AI Token 定价重估与算力硬件传导

推理需求、成本曲线与中国算力产业链投资框架

机构专题研究报告

Macro Thematic Research on AI Inference and Compute Infrastructure

报告日期	2026 年 4 月 29 日
数据时点	公开资料截至 2026 年 4 月 29 日；市场价格与产品定价以披露口径为准
研究范围	全球 AI 推理服务定价、Agent 工作负载、A 股/港股算力硬件映射
报告性质	宏观主题研究、产业链传导分析、情景测算
风险等级	中高风险；适合纳入组合研究框架，不作为单一交易依据

Contents

1 执行摘要

核心结论

1. **本轮变化不宜简单表述为“全行业 token 单价上涨”。**更准确的表述是：头部模型定价向高能力、长上下文、推理增强和企业级 SLA 分层，用户账单从“订阅感知”转向“用量感知”。OpenAI 官网披露的 GPT-5.5 标准 API 价格为输入 5.00 美元/百万 tokens、输出 30.00 美元/百万 tokens；GPT-5.4 为输入 2.50 美元/百万 tokens、输出 15.00 美元/百万 tokens。

2. **Agent 工作流是推理需求上行的核心乘数，但缓存与批处理优化构成重要的成本对冲机制。**单轮问答通常只消耗数百至数千 tokens；带工具调用、代码执行、检索与长上下文维护的 Agent 任务可达到数万至数百万 tokens，单任务账单会因工作负载结构变化而显著抬升。然而，KV cache 命中可将输入端成本降低最高 90%，Batch API 可对总用量打五折；实际企业账单增速取决于任务复杂度放大效应与效率优化降本效应之间的净差，不应简单等同于 token 消耗量增速。

3. **Scaling Law 未失效，但成本曲线需从“训练扩展”转向“推理时扩展”理解。**公开研究显示模型损失与模型规模、数据规模和训练计算量存在幂律关系；但商业化阶段的关键变量已扩展到推理 token 数、上下文长度、KV cache、工具调用、批处理效率和服务等级。

4. **硬件传导已经发生，但节奏存在显著分化。**光模块/高速互联最先反映 AI 资本开支；存储尤其 HBM 与企业级 SSD 受益于 AI 服务器与推理缓存；电力、液冷、变压器和储能属于更长周期的基础设施传导。IEA 对全球数据中心用电的基准预测为 2024 年约 415TWh、2030 年约 945TWh，应作为电力链条测算的约束锚。

5. **配置上建议采用“景气确认度与估值消化度”双轴。**光模块链条景气确定性较强，但部分标的已充分反映乐观预期；存储、液冷与电力设备处于业绩确认和订单兑现阶段，适合以基本面验证为前提分层配置。

资料来源：OpenAI API Pricing，https://openai.com/api/pricing/；IEA Energy and AI，https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

2 公开信息基础与关键假设

2.1 公开信息锚

Table 1: 核心研究事项的公开信息基础

事项	公开信息锚	分析判断	投资含义
OpenAI API 定价	GPT-5.5 标准价格为输入 5.00 美元/百万 tokens、输出 30.00 美元/百万 tokens；GPT-5.4 为输入 2.50 美元/百万 tokens、输出 15.00 美元/百万 tokens	高能力模型的输出端价格维持较高水平，结合长上下文和推理增强，企业总账单对任务复杂度高度敏感	价格研究应从单价比较转向任务级账单、模型组合和企业折扣率
GitHub Copilot 计费	GitHub Docs 披露 premium request 基础超额价为 0.04 美元/次，但须乘以模型乘数（GPT-5.5 为 7.5x，等效 0.30 美元/次；Claude Opus 4.6 fast mode 为 30x，等效 1.20 美元/次）；premium request billing 已于 2025 年 6 月 18 日开始；GitHub 已宣布自 2026 年 6 月 1 日起转向 usage-based 计费	计费模式正从” 套餐配额 + 超额按次” 向按实际用量直接计费根本性转变；模型乘数对高能力模型的实际成本放大效应在旧口径下容易被低估	2026 年 6 月计费切换后的账单结构将使企业 AI 开发工具真实用量首次系统性可见，是观察 AI 货币化程度的核心事件窗口
Anthropic Max Plan	Anthropic 官方帮助中心显示 Max 5x 为 100 美元/月，Max 20x 为 200 美元/月	高用量个人与专业用户被单独分层定价，反映头部 AI 助手的支付意愿分化	订阅升级与企业套餐共同构成 AI 服务 ARPU 提升路径
Agent 消耗量	公开产品形态显示 Agent 会持续调用工具、读取文件、执行代码并维护长上下文	单任务 token 消耗取决于上下文、工具返回、代码执行和重试次数，不能用固定倍数概括	需求测算需采用情景框架，区分轻量 Agent、复杂 Agent 与大型 agentic workflow

