DeepSeek V4-Flash：Token 价格锚下移之后，AI 基建的 ROI 怎么重算

DeepSeek V4-Flash——Token 价格锚下移之后，美国 AI 基建的 ROI 证明题怎么重算

DeepSeek 于 7 月 31 日发布的 V4-Flash-0731 不是一次普通降价，而是一份产业路线声明：在 284B 总参数、13B 激活参数与基础架构完全不变的前提下，仅重做后训练就让多项 Agent 基准大幅跃升——第三方独立复测（Artificial Analysis）给出智能指数 50 分（较 4 月版 +10 分、反超自家 1.6T 的 V4-Pro-Preview 6 分、距 GPT-5.6 Luna 仅 1 分），在"智能 vs 单任务成本"的 Pareto 前沿上呈同成本垂直跃升。它真正击中的是美股 AI 叙事里最容易被线性外推的三个假设：模型越大才越强、单次 Agent 任务必然消耗越来越多的 GPU 与 HBM、闭源模型可以长期维持高 Token 单价。价格锚的现状是：OpenAI 已在发布前一日把 GPT-5.6 Luna 降价 80%（输入 $1→$0.20、输出 $6→$1.20），把与 Flash 的目录价差从 7-21 倍压缩到 1.4-4.3 倍，但按"每个成功完成任务的总成本"口径，Flash（约 $0.03/任务）对旗舰闭源模型（约 $3/任务量级）仍有约百倍级优势——竞争的计量单位已经从每百万 Token 迁移到每个任务。市场的判决同样重要：与 2025 年 1 月 R1 时刻英伟达单日 −17% 形成镜像，V4-Flash 发布后英伟达连涨五日累计 +15.4%，主流机构无一将其作为交易触发器——市场已把"效率跃迁"从黑天鹅重新定价为常态变量，这与我们在 Kimi K3 事件中的判断一致：效率改写的是 capex 的构成表，不是总量表。ROI 的正确读法是分子分母同时被改：Token 单价与闭源溢价承压（分子），单任务 FLOPs 与 KV Cache 成本同步下降（分母），裁决变量是 Agent 调用量能否超比例增长——而 OpenRouter 周榜 7.22 万亿 Token 登顶、上线数日月下载 43 万、DeepSeek 官方页新增"近期整体显著涨价"措辞与 V4-Pro 跳票共同给出一个反直觉信号：模型商自己的推理算力先吃紧了，这在短期是算力需求侧的利多而非利空。基准情景维持"单位经济通缩 + 总需求加速"：利润从模型与硬件稀缺性，向高利用率推理工厂、数据、工作流与应用结果收费迁移。

> **涌现资本产业研究部 · 2026 年 8 月 6 日** 模型与价格数据为 2026-08-06 对各厂商官方页面的直接核验；行情为交易所日线（美股至 08-05 收盘）。方向评级 🟢 受益 / 🟡 中性 / 🔴 承压。本报告为信息聚合与研究判断，非投资建议，绝不构成跟单。

30秒结论卡

- **一句话判断**：V4-Flash-0731 敲掉的是"每 Token 资源线性增长"的信仰，不是 AI 基建总量——它让 ROI 证明题更严格，但没有给出 ROI 失败的答案；市场的实际反应（英伟达发布后五连涨 +15.4%，对照 R1 时刻单日 −17%）说明"效率跃迁"已从黑天鹅被重新定价为常态变量。
- **事件本体**：架构与参数完全不变（284B 总参/13B 激活），仅重做后训练即获得大幅 Agent 增益，第三方复测智能指数 +10 分至 50、距 GPT-5.6 Luna 仅 1 分；100 万 Token 长上下文场景计算量为上代 V3.2 的 10%、KV Cache 为 7%；权重 MIT 开源、上线约一周月下载 43 万、OpenRouter 周榜以 7.22 万亿 Token 登顶。
- **价格锚现状**：OpenAI 已在发布前一日把 Luna 降价 80%，目录价差被压缩到 1.4-4.3 倍；但按"每个成功任务的总成本"口径，Flash 约 $0.03 对旗舰闭源约 $3 量级，仍差约百倍——**竞争的计量单位已从每百万 Token 迁移到每个任务**。
- **反直觉信号**：DeepSeek 官方定价页新增"近期整体显著涨价"预告、峰谷计费（峰时 2 倍）公告待生效、V4-Pro 正式版跳票——模型商自己的推理算力先吃紧了。**最便宜的模型第一时间撞上算力墙，这在短期是算力需求侧的利多。**
- **ROI 判决**：分子（Token 单价、闭源溢价）与分母（单任务算力、内存成本）同时下移，裁决变量是 Agent 调用量弹性；当前证据（云积压、AI ARR、Token 用量全部加速）支持"总量继续增长、利润迁移"而非"回报恶化"。
- **最大反证**：若未来两个季度出现云厂商 capex 下修 + GPU 租价下行 + HBM 订单转弱三者同时发生，"效率被用量回补"的判断作废，本报告切换到"效率压过需求"情景。

