动化工作流逐步落地，Token 消耗不再只是用户数量扩张的自然结果，而会受到“机器自主调用”这一新变量影响。其长期演化路径可以概括为“双 S 曲线叠加 → J 型爆发与延续→ 次线性收敛”：

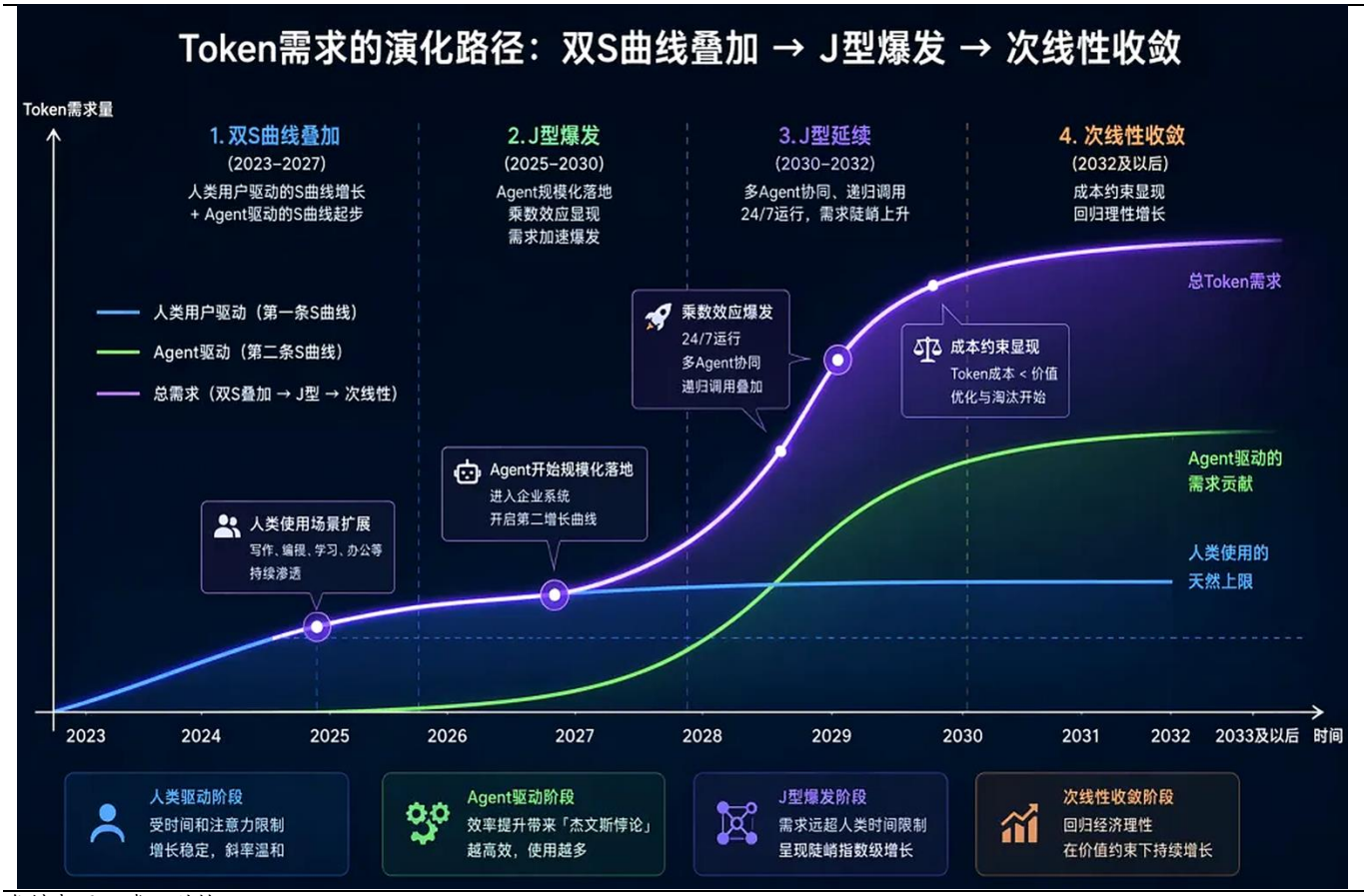
1) **第一阶段（2023—2027 年），Token 需求沿人类用户 S 曲线稳步增长。**AI 工具的用户渗透率和使用频次持续提升，推动 Token 消耗从早期低基数阶段向规模化应用阶段过渡。但这一阶段的增长仍主要受人类行为约束，单个用户每天能够发起的任务数量、交互时长和注意力投入存在上限。因此，Token 需求虽然会伴随 AI 应用普及而持续提升，但整体斜率相对可控，更接近互联网工具渗透过程中的 S 型增长。