事项	公开信息锚	分析判断	投资含义
模型推理成本	头部闭源模型通常不披露参数量、活跃参数、集群利用率和单位推理毛利	推理成本更适合用活跃参数、上下文长度、输出 token、缓存、批处理和 SLA 进行敏感性分析	投资结论不应依赖未公开模型参数，而应跟踪 API 定价、用量、capex 和硬件交期
数据中心电力	IEA 预计全球数据中心用电从 2024 年约 415TWh 增至 2030 年约 945TWh	AI 是数据中心用电增长的重要驱动，但电力测算需与全球数据中心总用电口径一致	电力、液冷和供配电链条具备长期空间，但项目节奏和区域并网约束决定兑现速度

资料来源：GitHub Docs：https://docs.github.com/en/copilot/concepts/billing/copilot-requests；Anthropic Help Center：https://support.anthropic.com/en/articles/11049744-how-much-does-the-max-plan-cost；OpenAI Scaling Laws：https://openai.com/index/scaling-laws-for-neural-language-models/

2.2 研究边界

本报告基于公开资料、行业预测和自上而下情景测算形成判断。由于企业级 AI 合同价格、闭源模型参数、推理集群利用率和单客户订单节奏通常不完全披露，相关测算用于刻画方向和弹性，不作为精确财务预测。

- API 价格以公开网页披露为准，企业折扣、保留容量和 SLA 价格可能与公开标价不同。
- 推理收入测算采用情景假设，核心变量包括用户数、任务频率、单任务 token 消耗、模型组合和折扣率。
- 硬件传导需结合订单、交期、毛利率和客户结构验证，股价表现可能领先或滞后于产业数据。
- 报告所列公司仅用于产业链映射，不构成个股买卖建议。

3 AI Token 定价：从单价比较到任务账单

3.1 主要服务定价与计费机制

Table 2: 主要 AI 服务与计费机制梳理

提供方	产品/模型	公开定价或机制	研究含义
OpenAI	GPT-5.5	输入 5.00 美元/百万 tokens；输出 30.00 美元/百万 tokens	头部能力模型保持高输出端价格，推理增强和长任务会放大总账单
OpenAI	GPT-5.4	输入 2.50 美元/百万 tokens；输出 15.00 美元/百万 tokens	可作为本报告任务成本测算的基准模型
OpenAI	GPT-5.4 mini	输入 0.75 美元/百万 tokens；输出 4.50 美元/百万 tokens	低成本模型承接高频轻量任务，体现分层定价而非统一涨价
GitHub Copilot	Premium requests	基础超额价为 0.04 美元/次；须乘以模型乘数：GPT-5.5 为 7.5x（等效 0.30 美元/次），Claude Opus 4.6 fast mode 为 30x（等效 1.20 美元/次）；GitHub 已宣布自 2026 年 6 月 1 日起转向 usage-based 计费	计费体系正从” 套餐配额 + 按次超额” 向按实际用量计费转变；高能力模型的实际单次成本远高于基础标价；2026 年 6 月后 usage-based 账单将使企业侧用量结构更透明
Anthropic	Claude Max Plan	Max 5x 为 100 美元/月；Max 20x 为 200 美元/月	高用量个人和专业用户被单独定价，说明需求端支付意愿分层

资料来源：OpenAI API Pricing，https://openai.com/api/pricing/；GitHub Docs，https://docs.github.com/en/copilot/concepts/billing/copilot-requests；Anthropic Help Center，https://support.anthropic.com/en/articles/11014257-about-claude-s-max-plan-usage

3.2 任务账单的结构性抬升

AI 服务成本不只由单位 token 价格决定，更由任务形态决定。对 Agent 工作流而言，系统提示、历史上下文、检索结果、代码执行输出、错误重试和长推理输出共同构成 token 消耗。

$$T_{\text{task}} = T_{\text{system}} + \sum_{i=1}^{n} (T_{\text{context},i} + T_{\text{tool},i} + T_{\text{reason},i}) + T_{\text{final}} \tag{1}$$

