

Token 价格战的毛利归宿

Token 价格战的毛利归宿 —— 塌的是商品档，收的是拥有硅的人

2026-07-30 的 API 降价不是全行业单价坍塌，而是一次严格限定在商品档的选择性重定价：Luna 两条腿各降 80%，旗舰 Sol 一分未降且新增 2 倍价的 Fast 档。降价也不是成本传导——同期公司自述的服务成本改善上限约 30%，而 DRAM 合约价与 GPU 租金双双上行。缺口由资产负债表承担，而承担能力取决于一件事：是否拥有自己的硅。拥有基建的 API 服务在摩根士丹利框架下 ROIC 约 46%，转租他人算力做 API 只有约 25%，且在每百万 token 售价 1 美元时即转为亏损——而降价后的商品档混合售价是 0.30 至 0.40 美元。上游五项定价权监测指标截至今日全部未触发。

塌的是商品档，收的是拥有硅的人

涌现资本产业研究部 · 2026 年 8 月 5 日

30秒结论卡

一句话判断：这不是一次单价坍塌，是一次分层重定价——商品档被打到成本地板附近，前沿档定价权完好，而毛利被打掉的那部分，最终由不拥有硅的中间层承担。

项目	内容
机会状态	● 上游与自有硅平台 · ● 长约型算力出租方 · ● 转租他人算力的纯推理服务商
市场忽略了什么	把「Luna 降 80%」读成全行业单价崩塌。同日旗舰 Sol 价格不动，并新增标准价 2 倍的 Fast 档；能力档替代口径下商品档累计压缩 200 至 300 倍，前沿档只有 12 倍——差一个数量级
市场误解了什么	把降价读成成本下降后的让利。公司自述的端到端服务成本改善约 20%、生成效率提升 15% 以上，组合后单位任务成本改善上限约 30%，填不上 80% 的价格缺口
重点标的	NVDA ● · AVGO ● · GOOGL ● · MSFT ● · CRWV ● · NBIS ●
最大反证	若 8 月下旬 NVIDIA 季度非 GAAP 毛利率指引跌破 72%，或 CoWoS 交期自 52 至 78 周明显缩短，则「上游定价权未受传导」这一支柱失效，全篇需重写

投资要点

推理的价格曲线在 2026 年裂成了两条：商品档趋近边际成本，前沿档保持垄断定价——毛利不是沿链均匀流失，而是集中落在「不拥有硅、只转租算力」的那一层。

判断	方向	依据
降价严格限定在商品档，前沿档定价权完好	●	Sol 维持每百万 token 输入 5 美元 / 输出 30 美元，并新增 10 / 60 美元的 Fast 档
这轮降价不是成本传导，是毛利让渡	●	价格降 80% 对成本改善上限约 30%；同期服务器 DRAM 合约价二季度环比涨 55 至 60%
拥有自有基建的 API 服务仍有厚利，转租算力者已在水下	●	三种经营模式的投入资本回报率分别为约 46% / 31% / 25%；转租模式在每百万 token 售价 1 美元时转亏
单位成本地板远低于市场直觉，前提是跑最新软件栈	●	同一 GB300 硬件、同一推理模型，软件栈半年内把每 GPU 吞吐从 2,907 提到 8,064 token/秒
上游定价权尚未被传导，五项监测指标全部未触发	●	毛利率指引、先进封装交期、高带宽存储合约方向、定制芯片积压、超大厂资本开支，方向一致向上
下游不是白拿红利，智能体用量把降幅吃回去	●	智能体编码任务的 token 消耗量级为普通对话的千倍，同一任务成本波动最高 30 倍

一、钱流与价值链五层图

这条链上真正在流动的钱有两笔，量级相差一到两个数量级，而市场常把它们当成同一笔。

第一笔是 token 账单：单一模型商的 API 年收入约 65 亿美元，整个商用推理 API 支出的合理量级是数百亿美元级。**第二笔是资本开支：**四大超大厂 2026 年合计约 7,250 亿美元，较 2025 年的 4,100 亿增长 77%，且在降价发生的同一个季度被上调。