一、投资要点

框架：一个事件、三层传导——价格锚（模型层毛利）、ROI 结构（超大厂分子分母）、硬件斜率（英伟达与 HBM）；每层分开裁决，不许一句"利空 AI"一锅端。

#	核心判断	方向
1	**后训练成为新战场**：不改架构、不加参数，仅后训练就把 Agent 能力抬上一个台阶（第三方复测 +10 分、Pareto 前沿垂直跃升）——前沿竞争从预训练规模转向"底座 + Agent 强化学习 + 工具环境 + 推理系统"的组合优化	🟢 范式确认
2	**但"预训练与算力不重要了"不成立**：V4 系列仍以超过 32 万亿 Token 预训练；Agent 强化学习需要数十万并发沙箱（GPU+CPU/微虚拟机+存储+网络）——capex 从纯预训练 GPU 向推理工厂与后训练基础设施**迁移**，不是消失	🟡 结构迁移
3	**对闭源模型 API 明确负面**：文本、代码、子代理推理的公开价格锚被系统性压低，OpenAI 已抢先降价 80% 应对；闭源厂商必须用多模态、托管工具、企业治理与分发解释剩余溢价	🔴 模型层毛利
4	**对超大规模云厂商不是单向利空**：Token 售价降压低单位收入，但效率提升同步降低单位成本（微软披露常用模型推理吞吐 +40%、谷歌 Gemini 服务成本一年降 78%）；真正的考题是 capex 增速能否持续被收入与现金流覆盖	🟡 双向
5	**对英伟达，被侵蚀的是稀缺性租金而非总需求**：数据中心收入最新季度 752 亿美元（+92%）、下季指引约 910 亿——订单尚未体现任何坍塌；长期看每 Token 算力强度与 CUDA 稀缺溢价承压，总收入取决于 Agent 调用量与系统级价值	🟡 斜率之争
6	**对 HBM，KV Cache 利空真实但被夹在更大的算术里**：KV Cache 降至上代 7% 只压缩了内存占用五项中的一项，284B 权重驻留（4bit 理论下限约 142GB）与并发、训练状态不受影响；海力士约 10 家长约与 HBM4 量产、美光 HBM4 高量出货显示订单侧未转弱	🟡 看弹性
7	**市场行为已切换**：从"DeepSeek 时刻"（R1：英伟达 −17%、蒸发 $5,890 亿）到"效率常态"（V4–Flash：英伟达五连涨、主流卖方无一作为交易触发器、覆盖逻辑从事件驱动转业绩驱动）——这本身是本轮周期成熟度的标志	🟢 脱敏确认