对应的直接 API 账单可写为：

$$Cost = P_{\text{in}} \cdot T_{\text{in}} + P_{\text{out}} \cdot T_{\text{out}} + P_{\text{tool}} \cdot N_{\text{tool}} + P_{\text{sla}} \tag{2}$$

其中，P_{sla} 代表优先处理、数据驻留、企业服务等级或保留容量等附加费用。该项在企业合同中通常不会完全体现在公开 API 单价中。

Table 3: GPT-5.4 标准价格下的单任务费用测算

任务场景	输入 tokens	输出 tokens	估算 API 费用
普通问答，单轮	500	300	0.006 美元
普通文档分析，单轮	5,000	1,000	0.028 美元
轻量 Agent，约 10 步并含少量工具调用	80,000	20,000	0.50 美元
复杂 Agent，约 50 步并含代码执行或检索	500,000	100,000	2.75 美元
大型 agentic workflow，长上下文、多轮重试	2,000,000	400,000	11.00 美元

注：按 GPT-5.4 输入 2.50 美元/百万 tokens、输出 15.00 美元/百万 tokens 标准价测算；不含工具调用单独收费、缓存折扣、批处理折扣、企业 SLA 和税费。实际账单中，KV cache 命中可将输入端降至 0.25 美元/百万 tokens（降幅 90%），Batch API 可对输入和输出均打五折；在缓存充分命中的长上下文 Agent 场景下，实际账单可低于上表测算值 50%–80%。

关键判断

单位 token 价格不是唯一变量。随着应用从 Chatbot 迁移至 Agent，单任务 token 消耗、上下文长度和工具调用次数上升，会使总账单对企业预算更敏感。但方向并非单向：KV cache 命中（输入端成本最高降低 90%）、Batch API（五折）和专用

推理架构（MoE 分流、投机解码）是对冲账单上涨的重要机制。实际企业账单增速取决于” 任务复杂度上升的放大效应” 与” 效率优化的降本效应” 之间的净差；高缓存命中率的企业场景下账单增速将显著低于 token 消耗量增速。投资研究中应跟踪” 每活跃用户 token 消耗量”” 每任务平均输出 token”” 缓存命中率”” 批处理比例” 和” 企业超额付费率”，而不仅是 API 标价。

4 推理成本曲线与需求拐点

4.1 Scaling Law 与推理时扩展

Kaplan 等人在 2020 年的研究中指出，语言模型损失与模型规模、数据规模和训练计算量存在稳定幂律关系。该结论支持“大模型能力随资源投入提升”的长期方向，但并不意味着商业推理成本只能由参数规模解释。

$$L(N) \approx \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \approx 0.076 \tag{3}$$

进入 2025–2026 年后，成本曲线的核心增量来自推理时扩展：

- **长上下文**：上下文窗口扩大提升可处理任务复杂度，但 KV cache 和注意力计算带来显存与带宽压力。
- **推理增强**：高 reasoning effort 或 extended thinking 会增加输出 token 与中间计算。
- **工具调用**：检索、浏览器、代码执行、文件处理等外部工具使 token 消耗与非 token 费用同时上升。
- **服务等级**：低延迟、高可用、数据驻留、保留容量会形成公开单价之外的企业付费项。

资料来源：OpenAI Scaling Laws，https://openai.com/index/scaling-laws-for-neural-language-models/

4.2 能力拐点与支付意愿

从需求侧看，AI 应用的可付费能力取决于模型输出能否替代或显著增强专业人员工作流。对企业而言，决策变量不是“每百万 tokens 是否便宜”，而是“每项工作交付的边际成本是否低于人力或传统软件流程”。

$$ROI_{\text{agent}} = \frac{V_{\text{workflow}} - Cost_{\text{AI}}}{Cost_{\text{AI}}} \tag{4}$$

当 Agent 可稳定完成代码修复、投研初稿、数据清洗、客服质检、合规初筛等任务

时，即使单任务 API 成本达到数美元，仍可能低于人工流程的边际成本。该支付意愿是高端模型维持较高输出端价格的需求基础。

需要约束的推导边界

公开资料未披露 GPT-5.5 等模型的参数量、活跃参数、集群利用率和单位推理毛利。因此，本报告不采用“未公开参数量推导成本缺口”的方式得出涨价结论。更稳健的框架是跟踪任务 token 消耗、企业合同价格、算力利用率、硬件供给和头部厂商资本开支。