价格战打的是第一笔钱。买单上游的是第二笔钱。这就是本篇全部分层结论的起点。

层	角色	收入锚定在哪	本轮价格战的暴露度
需求源	超大厂与企业 AI 预算	自有现金流与债务融资	受益，但被智能体用量吃回去
一级供应商	GPU、定制芯片、高带宽存储、先进封装	资本开支预算与多年期订单	几乎为零
真卡口	先进封装产能、高带宽存储配额、 推理软件栈迭代速度	物理产能与工程能力	零，且软件栈成为新增卡口
可投代理	自有硅平台、长约型算力出租方	长约与积压订单	低至中
已被挤压	无自有模型、转租算力的推理服务商	现货价差	最高

二、事件本体：一次被普遍读错的降价

2026-07-30（美东时间；北京时间 07-31），GPT-5.6 家族三档中的两档下调 API 价格。这是发生在同一批 SKU 上的真实降价，而非用更便宜的新模型替换旧模型——GPT-5.6 于 2026-07-09 发布，降价距发布仅 21 天，降价前后的型号未变。

档位	输入（美元/百万 token）	输出（美元/百万 token）	变动	缓存命中输入
Sol（前沿）	5.00 → 5.00	30.00 → 30.00	持平	0.50
Sol Fast（新增）	10.00	60.00	标准价 2 倍	—
Terra（中端）	2.50 → 2.00	15.00 → 12.00	-20.0%	0.20
Luna（商品档）	1.00 → 0.20	6.00 → 1.20	-80.0%	0.02

两条腿同为精确 -80%，因此「降 80%」这个标题与输入输出配比假设无关；变化的只有混合单价的绝对水平——按 8:2 输入输出配比，Luna 混合价约 0.40 美元，按 9:1 约 0.30 美元。

三项常被忽略的结构性条款，方向都与「无限通缩」相反：

1. **长上下文加价适用全系**。单次请求超过 272,000 token 后，输入与缓存按 2 倍、输出按 1.5 倍计价，且高价率适用于整个请求。对上下文不断滚雪球的智能体负载，Luna 的实际输入价是 0.40 而非 0.20 美元。
2. **前沿档出现向上加价档**。Sol 在标题价不动的同时，新增速度至多 2.5 倍、价格 2 倍的 Fast 模式。
3. **商品档阵营已出现反向定价动作**。DeepSeek 已宣布对 V4 Flash 实施高峰时段 2 倍计价（北京时间 09:00—12:00 与 14:00—18:00，输入实付将达每百万 token 0.28 美元），生效日尚未公布。这是低端阵营首次公开的提价机制。

同期主要供给方挂牌价（每百万 token，输入 / 输出）：

供给方	前沿档	中端	商品档
OpenAI	Sol 5.00 / 30.00	Terra 2.00 / 12.00	Luna 0.20 / 1.20
Anthropic	Fable 5 10.00 / 50.00 · Opus 5 5.00 / 25.00	Sonnet 5 2.00 / 10.00（促销至 2026-08-31，之后回 3.00 / 15.00）	Haiku 4.5 1.00 / 5.00
Google	Gemini 3.1 Pro 2.00 / 12.00（超 200K 转 4.00 / 18.00）	Gemini 3.5 Flash 1.50 / 9.00	—
xAI	Grok 4.5 2.00 / 6.00（超 200K 转 4.00 / 12.00）	—	—
Moonshot	Kimi K3 3.00 / 15.00	—	—
DeepSeek	—	V4 Pro 0.435 / 0.87	V4 Flash 0.14 / 0.28

形态一致：前沿档密集分布在每百万 token 输入 5 至 10 美元，商品档已下探到 0.14 至 0.20 美元，两者相差 35 至 70 倍。价格战发生在下面这一层，不是全层。

三、分水岭：两种通缩口径必须切开

推理成本的下降速度被系统性高估，根源是把两种口径混用。

- **同一模型降价口径**：Luna -80% (2026-07-30)、Terra -20% (同日)、DeepSeek V4 Pro -75% (2026-05-31)。单次降幅是 75 至 80% 量级。
- **同能力档替代口径**：GPT-4 质量档三年累计约 200 倍 (2023-03 的 30 美元 → 2024-07 GPT-4o mini 的 0.15 美元)；市面最便宜可用模型约 300 倍。

而同一份统计给出的第三个数字才是本篇的支点：**前沿档同期只压缩了约 12 倍** (GPT-4 输入 30 美元 → GPT-5.4 的 2.50 美元，单一输入价口径)。以混合价指数口径复核，前沿篮子在 2026 年 1 月对应约 5.62 美元、7 月对应约 11.25 美元——这不是同一模型涨价，而是「前沿」这个篮子的定价重心自 2.50 / 15 档上移到 5 / 30 档。