2) **第二阶段(2025—2032 年),AgentS 曲线叠加后带来 J 型爆发与延续。**Agent 开始从试用走向企业级部署，Token 需求的主导变量将从“人类发起请求”转向“系统自主执行任务”。这一变化意味着 Token 消耗不再按用户交互次数线性增长，而是随任务链条、自动调用频率和流程复杂度放大。尤其在 2026—2028 年，若 Agent 渗透率进入快速提升阶段，Token 需求可能出现一到两个数量级的跃迁，形成明显的 J 型增长。

3) **第三阶段（2032 年以后），Token 增长将逐步进入次线性收敛。**随着 Agent 应用进入成熟期，企业会更加关注单位 Token 的投入产出效率，而不是单纯扩大调用量。能够持续创造收入、节约成本或提升效率的 Agent 场景将继续扩张；低价值、低转化、过度消耗 Token 的流程则会被优化或淘汰。因此，Token 需求在经历高速扩张后，增速大概率会逐步回落，从爆发式增长转向由 ROI、算力供给和成本约束共同决定的次线性增长。

仅供内部参考，请勿外传

图 20：Token 需求曲线将呈现阶段性跃迁



资料来源：腾讯科技

3.4. 场景质量和业务结果决定 Token 真实经济意义

Token 表观消耗量不等同于真实需求量。在 AI 转型逐渐成为企业组织升级不可或缺的阶段，Token 调用量容易被简单纳入员工考核、部门 AI 渗透率或业务数字化进度指标，由此产生“Token 伪需求”。例如，Meta 和亚马逊曾建立内部 AI 工具使用排行榜，但因部分员工通过不必要操作人为提高得分、推高算力成本，亚马逊已于 2026 年 5 月叫停相关榜单。腾讯部分业务过去也会统计各部门、团队的 Token 用量并排名，一些员工担心 Token 消耗不够充分，在工作时间搭建无实际意义的工作流、让 Agent 重复完成任务，甚至处理个人需求。此类消耗表面上抬高了平台调用量，但未必对应真实业务需求，也不一定带来可衡量的效率提升或收入贡献。因此，分析 Token 经济时不能只看调用量绝对规模，而应进一步区分真实业务调用、探索性试用、考核驱动消耗与无效冗余调用，否则容易高估 AI 渗透进度和商业化质量。

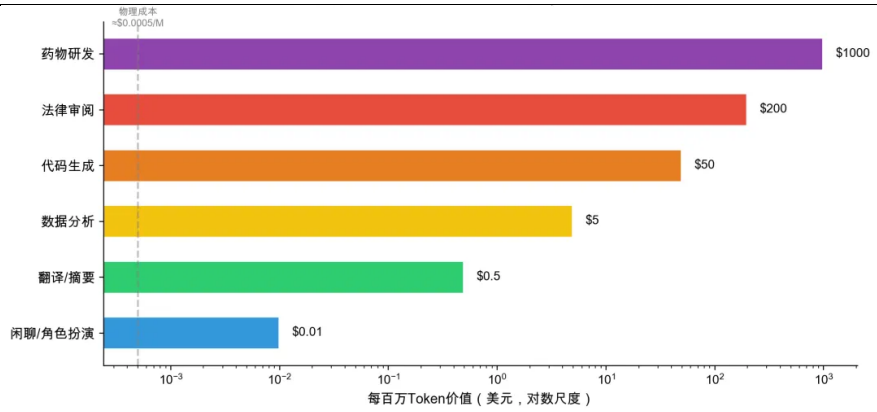
企业内部 Token 管理正在从“大锅饭”走向预算化、部门化和 ROI 约束。据经济观察报，2026 年 6 月起，腾讯多个业务员工的 Token 额度均有下降，且部门之间差异较大：AI 需求较大的混元大模型团队员工月 Token 额度约 7000 元，聚焦计算机视觉领域的优图实验室额度约 5250 元，另有腾讯娱乐外包员工月 Token 额度仅 1000 元。员工 Token 额度分配后由组内管理者再分给员工个人，额度不足时员工可向管理者“举手申请”提高额度；内部反对 Token 使用量排名，不单纯以 Token 消耗衡量员工产出。此外，阿里、字节也在不同程度上对 Token 消耗量进行管控。这说明企业侧 Token 支出已经开始从“鼓励使用 AI”的扩张阶段，进入“预算分配、额度控制、使用审计和结果评估”并重的精细化阶段。