5 产业链传导：从推理账单到硬件需求

5.1 传导路径

Table 4: AI 推理需求向硬件产业链传导路径

环节	需求驱动	受益方向	主要跟踪指标
模型与应用	Agent、多模态、长上下文、企业 SLA	API 收入、订阅升级、企业合同重签	每用户 token 消耗、超额付费率、企业续约价
算力集群	推理吞吐、低延迟、并发用户增长	GPU/ASIC、服务器、交换机	CSP capex、AI 服务器出货、GPU 交付周期
高速互联	集群规模扩大与 scale-up/scale-out 并行	800G/1.6T 光模块、CPO、DSP、交换芯片	800G/1.6T 出货、客户认证、良率、ASP
存储与内存	HBM 容量、KV cache、模型权重、企业级 SSD	HBM、DRAM、eSSD、内存接口芯片	HBM 需求增速、DRAM/NAND 合约价、服务器内存容量
电力与热管理	机柜功率密度提升、数据中心并网与散热	液冷、UPS、变压器、储能、电网设备	数据中心用电、PUE、液冷渗透率、并网项目

5.2 光模块与 CPO：景气先行，估值敏感

AI 集群的高速互联需求是硬件端最先兑现的方向之一。LightCounting 在 2026 年 2 月的行业更新中指出，受 AI 基础设施投资驱动，2025 年 800G PAM4 芯片出货接近三倍增长，2026 年 800G 出货预计继续翻倍以上，1.6T 将从 2025 年的小基数增长至数千万端口量级。

Table 5: 光模块链条研究框架

维度	正向因素	约束因素
需求	北美 CSP 和模型厂商资本开支维持高位；800G 到 1.6T 升级提升单端口价值量	大客户资本开支节奏可能季度波动；技术路线从可插拔、LPO、CPO 到硅光存在切换风险
供给	中国光模块厂商在制造、封装、交付和成本上具备全球竞争力	高速 DSP、EML、硅光、先进封装等环节仍受上游能力和出口管制影响
估值	业绩高增长提供基本面支撑	领先标的股价已反映较多乐观预期，若订单、毛利率或客户认证不及预期，回撤风险较高

资料来源：LightCounting，https://www.lightcounting.com/newsletter/en/february-2026-pam4-and-coherent-dsps-381

5.3 存储半导体：从 HBM 到企业级 SSD

AI 对存储的拉动不只体现在训练侧 HBM，也体现在推理侧 KV cache、长上下文、向量检索、模型权重加载和企业级 SSD。TrendForce 预计，2026 年全球 AI 服务器出货量增长超过 20%，AI 服务器收入增长超过 30%；HBM 消耗量在 2026 年仍将保持超过 70% 的年增长。需注意量价分化风险：随着三星完成 HBM3e 资质认证，HBM 供给端将从两厂商格局转为三厂商竞争，买方议价能力上升，合约价格面临下行压力；HBM4 量产资质认证的三家同期完成时点是价格中枢的关键变量。存储链条的投资逻辑需区分” 消耗量增长” 与” 价格增长”，两者不同步时利润分布将发生分化。

Table 6: AI 存储需求的主要受益环节

环节	需求来源	A 股/港股映射
HBM	GPU/ASIC 每颗芯片搭载容量提升，HBM3e/HBM4 切换带来价值量提升	内存接口芯片、先进封装材料、测试设备、国产 DRAM 产业链
服务器 DRAM	推理并发、长上下文缓存、CPU 侧内存扩容	DDR5、RCD/DB、内存模组、服务器供应链
企业级 SSD/-NAND	模型权重、训练数据、向量数据库、检索增强生成	NAND 模组、主控芯片、企业级 SSD、存储模组品牌
封装与测试	HBM 堆叠、先进封装、良率控制与热管理	TSV、封装基板、测试设备、材料供应商

资料来源：TrendForce，https://www.trendforce.com/presscenter/news/20251030-12762.html

5.4 电力与热管理：空间大但传导周期更长

电力链条的研究需要以数据中心总用电口径为约束。IEA 预计全球数据中心用电 2024 年约 415TWh，占全球用电约 1.5%；2030 年基准情景约 945TWh，占全球用电接近 3%。在该框架下，AI 推理需求上升仍将显著拉动高功率机柜、液冷系统、供配电设备和区域电网扩容，但项目兑现节奏取决于并网条件、审批周期和数据中心资本开支安排。