结论：通缩是真的，但它几乎全部发生在商品档。把 300 倍套用到整条价格曲线上再推导毛利，会得出与事实相反的分层结论。

四、这次降价不是成本传导

降价方给出的归因是效率提升后的让利：在 Codex 内自主优化生产 GPU kernel、重构部分推测解码系统，使端到端服务成本下降约 20%、token 生成效率提升 15% 以上。

把这两个公开数字组合成单位任务成本的改善上限： $1 - 0.80 \times (1 \div 1.15) \approx -30\%$ 。而价格降了 80%。要在毛利率不变的前提下把每百万 token 从 1.00 美元降到 0.20 美元，需要的是 80% 的成本改善。20 至 30% 的改善填不上这个缺口。(本部测算，基于两项公开自述数字。)

三条独立证据把「成本传导」这条解释进一步排除：

1. **同日旗舰零降价**。若是全机队效率红利，Sol 没有理由一分不降。选择性重定价是竞争行为，不是成本行为。
2. **两项关键成本输入同期反向上行**。服务器 DRAM 合约价一季度环比涨 93 至 98%、二季度涨 55 至 60%，三季度指引再涨 13 至 18%；GPU 租金指数在 2026-08-03 报 111，同比涨约 6%，H100 中位在岸价 3.85 美元/GPU·小时。
3. **竞争压力有独立读数**。中国模型在第三方路由平台上已占美国企业 token 用量的 46%。

判定：这一层的毛利不是让出去的，是被打掉的。●

五、毛利到底落在哪一层：三条 ROIC 阶梯

摩根士丹利 2026-07-27 的 1GW 数据中心框架给出了本篇最硬的一把标尺。基准配置：约 41 万张 GB300、75% 利用率、年运营支出约 76 亿美元、单 GW 建设成本约 390 亿美元（含网络、存储、液冷、供配电的全包口径）。

经营模式	年收入（每 GW）	息税前利润率	投入资本回报率	本质
出租 GPU（超大厂基础设施即服务）	约 229 亿美元	约 67%	约 31%	收租金
自有基建做 API 服务	约 304 亿美元	约 75%	约 46%	收租金 + 收溢价
转租他人算力做 API 服务	约 405 亿美元	约 31%	约 25%	只收价差

收入侧算术闭合可复核： $410,000 \times 8.50 \text{ 美元} \times 8,760 \times 0.75 = 228.96 \text{ 亿美元}$ ，与公开值 229 亿一致；息税前利润率 $1 - 7.6 \div 22.9 = 66.8\%$ 。

最关键的一条是同一框架给出的敏感性：转租算力做 API 的模式，在每百万 token 售价 1.00 美元时转为亏损，在 2.50 美元时净营业利润率为 32%。

而降价后商品档的混合售价是 0.30 至 0.40 美元。

这不是说这一层今天就在亏——档位不同、负载不同、大客户折扣不同。这是说：在不拥有硅的经营模式下，商品档的公开挂牌价已经跌破了这套框架给出的盈亏线，价差空间被从上下两端同时压掉。

成本地板远低于市场直觉——前提是拥有硅并跑最新软件栈

前述框架中自有基建做 API 的单位成本地板可精确重建：推理 GPU 小时 = $410,000 \times 65\%$ 推理配比 $\times 75\%$ 利用率 $\times 8,760 = 17.51 \text{ 亿}$ ；按框架采用的每 GPU 2,750 token/秒吞吐，得年产 1.733×10^{16} token； $76 \text{ 亿美元} \div \text{该产量} = \text{每百万 token } 0.438 \text{ 美元}$ 。三重锁定：收入侧 304 亿美元（对公开值 304 亿）、息税前利润率 75.0%（对公开值约 75%）、投入资本回报率约 46%，全部闭合。

但 2,750 token/秒是六个月前的吞吐水平。同一 GB300 NVL72 硬件、同一推理模型，在 MLPerf Inference v6.0（2026-04-01 发布）中服务器场景已从 2,907 提到 8,064 token/秒/GPU（2.77 倍），离线场景从 5,842 提到 9,821——纯软件栈提升，硬件未变。以同一成本基数、当期吞吐重算： $76 \text{ 亿美元} \div 5.083 \times 10^{16} = \text{每百万 token } 0.15 \text{ 美元}$ 。