二、时间线与事件本体

时点	事件
7-30	OpenAI 将 GPT-5.6 Luna 降价 80%（$1/$6 → $0.20/$1.20）、Terra 降 20%，应对成本敏感客户与中国厂商竞争
7-31	DeepSeek 发布 V4-Flash-0731：架构与参数不变、仅重做后训练；API 定价维持 $0.14/$0.28（缓存命中输入 $0.0028）；权重 MIT 开源；API 处于 public beta
8-01~03	生态快速接入：OpenCode 用量日增 30%；国家超算互联网平台 8-02 上线 0731 API；OpenRouter 周榜（7/27-8/2）V4 Flash 以 7.22 万亿 Token 登顶全球第一
8-03~05	第三方复测落地：Artificial Analysis 智能指数 50（4 月版 40），反超 V4-Pro-Preview 6 分、距 GPT-5.6 Luna（51）1 分；高盛上调中国大模型厂商 ARR 预期（2026 年 $13B、2030 年看 $125B）；大摩定性"需求结构明显向低成本模型倾斜"
8-06 现状	V4-Pro 正式版仍未发布（API 仍为 4 月 Preview 权重，官方口径 soon）；峰谷计费（北京时间峰时 2 倍）已公告未生效；官方定价页新增"近期整体显著涨价"预告

技术本体的三个关键读数：① 284B 总参数、13B 激活参数、超过 32 万亿 Token 预训练的底座不变，0731 的全部增益来自新后训练管线（工具使用/编码/推理链/自主 Agent 工作流），后训练计算量未披露；② 100 万 Token 长上下文场景下，单 Token 计算量为上代 V3.2 的 10%、KV Cache 为 7%——长上下文的边际成本曲线被实质性压平；③ 13B 激活参数压缩的是**计算量**，不是模型驻留内存——284B 权重即使全按 4bit 存储、理论下限也约 142GB，生产级部署仍需多卡高带宽互联，消费级单卡承载不成立。

基准数字的采信纪律：官方九项 Agent 基准部分跑在内部 harness、两项为内部数据集，本报告以第三方复测为准（AA 复测 Terminal-Bench 2.1 为 79%，官方自报 82.7，差异属正常 harness 口径）。

三、价格锚：从"每百万 Token"到"每个任务"

3.1 目录价现状（2026-08-06 各厂商官方页直查，美元/百万 Token）

模型	普通输入	缓存输入	输出	对 Flash 倍数（输入/输出）
DeepSeek V4-Flash-0731	0.14	0.0028	0.28	—
OpenAI GPT-5.6 Luna（7-30 降价后）	0.20	0.02	1.20	1.4x / 4.3x
OpenAI GPT-5.6 Terra	2.00	0.20	12.00	14x / 43x
Google Gemini 3.6 Flash	1.50	—	7.50	11x / 27x
Anthropic Claude Opus 4.8	5.00	0.50	25.00	36x / 89x

OpenAI 的抢先降价把入门级旗舰与 Flash 的目录价差压缩到一个量级以内——**价格战第一回合已经打完，闭源阵营选择跟牌而非死守**。但两个纵深仍在：缓存命中价 Flash（$0.0028）仍为 Luna（$0.02）的约七分之一，而 DeepSeek 的缓存折扣深度（约 98%）远超行业通行的 90%；叠加任务完成效率，第三方测算的**单任务成本约 $0.03，对旗舰闭源模型约 $3 量级仍差约百倍**。

3.2 Agent 经济学的正确算法

缓存命中率为 h 时，Flash 每百万输入 Token 混合价 = 0.14×(1−h) + 0.0028×h：命中 50% 为 $0.071、90% 为 $0.017、99% 为 $0.004。按日耗 1 亿输入 Token + 100 万输出 Token 计，模型账单在 $0.70–7.42 之间——"一整天 Agent 不到一杯奶茶钱"**可以发生但不是普遍事实**，它依赖极高重复前缀、低输出比例与低重试率，且未计入沙箱、检索、存储、工具与观测系统成本。真正的竞争指标是：

> **每个成功完成任务的总成本 = 模型 Token + 重试 + 工具调用 + 沙箱执行 + 延迟损失 + 人工接管**

闭源厂商的溢价辩护必须落在这个公式的后半段——更完整的多模态、托管工具、状态管理、企业治理与分发（DeepSeek 的兼容层对 MCP、图像、文档等仍非全面支持）。**若任务只是文本、代码与工具调用，10–30 倍的目录溢价已经无法自圆。**

四、市场的判决：从"DeepSeek 时刻"到"效率常态"