Token 价值不能单纯看消耗数量，场景差异会带来极大的价值分化。同样数

仅供内部参考，请勿外传

量的 Token，在不同应用场景中的价值可能存在数量级差异：例如，闲聊场景中百万 Token 价值仅约 0.01 美元，而用于代码生成可达 200 美元，用于法律文档审阅可达 1000 美元，价值差异可达十万倍。背后核心驱动因素是 Token 独有的可编程属性，这是传统生产要素不具备的优势：依托同一基础大模型，借助差异化提示词、上下文设定与工具链组合，模型可充当翻译、代码工程师、法务专员、科研助理等多种角色，输出附加值截然不同的智力服务成果。Token 是可合同化的计量单位，数量可以被精确记录，但其价值高度取决于被编程执行的任务。这也催生了行业反常识特征：不足 5%的 Token 使用量，创造了超 80%可量化经济价值。用平均 Token 价格理解市场，就像用平均房价描述同时存在低端住房和核心资产的城市，数字本身成立，但难以解释真实价值分层。

图 21：不同场景的 Token 价值差异可达十万倍



数据来源：Litowitz et al. ，Bergemann et al. ，各厂商 API 定价页，腾讯研究院

企业客户购买的是业务结果，而不是 Token 数量。对企业而言，Token 支出本质上不是单纯 IT 预算，而越来越接近生产、研发和运营环节中的新型投入品。若把 Token 类比为“油耗”，只关注油耗而不关注“引擎”的经济性，用户端成本可能过高，客户最终也会放弃。企业真正关心的是 Token 是否带来业务改变，例如基金公司通过 AI 改造多模态数据治理，保险公司实现车险自动定损，重工业企业将高级技师培养周期从几年缩短至几个月。由此看，未来厂商竞争重点不应只是降低单位 Token 价格或提高 Token 消耗量，而应在更好模型引擎、更低综合成本和更高质量输出之间取得平衡，帮助客户获得可验证的生产力提升、成本下降和效率优化。

4. Token 经济打开 AI 基础设施、模型服务和应用场景的机会窗口

4.1. Token 需求放量使算力、模型和应用迎来价值重估

4.1.1. 推理需求快速增长加剧算力供给约束

算力供给趋紧已开始体现在模型服务稳定性上。随着全球 AI 应用热度持续升温，大模型调用需求快速上行，部分头部模型厂商已出现服务可用性下降、系统宕机和用户体验波动等问题。以 Anthropic 为例，截至 6 月 11 日，其 Claude API 过去 90 天正常运行时间为 99.21%，低于软件企业通常面向客户承诺的 99.99%可用性水平，并由此引发用户投诉和部分企业客户流失。国内模型厂商也面临类似压力，DeepSeek V4 系列 5 月上线后，DeepSeek 曾多次因服务中断登上热搜，宕机频次较此前明显增加。除 DeepSeek 外，千问曾在 2 月因“免费送奶茶”活动带来访问量激增，导致服务器一度拥堵；Kimi K2.6 在 4 月发布后不到 48 小时，也因大量用户涌入导致服务器承压；海外 ChatGPT 同样曾出现服务中断。上述现象表明，模型能力提升和用户