于是分层结论清晰了：

- 拥有硅、跑最新软件栈、且卖的是自家模型的一方，成本地板约 0.15 美元，商品档 0.30 至 0.40 美元的售价仍在地板之上——薄，但不是负毛利。
- 转租他人算力的一方，成本是别人的售价（框架中每 GPU·小时 7.75 美元），地板天然抬高一档，商品档定价直接击穿其盈亏线。

软件栈的迭代速度，正在成为决定谁能在商品档活下来的第一变量。这是本轮价格战里最容易被忽略的技术卡点。

六、上游：五项监测指标全部未触发

前端单价坍塌是否会向上传导，取决于五个可观测量。截至 2026-08-05：

监测指标	当前读数	是否触发
NVIDIA 非 GAAP 毛利率指引	二季度指引 75.0% ±0.5 个百分点（一季度实际 75.0%，营收 816 亿美元、数据中心 752 亿美元）	● 未触发
CoWoS 先进封装交期	52 至 78 周，2026 年产能售罄；全行业月产能约 20 万片，供需缺口自约 20% 收敛至年底约 10%	● 未触发
高带宽存储 2027 年合约方向	涨价。HBM4 单价预期自 2026 下半年约 2 美元/Gb 升至 2027 年 4 至 5 美元/Gb	● 未触发
Broadcom 定制芯片积压	730 亿美元、覆盖未来 18 个月；2027 年芯片 AI 收入指向 1,000 亿美元以上	● 未触发
四大超大厂资本开支	2026 年约 7,250 亿美元，较 2025 年 4,100 亿增长 77%，且在降价同期被上调	● 未触发

为什么传导不过去，有一个量级层面的解释：单一模型商的 API 年收入约 65 亿美元（约 250 亿美元年化收入的 26%），对四大超大厂 7,250 亿美元的资本开支，比值约 1:110。把其他模型商与云厂直销并入，整个商用推理 API 支出的合理量级是数百亿美元级——**仍比资本开支小一到两个数量级。**

上游的现金流由资产负债表买单，不由 token 账单买单。token 单价怎么打，短期内改变不了先进封装交期和高带宽存储的合约方向。●

同期的量增侧也不支持「降价换不来量」：第三方路由平台的年化处理量约 1,500 万亿 token（2026-05），同比约 15 倍；单一超大厂全表面月处理量自 480 万亿（2025-05）增至 3,200 万亿（2026-05），同比 6.67 倍。**注意口径：**全表面口径约为 API 计费口径的三倍多——该平台 API 吞吐在二季度约 220 亿 token/分钟，折合约 950 万亿/月，占全表面口径约三成。任何「token 量 × 单价 = 市场规模」的推算，在拿到 API 计费口径的绝对金额之前都不成立。

七、中间层：三种命运

（一）自有模型 + 自有基建：仍能扩利，但证据等级最低

行业估算给出的读数是推理基础设施毛利率自 38% 升至 70% 以上；同一家的投资人材料预测其二季度收入 109 亿美元、经营利润 5.59 亿美元，为其首个经营利润季，算力成本自每美元收入 0.71 降至 0.56 美元。

必须显式声明的口径差：推理基础设施毛利率与公司整体毛利率不是同一件事。同期另一家模型商的调整后毛利率自 2024 年 40% 降至 2025 年 33%，推理成本 2025 年 84 亿美元、2026 年预计 141 亿美元。

70% 与 33% 不矛盾，它们量的是不同的东西——前者剥离了消费级产品、训练与人力，后者没有。

这一支的全部量化支柱都来自未审计的第三方估算或融资材料，**只能作为假说提出。** ●

（二）无自有模型的纯推理服务商：结构最脆

两家代表性私营服务商的第三方估算毛利率分别约 45% 与 50%（GPU 成本计入营业成本），年化收入分别约 10 亿与 8 亿美元。它们同时面对上游租金上行与下游售价坍塌，且没有前沿档可以用来补贴商品档。这一层与 ROIC 阶梯中「转租算力做 API」的经营模式高度重合。 ●