Table 7: 数据中心电力链条投资判断

环节	需求逻辑	投资研究重点
液冷散热	DGX B200 等新一代系统功率密度显著提升，传统风冷方案边际受限	冷板、CDU、液冷机柜、温控集成商的订单兑现与毛利率
供配电设备	单园区功率需求提升，数据中心并网、变压器、开关设备需求增加	高压/特高压设备、变压器、配电柜、UPS 的交付周期
储能与电源	数据中心对稳定性、峰谷调节和绿电消纳需求提高	储能系统、PCS、备用电源、长期电力采购协议
电网投资	数据中心负荷具有区域集中性，对局部电网形成压力	区域电网扩容、审批节奏、电价机制与负荷管理政策

资料来源：IEA Energy and AI, https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai; NVIDIA DGX B200，https://www.nvidia.com/en-us/data-center/dgx-b200/

6 中国算力硬件板块配置框架

6.1 板块优先级

Table 8: 中国算力硬件板块配置优先级

板块	景气确认度	估值压力	配置观点	核心风险
光模块/CPO	高	高	维持核心跟踪，等待估值消化或订单超预期验证	大客户砍单、技术路线切换、出口管制、毛利率回落
存储半导体	中高	中	作为中期重点方向，关注 HBM、DDR5、企业级 SSD 与国产替代	周期价格回落、海外龙头扩产、国产 HBM 进度不及预期
液冷与热管理	中高	中	关注从主题交易向订单和收入兑现迁移的公司	标准不统一、价格竞争、项目验收周期拉长
电力设备	中	中低	作为滞后传导方向配置，重视并网与供配电项目落地	电网审批慢、项目资本开支推迟、原材料波动
国产算力芯片	中	高	事件驱动与政策驱动并存，需以生态、良率和客户放量验证	软件生态、先进制程、供应链限制、商业化节奏

6.2 代表公司跟踪口径

Table 9: 代表公司与跟踪指标

方向	代表公司	重点跟踪指标
高速光模块	中际旭创、天孚通信、新易盛、长飞光纤等	800G/1.6T 订单、北美客户集中度、毛利率、汇率与出口管制影响
存储与接口	江波龙、澜起科技、佰维存储、聚辰股份等	AI 服务器内存需求、DRAM/NAND 合约价、企业级 SSD、DDR5 接口芯片
液冷与温控	英维克、曙光数创、高澜股份等	液冷项目中标、数据中心客户、CDU 与冷板出货、单 GW 投资强度
电力设备	特变电工、保变电气、金盘科技、科华数据等	变压器订单、数据中心供配电项目、UPS 与储能配套、交付周期
国产算力链	昇腾产业链、服务器与 PCB/覆铜板相关公司	国产 AI 集群招标、适配生态、先进封装、服务器出货与 PCB 层数升级

注：公司列示仅为产业链研究样本，不构成买入、卖出或持有建议。

7 情景测算

7.1 全球 AI 推理收入情景

本报告采用自上而下框架估算 AI 推理收入，核心变量包括企业用户数、日均任务数、单任务 token 消耗、输出占比、模型组合和折扣。由于企业合同价格不公开，测算只用于判断弹性，不用于精确预测。

$$Revenue_{\text{infer}} = \sum_{i} Users_i \times Tasks_i \times (P_{\text{in},i} T_{\text{in},i} + P_{\text{out},i} T_{\text{out},i}) \tag{5}$$

Table 10: 全球 AI 推理收入情景测算

场景	2025A	2026E	2027E	核心假设
保守	200 亿美元	300 亿美元	430 亿美元	Agent 渗透慢，企业预算受限；缓存与批处理优化抵消较多需求增量；计费模式切换引发企业侧用量监控收紧
基准	200 亿美元	420 亿美元	680 亿美元	企业 Agent 进入部门级部署，模型分层与超额计费提高 ARPU；缓存与批处理优化部分对冲账单增速
乐观	200 亿美元	560 亿美元	900 亿美元	代码、客服、数据分析等场景快速放量，长上下文与多模态成为主流；该情景对应 2025A–2027E 复合增速约 112%，需持续以 API 用量、CSP capex 和企业合同数据验证