仅供内部参考，请勿外传

需求放大正在持续抬高推理侧负载，算力、带宽、调度和系统冗余能力成为大模型商业化落地的关键约束。

表3：算力供给趋紧导致部分头部模型厂商已出现服务可用性下降、系统宕机和用户体验波动等问题

时间	公司/平台	涉及模型/服务	故障表现
2026/1/23 起	智谱	GLM Coding Plan	智谱因算力资源阶段性紧张宣布 GLM Coding Plan 暂时限量发售；自 1 月 23 日上午 10:00 起，每日可销售量降至当前的 20%，后续每日 10:00 刷新购买额度，停止限售时间另行通知。
2026/2/6	阿里通义千问	千问 App / AI Agent 点奶茶活动	下单过程中多次卡顿，页面提示“活动太火爆了，请稍后再试”；活动上线不到 3 小时送出超 100 万单奶茶
2026/3/29—3/31	DeepSeek	网页对话、App、API 等	连续三天出现不同程度异常；三次故障分别持续约 1 小时 48 分、10 小时 13 分和 1 小时 3 分
2026/4/13	Anthropic	Claude.ai	Claude 出现较大范围访问故障，主要影响登录和连接；一周内第二次重大故障
2026/4/15	Anthropic	Claude.ai、Claude Code、Platform 等	Claude.ai 和 Platform 一度 down；用户出现登录失败、提示词处理失败、误报达到使用限制等问题；Downdetector 峰值超 5,100 条报告
2026/4/20	OpenAI	ChatGPT、Codex、API Platform	OpenAI 状态页显示用户无法加载 ChatGPT、Codex 和 API Platform；问题涉及登录、对话、语音、图像生成等功能，Downdetector 英国峰值超 8,700 条、美国约 1,900 条报告
2026/4/22	月之暗面 Kimi	Kimi K2.6 / Kimi 服务、Agent 额度统计	Kimi 官方称 K2.6 上线期间“访问量瞬间激增”，部分用户遇到会员排队、功能短暂不可用；后台统计 Agent 额度出现偏差，导致权益误扣
2026/4/28	Anthropic	Claude.ai、Claude API、Claude Haiku 4.5、Claude Design	短时但明显故障；Downdetector 报告约在 2:05 p.m. ET 达到约 2,700 条；影响 Claude.ai、API、Haiku 4.5 和 Design
2026/5/28	DeepSeek	DeepSeek 网页对话、API	上午 10:21 起服务中断，10:50 恢复；这是 DeepSeek 2026 年以来第 18 次部分性能异常、5 月第 5 次
2026/6/5	Anthropic	Claude 多个模型：Opus 4.6、Sonnet 4.6、Opus 4.8、Opus 4.7、Opus 4.5 等	Claude 多个模型出现 elevated errors；状态页给出各模型恢复时间，最早 15:25 UTC 恢复，最晚 Opus 4.5 于 17:29 UTC 恢复
2026/6/7—6/10	Anthropic	Claude Opus 4.7、Opus 4.8、Haiku 4.5、多个 Claude 模型等	连续多日出现 elevated errors、degraded performance 或 multiple models degraded；状态页显示 6/7 多模型性能降级，6/8 多次 Opus/Haiku 错误，6/9 Opus 4.6/4.8 错误，6/10 Haiku 4.5 错误
2026/6/11	Google	Gemini / Gemini App，多端	Google 确认 Gemini 出现大范围故障，用户在 Web、macOS、iOS、Android、Chrome 内 Gemini 等多端遇到“Something went wrong”及错误码 1099、1076

资料来源：品玩，财联社，证券时报，The Times，TechRadar，OpenAI，Claude，IT 之家，Tom’s Guide，时代在线，国泰海通证券研究