（三）长约型算力出租方：是收费公路，但收的是杠杆化的租金

一季度（截至 2026-03-31）读数：收入 20.78 亿美元、营业成本 7.16 亿美元、毛利率 65.5%；积压订单 994 亿美元（同比增 284%）、剩余履约义务 988 亿美元、合同电力超 3.5GW；管理层称 A100 / H100 / H200 / L40 均价环比全部上升，近端产能「基本售罄」，新增产能加权平均合约期限约 5 年。

65.5% 这个数字必须拆开看：折旧与数据中心运营成本落在营业费用行（技术与基础设施 12.73 亿美元），不在营业成本内。把折旧计回： $20.78 - 7.16 - 12.73 = 0.89$ 亿美元，即 **4.3%**。同期净利息支出 5.36 亿美元单季吞掉全部经营利润，净亏损 7.40 亿美元；调整后经营利润率 1%。2026 年资本开支指引 310 至 350 亿美元，对应全年收入指引 120 至 130 亿美元。

判定：这不是软件毛利，是杠杆化的电力与地产毛利。它对 token 价格战有相当强的免疫力（长约、售罄、均价上行），但真实风险不在价格战，在客户信用与再融资。 ●

另一家欧洲同业一季度收入 3.99 亿美元（同比增 684%）、调整后 EBITDA 1.295 亿美元 / 32%（其 AI 云业务口径为 45%），为首个调整后 EBITDA 转正季。管理层明确表示二季度利润率会回落、三季度回到一季度水平、四季度再上台阶——**中间层的利润率是装机节奏的函数，不是稳态。**其二季度业绩定于 2026-08-12 盘前，是最近的一次验证窗口。 ●

八、下游：红利不是白拿的

降价 80% 是否变成应用层的利润，取决于用量弹性。现有证据指向：**不是。**

- **智能体负载的 token 消耗是数量级跳变。**基于 SWE-bench Verified 上八个前沿模型轨迹的研究显示，智能体编码任务的 token 消耗量为代码推理与代码对话的约 1,000 倍，成本主要来自上下文滚雪球式增长的输入 token；同一智能体跑同一任务，成本波动最高达 30 倍，且模型无法预测自身成本。
- **管理层层面已有明确归因。**某企业软件平台二季度非 GAAP 订阅毛利率自 83.0% 降至 80.5%（同比降 2.5 个百分点），归因于「客户对超大厂合作的使用增加与 AI 采纳加速」，并称该逆风将随平台规模扩大而持续；全年订阅毛利率指引 81%。同期其 AI 年度合同价值突破 10 亿美元、生产环境智能体客户九个月增长 9 倍。**用量在涨，毛利在掉，两件事同时发生。**
- **单位成本下降与总成本上升可以并存。**某语言学习平台一季度毛利率 73.0%（同比升 1.9 个百分点），归因为「单位 AI 成本持续下降」；同一场电话会里，管理层同时表示「内部 AI 成本已开始出现相当大的增长」，并指引二季度约 71%、四季度约 69%。

- **行业横截面同向。**一项覆盖 269 家公司的调研显示，AI 产品毛利率自 2024 年 41% 升至 2025 年 45%、2026 年预测 52%；但模型推理占总成本的比例随阶段单调上升，自上线前的 20% 升至规模化阶段的 23%，同期人力占比自 32% 降至 26%——**推理正在成为规模化后的主导成本项。**

存在一个显著的反例：某客户关系管理平台四季度累计处理 token 超 19 万亿（同比 5 倍），智能体产品年度经常性收入 8 亿美元（同比增 169%），经营利润率同期扩张。用量暴增与利润率扩张并存是可能的——条件是 token 成本占其收入的比例本就很低。

一个必须点破的乘法陷阱：智能体负载的 5 至 30 倍用量弹性跑在前沿推理模型上，而前沿档 Sol 此次一分未降。用商品档的价格弹性去乘前沿档的用量弹性，两者不在同一档位，乘积没有意义。

真正有效的下游成本武器不是等降价，是路由与自研。某代码编辑器公司自研推理模型于 2025-11 上线，年底已承接约一半的自动补全请求；此前在约 20 亿美元年化收入规模上，每 1 美元收入约有 0.40 至 0.70 美元用于第三方模型推理。2026 年初它进一步接入第三方长上下文低成本模型作兜底，形成「自研快速编辑 / 前沿复杂推理 / 廉价长上下文」三层路由。毛利结构自「重度用户盈亏平衡」转为「企业分部整体为正」。