注：2025A 为研究假设锚，不代表官方统计，实际市场口径存在重大不确定性（主要头部厂商均未单独披露 API 推理收入）；收入口径覆盖 OpenAI、Anthropic、Google、Microsoft Azure AI、AWS 等主要提供商的 API 与企业推理服务，不含全部 B2C 订阅收入。乐观情景所需增速在历史上仅有极少数数字平台实现过，应将其视为上界参考而非基准预测。

7.2 硬件传导弹性

Table 11: 推理收入增长对硬件需求的弹性映射

硬件环节	弹性方向	传导机制	确认指标
GPU/ASIC	高	推理并发与低延迟要求带来新增集群与替换需求	CSP capex、服务器订单、芯片交期
光模块/互联	高	集群规模扩大带来端口数量和速率升级双重弹性	800G/1.6T 出货、交换机端口、客户认证
存储芯片	中高	模型容量、上下文缓存和数据检索需求提升存储价值量	HBM 合约、DRAM/-NAND 价格、企业级 SSD 出货
液冷/电力设备	中	单机柜功率密度和园区负荷提升，项目建设周期较长	数据中心并网、液冷项目、变压器订单

8 投资策略与组合建议

8.1 阶段性策略

2026 年二季度至三季度

建议以业绩兑现度为核心，优先跟踪存储半导体、液冷温控和供配电设备。光模块仍是景气度最高方向之一，但需将估值消化、客户结构和 1.6T 放量节奏纳入仓位管理。

2026 年四季度至 2027 年

若企业 AI 预算、API 用量和 CSP capex 同步上修，可重新提高对光模块、PCB/覆铜板、服务器电源与电力设备的弹性暴露。若 API 价格上行导致需求被抑制，则应转向效率提升和降本受益链条。

8.2 核心跟踪指标

1. **价格端**：OpenAI、Anthropic、Google、GitHub 等公开 API 价格与套餐规则，企业合同重签价格；重点关注 GitHub Copilot 于 2026 年 6 月完成从 request-based 向 usage-based 的计费切换，该事件将提供企业 AI 开发工具真实用量结构的首批系统

性数据。

2. **用量端**：Agent 产品活跃用户数、premium request 消耗、上下文窗口使用量、企业超额付费率。
3. **资本开支端**：北美 CSP capex、AI 服务器订单、GPU/ASIC 交付周期。
4. **硬件端**：800G/1.6T 光模块出货、HBM 供需、DRAM/NAND 合约价、液冷项目中标。
5. **政策端**：美国出口管制、中国算力政策、数据中心能耗审批、电价与绿电交易机制。

9 风险因素

主要风险

1. **需求不及预期**：企业 Agent 投资回报率低于预期，导致 API 用量与硬件订单低于测算。
2. **价格弹性风险**：高端模型涨价或套餐限制过严可能抑制使用量，开源模型和本地部署形成替代。
3. **技术降本风险**：MoE、量化、投机解码、KV cache 优化（当前已可将输入端成本降低最高 90%）、Batch API 折扣和专用推理芯片可能持续压缩单位推理成本；若缓存命中率大幅提升，Agent 工作流的实际账单增速将显著低于 token 消耗增速，对 API 收入弹性预期形成约束，并间接影响算力扩容节奏。
4. **供应链风险**：出口管制、先进封装产能、HBM 供给、光模块关键器件限制可能影响交付。
5. **估值风险**：部分 AI 硬件标的已反映高增长预期，业绩兑现节奏稍有偏差即可能引发估值收缩。
6. **电力与监管风险**：数据中心能耗审批、并网周期、电价政策和 AI 监管变化可能延缓建设节奏。

10 结论

AI 推理服务正在从“低摩擦订阅使用”进入“高能力模型分层、长任务用量敏感、企业服务等级定价”的阶段。公开价格本身并不能证明全行业单位 token 单价持续上行，但 Agent 工作流带来的 token 消耗倍增、工具调用和 SLA 溢价足以推动企业 AI 账单上行，并向算力硬件资本开支形成传导。

对中国算力硬件而言，光模块和高速互联已经率先兑现景气，存储半导体与液冷温

控处于从主题到业绩的验证阶段，电力设备属于更长周期但空间较大的基础设施方向。组合配置上，应以公开价格、实际用量、CSP capex、订单兑现和估值消化五类指标进行动态跟踪。

本报告数据截至 2026 年 4 月 29 日。公开资料来源包括 OpenAI、GitHub Docs、Anthropic Help Center、IEA、LightCounting、TrendForce、NVIDIA 及上市公司公告/公开财经资讯。