两次发布，两种世界：

	R1（2025-01）	V4-Flash-0731（2026-07-31）
英伟达	单日 −17%、蒸发 $5,890 亿	发布后五连涨、累计 +15.4% 收 $219.22
板块	全球算力链恐慌	大涨（主因是微软/亚马逊财报：MSFT 单日 +15.5%、AMZN +15.3%、AWS +37%）
机构行为	紧急下修与对冲	无一家以发布为交易触发器；覆盖从事件驱动转业绩驱动
叙事	"算力需求要崩"	"效率常态"：天风"大超预期"看多国产应用链、高盛上调中国 LLM ARR、大摩转向 ROIC 结构分析

这不是市场麻木，是市场学会了正确的算术：Kimi K3 事件（7 月中旬）已经完成了一轮教育——效率跃迁后没有任何一家机构下修数据中心用电与算力预测，BNEF 反而上修。V4-Flash 是同一逻辑的第二发：**效率改写 capex 的构成表（预训练 GPU → 推理工厂、沙箱、缓存与调度系统），不改写总量表**（详见 kimi-k3-efficiency-shock-capex-composition-2026-07-27 与 AIDC v2 的效率账）。大摩给出了结构分析的正确形状：GPU 出租、自有算力跑 API、租赁算力跑 API 三种商业模式的 ROIC 约 31%/46%/25%——**价格战淘汰的是低利用率的转售层，不是高利用率的推理工厂**。

五、ROI 重算：分子分母同时被改

> ROI ≈（模型/API 收入 + 云与软件增量 + 广告/订阅增量 + 内部降本）÷（芯片 + 数据中心 + 电力 + 网络 + 折旧 + 运维）

V4-Flash 同时压低分子（Token 单价、闭源溢价）与分母（单任务 FLOPs、KV Cache、推理成本），裁决变量是**量**——调用量弹性能否超过单位效率提升。最新一轮财报给出的是"收入证明增强、现金流压力也增强"的双升：

- **微软**：AI 业务 ARR 超 $370 亿（+123%）、Azure 供不应求、常用模型推理吞吐 +40%（效率红利首先被平台自己吸收）；2026 年 capex 约 $1,900 亿，云毛利率受 AI 投资压制。
- **Alphabet**：云积压约 $4,620 亿；Gemini 服务成本一年下降 78% 的同时 Token 处理量两年增长超 300 倍——**当前最强的 Jevons 样本**；2026 年 capex $1,800–1,900 亿仍需证明长期资本回报。
- **亚马逊**：AWS +37% 至 $422 亿/季、AI 业务年化收入超 $250 亿；但过去 12 个月自由现金流转负 $76 亿——**capex 的现金流代价已在被定价**。

结构迁移的方向同样清晰：后训练与 Agent 强化学习需要数十万并发沙箱（GPU 推理 + CPU/微虚拟机 + 分布式存储 + 调度），资本开支从纯预训练集群向这套"推理工厂"体系迁移。**ROI 证明题一季比一季严格，但云收入、AI ARR、积压与 Token 用量仍在加速——回报恶化的结论下不出来。**

六、传导映射：英伟达与 HBM

英伟达——稀缺性租金 vs 总需求。负面传导有三：长上下文单 Token 算力降至上代 10%、13B 激活让推理更易分配到国产加速器与定制 ASIC、开源权重压低云端 GPU 转售溢价。对冲同样有三：284B 权重的生产级部署仍需多卡高带宽互联；Agent 是多轮规划–调用–验证–重试的循环，任务数与步骤数的增速可能远超单步算力降幅；英伟达卖的是整套系统（网络+CPU+存储+调度），不只卖 Token 计算。订单现实：数据中心收入 $752 亿（+92%）、下季指引约 $910 亿。**长期承压的是"每任务算力线性增长"的估值假设，不是近两个季度的收入。**