九、受益标的

大公司受益链

需求源（超大厂 AI 预算与企业 token 支出）→ **直接受益：**NVDA（加速计算）、AVGO（定制芯片）、高带宽存储与先进封装供应商 → **二阶受益：**GOOGL（自有硅 + 自有模型 + 自有云，三层利润同表）、MSFT（云分部加价）→ **隐形受益：**推理软件栈与编译器能力持有方（吞吐提升 2.77 倍即成本地板下移 65%，不体现在任何单独报表科目里）→ **伪受益：**仅以「token 用量暴增」为叙事、拿不出毛利率轨迹的应用层标的 → **可投代理：**CRWV、NBIS（长约算力，收的是杠杆化租金）→ **不可投但需监控：**模型商与纯推理服务商（均为私营，无干净上市敞口）。

三张公司卡

	GOOGL 	AVGO 	CRWV 
位置	自有硅 + 自有模型 + 自有云	定制加速芯片与网络	长约型算力出租
为什么重要	唯一在三层都不向外付费的一方，成本地板天然最低	与前端单价完全脱钩，收入锁在多年期订单里	掌握合同电力超 3.5GW，机架级供给稀缺
为什么被低估	市盈率 18.9，为本表最低，市场仍按传统搜索资产定价	—	—
证据	二季度云收入 248 亿美元、经营利润 88 亿（去年同期 28 亿）、利润率 35.5%	积压 730 亿美元覆盖 18 个月，2027 年芯片 AI 收入指向 1,000 亿以上	积压 994 亿美元、剩余履约义务 988 亿、加权合约期约 5 年、近端售罄且均价环比上升
估值状态	市盈率 18.9，本表最低	市盈率 69.7，本表最贵	当期亏损，按资产与合约估值
催化	三季度云利润率是否延续扩张	2027 年指引的季度兑现节奏	客户信用事件与再融资窗口
反证	云资本开支若压制合并利润率	单一大客户订单延期	折旧计回后经营利润率仅 4.3%，单季净利息 5.36 亿美元

情景表

情景	核心假设	结果	标的动作
熊（约 20%）	前沿档跟随降价，或上游五项指标中出现两项触发	分层框架失效，价格战自下而上贯通全链	全面降低 AI 链敞口，优先减持估值最贵的一端
基准（约 60%）	商品档在成本地板附近横盘，前沿档定价权保持	毛利持续向拥有硅的一方集中，中间层分化加剧	持有上游与自有硅平台，长约出租方只做观察
牛（约 20%）	商品档提价机制扩散，同时智能体用量弹性超 5 倍	单价与用量同时向上，全链毛利修复	长约型出租方与推理服务商具备重估空间

标的	现价 (2026-08-05)	市值	市盈率	在本轮价格战中的位置	方向
NVDA	211.94 美元	51,334 亿美元	32.5	收入来自资本开支预算而非 token 账单，量级差 1 至 2 个数量级；毛利率指引 75.0% ±0.5 个百分点	●
AVGO	418.16 美元	19,894 亿美元	69.7	730 亿美元积压覆盖 18 个月，2027 年指向 1,000 亿美元以上，与前端单价完全脱钩；估值为本表最贵	●
GOOGL	377.65 美元	46,186 亿美元	18.9	唯一同时拥有自有硅、自有模型、自有云的一方，三层利润都记在自己账上；二季度云收入 248 亿美元、经营利润 88 亿美元（去年同期 28 亿）、利润率 35.5%	●
MSFT	492.81 美元	36,594 亿美元	27.5	云毛利率 65%，好于此此前约 64% 的指引；但公司毛利率回落至 67%、季度资本开支约 410 亿美元、自由现金流同比下降 23%。买方与卖方双重身份	●
CRWV	91.90 美元	501 亿美元	亏损	长约收费公路，均价环比上升、近端售罄；但折旧计回后当期经营利润率 4.3%，单季净利息 5.36 亿美元，全年资本开支指引 310 至 350 亿对全年收入 120 至 130 亿	●
NBIS	225.74 美元	573 亿美元	87.2	首个调整后 EBITDA 转正季；管理层已预告二季度利润率环比回落，利润率是装机节奏函数。8/12 为验证窗口	●