我国算力基础设施建设提速，但 Token 需求增长更快。从供给端看，我国算力中心建设已经进入快速扩张阶段。中国信通院数据显示，截至 2025 年 6 月，我国计算设备算力总规模达到 962 EFlops，全球占比约 21%，同比增长 73%；其中，计算设备智能算力规模达到 782 EFlops，同比增长 96%。拉长周期看，近 8 年我国算力规模年均增长 48%，说明国内算力基础设施正在持续夯实。不过，需求端扩张速度更为陡峭。国家数据局数据显示，截至 2026 年 3 月，我国日均 Toke 调用量突破 140 万亿，较 2025 年底的 100 万亿，增长超过 40%；相较 2024 年初约 1000 亿的日均调用水平，已实现千倍级增长。随着海量用户和企业同时启动 Agent 工作流，推理需求由持续运

仅供内部参考，请勿外传

行、多轮调用和自动化流程共同放大，需求增速明显快于供给扩张节奏。由此看，GPU 产能、电力保障、数据中心建设和算力调度能力仍需继续夯实，才能支撑 Token 经济进入规模化阶段。

4.1.2. Token 工厂标志 AI 竞争从模型能力转向生产效率

Token 工厂是以规模化生产 Token 为目标的新型 AI 基础设施。与传统数据中心或算力中心不同，Token 工厂并不只是出租机房、服务器、GPU 集群、电力和网络等硬件资源，而是以模型推理和训练任务为核心，将电力、芯片、数据等输入要素转化为可计量、可定价、可交易的标准化智能服务输出，即 Token。其本质是把“算—电—热—网”视为一个统一系统进行协同优化，通过高密度算力集群、低成本电力、散热系统、网络调度和模型服务能力，实现更高效率、更低成本、更稳定的 Token 生产。相较于泛化的“计算资源供给”，Token 工厂更强调工业化生产过程：电力、芯片和数据是原材料，Token 是产成品，客户按 Token 消耗付费，平台也可能与模型厂商按收入分成。由此，Token 工厂的价值不只来自重资产底座，而来自能否持续承接模型推理任务，并以规模经济降低单位 Token 成本。

表4：Token 工厂与传统数据中心或算力中心有较大不同

维度	传统数据中心	算力中心	Token 工厂
本质定位	资源仓库	算力枢纽	智能生产线
核心产出	计算能力、存储空间、网络带宽	算力资源（FLOPS）	Token（智能服务单位）
优化目标	高可用性、资源利用率	峰值算力、调度效率	Token 吞吐量、单位功耗产出、成本
架构逻辑	通用 IT 基础设施	高性能计算集群	算—电—热—网深度协同
能源策略	依赖电网、采购绿电	高密度供电、部分绿电	自建微电网、捆绑核电/地热、液冷标配
商业模式	出租资源（机柜、服务器）	出租算力（按卡时、按核时）	出售 Token（按百万 Token 计费）

资料来源：融中研究，国泰海通证券研究

厂商布局 Token 工厂标志着 AI 竞争从模型能力转向生产效率。随着模型能力差距逐步收敛，模型本身正从稀缺能力走向标准化供给，竞争焦点开始转向“谁能更快、更便宜、更稳定地生产 Token”。在这一过程中，AI 产业也从实验室经济走向工厂经济，价值锚点从单一模型智能水平，迁移至 Token 生产效率、单位成本和服务稳定性。算电协同因此成为关键范式：电力不再只是成本项，而是影响 Token 工厂选址、运营和利润率的战略资源；通过柔性负载调度、自建微电网乃至绑定长期低成本电力资源，算力基础设施进一步与地理位置、能源结构和电网能力深度绑定。头部厂商已经开始将这一逻辑产品化，例如百度于 5 月 13 日宣布，将百度智能云过去的“MaaS 模型服务”升级为“Token Factory”（词元工厂）。Token 工厂正从概念走向商业化落地，并成为云厂商争夺 AI 服务收入和算力运营效率的重要抓手。