HBM——KV Cache 只是五项之一。推理系统的 HBM 占用 = 权重驻留 + KV Cache×并发 + 激活值 + 并行冗余 + 训练状态；V4-Flash 把第二项压到上代的 7%，其余四项不动。产业数据也未见转弱：海力士约 10 家长约 + HBM4 量产出货、美光 HBM4 高量出货且 2027 年量产 HBM4E（格局详见 hbm-landscape-share-lta-zhbm-cxmt-2026-08-06）。对存储仓位的正确姿势不是"利空/利多"二选一，而是盯四个量：每 Agent 任务的平均上下文与步骤数、云厂商 Token 总量增速 vs 每 Token 内存降速、HBM4/4E 长约与 ASP、推理服务器内权重/KV Cache/SOCAMM/SSD 的层级迁移。

反直觉的一条：DeepSeek 官方页的涨价预告 + 峰时 2 倍计费公告 + V4-Pro 跳票，组合起来指向**模型商自身推理算力吃紧**——最便宜的模型第一时间撞上产能墙（与 K3 上线数日暂停注册如出一辙）。效率最高的玩家在超卖，这是算力与内存需求侧的短期利多信号，不是利空。

七、三种产业情景

情景	核心条件	结果	当前证据
效率压过需求	Token/Agent 量增长低于单位算力与内存降幅	闭源毛利、GPU 云租金、HBM 增速承压；应用层受益	未观测到（用量榜、积压、ARR 全部加速）
🎯 价格与成本同步下移（基准）	价格降、成本同步降、Agent 量覆盖大部分效率	基建总量继续增长但估值与利润率下修，价值向高利用率推理工厂与全栈平台集中	大摩 ROIC 结构、三家超大厂"双升"财报
Jevons 主导	Agent 成默认负载，步骤数与并发数量级增长	Token 单价暴跌但总计算/内存/电力需求继续上升	Alphabet 300 倍 Token 样本、OpenRouter 登顶、DeepSeek 自身超卖

八、受益标的落地

商业模式	价格锚下移的含义	代表方向	评级
Agent 应用层（吃 $9 劳动力预算、按结果收费）	底层智能成本下降 = 毛利扩张 + 功能覆盖扩大；价值向工作流、数据与分发迁移	治理编排与 Agent 化平台（NOW/CRM/SNOW 类，参 Agent 1:9 框架）	🟢
高利用率云平台 / 推理工厂	模型降价压推理单价，但拉高利用率并带动存储、数据库、网络收入；自有算力跑 API 的 ROIC 结构最优	超大规模云 + 全栈平台	🟢🟡
与模型效率解耦的物理层	效率事件不改电力/土地/并网的年级约束（AIDC v2 效率账结论）	电力硬卡点（燃机/变压器/开关/现场电源）	🟢
硬件斜率层（GPU/HBM）	总需求未坍塌但"每任务线性增长"估值假设承压；盯 Token 弹性四指标	NVDA/HBM 三家——订单强劲、叙事重估	🟡
纯模型 API（靠高 Token 毛利）	最大负面：价格降速快于成本降速时训练成本难摊销	未上市闭源模型商	🔴
GPU 转售 / 低利用率租赁	稀缺溢价消失后 ROIC 垫底（约 25%），价格战首先出清这一层	neocloud 中无长约低利用率者	🔴

九、行动清单与监控

- **现在可做**：维持"效率事件不减仓电力层与高利用率平台"的既定姿态；Agent 应用层按 Agent 判别式（收费模式迁移/吃 $9 不吃 $1）继续筛选加仓候选。
- **重点观察·V4-Pro 事件窗口**：正式版仍未发布（传闻窗口 8 月中旬，未获官方确认）。届时盯七件事：官方日志/模型卡/权重/API 是否同步；第三方同 harness 复测的每任务成功率；真实 Token 量与重试率；Responses API 是否扩展到 Pro；峰谷价格生效与实际利用率；三巨头是否再调价；**市场反应是否伴随云 capex 下修、GPU 租价下降、HBM 订单变化——只有价格与基本面同时出现，才构成第二波产业冲击。**
- **只监控**：DeepSeek 涨价预告的落地形态（整体涨价 + 峰时 2 倍）——它是"效率玩家超卖"信号的确认器；第三方托管商对 0731 权重的接入速度（当前仅 2 家）。
- **不做**：把"一天 Agent 一杯奶茶钱"当普遍成本事实外推；把 13B 激活当消费级部署可行性；在 V4-Pro 未发布前为"接近旗舰闭源"的传闻定价。
- **下次看这五个数**：① OpenRouter 周度 Token 榜的国产模型份额；② 三巨头缓存价与批量价的下一次变动；③ 微软/谷歌下季披露的推理吞吐与服务成本降幅；④ HBM 长约 ASP 与验收节奏；⑤ DeepSeek 峰谷计费生效后的高峰利用率信号。