不在表内但结论明确的一层：无自有模型、转租第三方算力做推理 API 的服务商（第三方估算毛利率 45 至 50%，均为私营）。这一层与 ROIC 阶梯中回报最低、且在每百万 token 售价 1 美元即转亏的经营模式重合，商品档定价已击穿其盈亏线。●

十、风险与反证

风险在此处一次性说清，正文其余部分不重复对冲。

反证条件	触发后如何改变结论	当前状态
NVIDIA 非 GAAP 毛利率指引跌破 72%	上游定价权未受传导这一支柱失效，全篇分层结论需重写	未触发（75.0% ±0.5 个百分点，8 月下旬验证）
CoWoS 交期自 52 至 78 周明显缩短，或出现退单	物理闸门松动，价格战开始向上游传导	未触发
前沿档跟随降价（Sol 下调，或 Fable 5 / Opus 5 大幅下调）	「分层重定价」重新变成「全层坍塌」，本篇核心分水岭消失	未触发
商品档出现更多反向定价（继高峰时段 2 倍计价后再有提价）	底部已触及边际成本，价格战提前结束	部分出现（已宣布，未生效）
长约型出租方的大客户信用事件或再融资受阻	该层的风险从「价格战免疫」切换到「杠杆清算」，评级须下调	未触发
智能体用量弹性被证明低于 5 倍	下游确实拿到净红利，应用层毛利判断需上调	数据不足，事件距今 6 天

必须同时说明的三项方法学限制：

1. 本篇基于挂牌价。企业大客户折扣、承诺用量协议、云平台转售价差可能与挂牌价偏离数十个百分点，而毛利归属恰恰依赖真实成交价。
2. 本篇拒绝采信降价方给出的「效率让利」归因，理由是其自述效率数字与降幅存在 50 个百分点以上的量级缺口（见第三节测算）。这是一项判断，不是事实。
3. 模型商未上市，无审计报告。第六节（一）的全部数字来自第三方估算与融资材料，证据等级低于本篇其他章节。

十一、数据口径与冲突

冲突项	两套读数	取舍
降价日期	2026-07-30（美东） / 2026-07-31（北京时间与部分媒体跟进日）	同一事件的时区差。全文统一记 07-30 美东
单位成本地板	每百万 token 0.44 美元（按 2,750 token/秒） / 0.15 美元（按 8,064 token/秒）	两者均正确，差别在吞吐假设的时点。前者对应六个月前的软件栈
「99% 更便宜」	每任务成本口径 -99% / 纯单价口径 -98.0%（输入）与 -97.6%（输出）	每任务成本混合了单价与消耗量，不可与每百万 token 单价并列。基准已具名，token 消耗假设未公开
前沿档「涨价 100%」	混合价指数自 2026-01 的 667 降至 07 的 333	这是篮子构成上移（自 2.50 / 15 档换到 5 / 30 档），不是同一模型涨价。表述为「定价重心上移一倍」
分层毛利率对照	70%+ / 65.5% / 45-50% / 32% / 35.5% / 39% 等	六种口径（推理基础设施毛利、不含折旧的毛利、私营公司估算毛利、调整后 EBITDA、经营利润率）， 不可排成同一张表跨层比较 。本篇按口径分节处理
三季度 DRAM 涨幅	环比 13 至 18%	采用该区间；三季度未结束，属预测
长期价格下降速度	约 10 倍/年 / 约 4 倍/年	按端点值复核为 3.6 至 4.0 倍/年（-75%/年）。10 倍/年不成立

十二、行动清单

- **现在的核心持仓逻辑不变：**上游（NVDA / AVGO）与自有硅平台（GOOGL）的收入锚在资本开支预算与已签订单上，token 单价怎么打都够不到。
- **重点观察：**8/12 的欧洲算力出租方二季度业绩（验证「利润率是装机节奏函数」）；8 月下旬 NVIDIA 季度业绩与毛利率指引（本篇最大的单点反证）。
- **只监控不买：**无自有模型、转租算力的纯推理服务商——结构最脆且多为私营，无干净的上市敞口。
- **绝不追：**以「token 用量暴增」为唯一叙事、但拿不出毛利率轨迹的标的。用量与毛利在 2026 年已经解耦。
- **下次要看的五个数：**① NVIDIA 非 GAAP 毛利率指引 ② CoWoS 交期周数 ③ HBM4 2027 年合约价方向 ④ 前沿档是否跟随降价 ⑤ 企业软件公司的订阅毛利率同比变动。
- **触发重写的信号：**前沿档跟随降价，或商品档出现全面提价。任一发生，本篇的分层框架即失效。