仅供内部参考，请勿外传

图 22：百度千帆将 MaaS 模型服务升级为 Token Factory



数据来源：百度智能云

4.1.3. Token 出海打开数字贸易空间

Token 出海本质上是 AI 推理能力的跨境输出。“Token 出海”本质上是我国人工智能推理能力以数据要素形式的跨境输出：海外开发者或企业通过 API 调用中国模型，请求和数据经网络传输至中国数据中心，消耗中国的电力、芯片、算力基础设施和模型能力完成推理，再将结果返回海外客户。与传统商品出口不同，Token 出海没有集装箱、海关和实物交割，交付物是可计量、可定价、可持续调用的智能服务单位。Token 出海并非孤立的技术现象，而是建立在电力、AI 芯片、算力基础设施、大模型和应用生态五个层次之上的系统工程，集中体现了中国在能源供给、数据与场景、基础设施建设、工程迭代和应用落地方面的综合竞争优势。

中国具备 Token 出海的成本优势和产业基础。首先，电力优势正在转化为 Token 的价格优势。AI 推理属于高耗能环节，电力成本对单位 Token 成本影响显著。目前中国工业用电成本约 0.4—0.6 元/度，低于美国约 0.8—1.2 元/度和欧洲约 1.5—3 元/度；在贵州、云南等新能源富集区，通过电力市场化交易，风电、光伏上网电价约 0.3 元/度，部分西部绿电直供合约价格更低。以当前主流大模型高强度推理测算，生成 100 万个 Token 平均耗电约 15—20 度，对应电力成本仅为个位数人民币；而国际市场同等质量 Token 输出定价约 60—168 美元/百万 Token，即使考虑服务器折旧、网络带宽和研发成本，仍存在数量级的价值提升空间。其次，中国模型具备显著性价比优势。海外开发者在全自动 Agent 场景中可能面临一夜数千美元账单，因此更倾向于寻找低成本替代方案；粗略测算，DeepSeek R1 推理费用较 OpenAI o1 便宜 5—27 倍，Qwen 3.5-Plus 较海外可比模型便宜 23—45 倍。电力成本、模型价格和工程能力共同作用，使中国具备将低成本算力转化为全球智能服务供给的基础。

仅供内部参考，请勿外传

5. 风险提示

1）AI 应用商业化落地不及预期风险

若企业客户 AI 预算释放慢于预期，Agent、AI Coding、智能办公、行业大模型等应用场景的付费意愿和使用频次低于预期，可能导致 Token 调用量增长、模型 API 收入和相关公司商业化进展不及预期。

2）算力供给、能源资源及基础设施建设不及预期风险

Token 需求增长依赖 GPU、数据中心、电力、网络和调度系统等基础设施支撑。若高端 GPU 供给受限、国产算力生态适配进度较慢、数据中心建设周期拉长，或能耗指标、电力资源和并网条件受限，可能影响 Token 工厂、算力租赁和模型推理服务的扩张节奏。

3）模型价格竞争及技术路线变化风险

大模型 API 价格受模型能力、供需格局、开源生态、厂商补贴和商业策略共同影响。若行业价格竞争加剧，或计费模式由按 Token 计费加速转向按任务、按结果、订阅制等模式，现有 Token 价格比较框架和相关公司盈利模式可能面临调整压力。

仅供内部参考，请勿外传

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地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，特此声明。

免责声明

本报告仅供国泰海通证券股份有限公司（以下简称“本公司”）授权客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息保持在最新状态。同时，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。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也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。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

本报告版权仅为本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“国泰海通证券研究”，且不得对本报告进行任何有悖原意的引用、删节和修改。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

评级说明

投资建议的比较标准

投资评级分为股票评级和行业评级。

以报告发布后的 12 个月内的市场表现为比较标准，报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的沪深 300 指数涨跌幅为基准。

	评级	说明
股票投资评级	增持	相对沪深 300 指数涨幅 10%以上
	谨慎增持	相对沪深 300 指数涨幅介于 5%～10%之间
	中性	相对沪深 300 指数涨幅介于-5%～5%
	减持	相对沪深 300 指数下跌 5%以上
	无评级	对于个股未来市场表现与基准指数相比无明确观点
行业投资评级	增持	明显强于沪深 300 指数
	中性	基本与沪深 300 指数持平
	减持	明显弱于沪深 300 指数

国泰海通证券研究所

地址 上海市黄浦区中山南路 888 号 邮编 200011 电话 （021）38676666