十、风险与反证

#	风险	反证信号（出现即证明本报告结论错）	方向
1	效率压过需求	云厂商 capex 下修 + GPU 租价下行 + HBM 订单转弱三者同时出现	🔴 切换情景一
2	Agent 量弹性不足	OpenRouter/云厂商 Token 总量增速跌破单位成本降速	🔴
3	价格战二次降维	V4-Pro 以旗舰能力 + Flash 级定价发布，闭源阵营被迫再降 50%+	🟡 模型层加速出清
4	基准可信度	第三方复测大幅低于官方口径、内部数据集问题发酵	🟡
5	超卖信号反转	DeepSeek 涨价落地后用量与利用率双双回落——"算力吃紧"判断不再成立	🟡
6	地缘与合规	主要市场对开源中国模型的企业采用设限	🟡

十一、来源置信度表 · 数字三道关自评 · 免责

结论 / 数字	来源	日期	口径	置信度
V4–Flash–0731 本体（架构不变/价格/缓存/public beta/峰谷公告/涨价预告）	DeepSeek 官方文档与定价页	2026–08–06 直查	官方	A
各厂商目录价（Luna $0.20/$1.20 等）	OpenAI/Google/Anthropic 官方定价页	2026–08–06 直查	官方	A
OpenAI 7–30 降价 80%	CNBC + 官方页印证	2026–07–30	事件	A/B
AA 独立复测（智能指数 50、Pareto 跃升、Terminal–Bench 79%）	Artificial Analysis	2026–08	第三方复测	A
OpenRouter 周榜 7.22 万亿 Token 登顶、HF 月下载 43 万	OpenRouter/HF + 多源	2026–08	平台数据	B
长上下文 10% FLOPs / 7% KV Cache、32 万亿 Token 预训练、DSec 沙箱	技术报告及其第三方拆解	2026	技术口径	B
微软/Alphabet/亚马逊/英伟达/海力士/美光财务读数	各公司官方披露	2026–07	财报	A
单任务成本 $0.03 vs $3 量级、大摩 ROIC 三情景	Artificial Analysis / 大摩（经转述）	2026–08	测算	C
高盛中国 LLM ARR 上调、天风定性	券商报告（经转述）	2026–08	卖方	C
V4–Pro GA 窗口 8 月中旬	中文媒体汇总	2026–08	**未获官方确认**	C
行情（R1 对照、NVDA 五连涨、财报日归因）	EODHD 交易所日线 + 多源	2026–08–05	收盘	A

⚠ **数据分歧显式披露**：① 官方九项 Agent 基准与第三方复测存在 harness 口径差（82.7 vs 79%），本报告以第三方为准；② "缓存命中率 99%"为情景假设而非官方承诺（官方仅给命中单价、不承诺命中率）；③ 单任务成本百倍差为单一测算机构口径，量级可信、精确值不做载重；④ V4–Pro 相关一切能力传闻（"接近旗舰闭源"）未获官方确认，不进入结论。 **数字三道关自评**：①量级一致性——混合价三档（$0.071/$0.017/$0.004）按公式复算自洽；284B×4bit≈142GB 下限复算自洽；目录价倍数逐格按 8–06 直查价重算。②单位显式——每百万 Token 与每任务两套计量分列；FLOPs 压缩（10%）与内存压缩（7%）分列且注明对象为对 V3.2。③关键数字 ≥2 源——价格全部官方页直查；财务数取公司官方；单源测算（ROIC、单任务成本、GA 窗口）标 C 级不作载重论据。

涌现资本产业研究部 · 2026-08-06