十三、来源置信度表与口径自评

结论 / 数字	来源	日期	置信度
GPT-5.6 三档价格与降幅	OpenAI 官方公告；CNBC；VentureBeat；InfoWorld；eWeek；Simon Willison（发布日基线）	2026-07-30	A（官方页面本轮返回403，由5家以上二手源数值互证）
三条 ROIC 阶梯、1GW 框架、每百万 token 1 美元转亏	摩根士丹利报告，经 BigGo Finance 转载	2026-07-27	B
GB300 吞吐 2,907 → 8,064 token/秒/GPU	MLPerf Inference v6.0, StorageReview	2026-04-01	A
DRAM 合约价环比涨幅	TrendForce	2026-03-31 / 07-09	A（二、三季度为预测）
GPU 租金指数 111、同比 +6%	getdeploying GPU Price Index	2026-08-03	B
NVIDIA 一季度业绩与二季度指引	NVIDIA 官方新闻稿；CNBC	2026-05-20	A
CoWoS 产能、缺口与交期	TrendForce；TechTimes 转述台积电股东会	2026-06-15 / 06-05	B
HBM4 单价方向	Seoul Economic Daily；DigiTimes；TrendForce	2026-07	B
Broadcom 积压与 2027 指引	CNBC；TrendForce	2025-12-11 起	B
四大超大厂 2026 年资本开支 7,250 亿美元	Tom's Hardware；CreditSights	2026	B（单一分项存在源冲突）
算力出租方一季度财务	公司 8-K 与财报电话会	2026-05-07	A
欧洲同业一季度财务与利润率指引	公司 6-K 与电话会实录	2026-05-13	A
企业软件订阅毛利率 2.5 个百分点 及其归因	公司二季度材料与电话会	2026-07-22	A
语言学习平台毛利率与指引	公司 8-K 股东信与电话会	2026-05-04	A
智能体任务 token 消耗千倍、成本波动 30 倍	Stanford Digital Economy Lab, arXiv:2604.22750	2026-05-05	A

结论 / 数字	来源	日期	置信度
AI 产品毛利率 41 / 45 / 52% 与推理成本占比	ICONIQ Growth 半年度调研 (N=269 / N=202)	2026-01	B
模型商推理毛利率 38% → 70%+	SemiAnalysis	2026-05-01	C (未审计估算, 成本剔除项未公开)
模型商调整后毛利率 40% → 33%	Sacra; Forbes 独立复述	2026-07-28	C
私营推理服务商毛利率 45% / 50%	Sacra 估算	2026	C
第三方路由平台量增 15 倍	平台官方 Series B 公告; Menlo Ventures	2026-05-26	B
云分部二季度财务	Alphabet 8-K; Amazon 官方; Microsoft IR	2026-07-22 / 07-29 / 07-30	A
现价、市值、市盈率	实时行情接口	2026-08-05	A

数字三道关自评：①量级一致性——第四节三条 ROIC 阶梯的收入侧、息税前利润率、单位成本地板三项互为交叉校验并全部闭合；第一节四项降幅（80%/80%/20%/20%）逐项复核。②单位显式化——全文单价统一为「美元/百万 token」，token 处理量统一为「万亿」，避免与「千 token」及 quadrillion 口径混用；毛利率、经营利润率、调整后 EBITDA 利润率分节陈述，不同口径不并表。③关键数字走公司公告、财报与官方基准，第三方估算与私营公司数字在下表中单独标注为 C 级。

置信度定义：A = 公司公告 / 财报 / 监管文件 / 官方基准；B = 产业数据机构或多源交叉；C = 第三方估算或二手转述；D = 推演假设。

免责声明：本报告由涌现资本产业研究部独立编制，基于截至 2026 年 8 月 5 日的公开信息，仅供研究参考，不构成任何证券的买卖要约或投资建议。文中前瞻性判断存在不确定性，实际结果可能与预期存在重大差异。报告中涉及的标的价格与市值取自当日实时行情接口，市场波动可能导致数据变化。读者应独立判断并自行承担投资风险。

涌现资本产业研究部 · 2026 年 8 月 5 日