



探索性研究论文

编辑日期：2026 年 10 月 24 日

令牌价格通缩陷阱：加速计算商业寿命经济价值的初步模型

The Token Deflation Trap: A Reproducible Synthetic Study of the Commercial Life of Accelerated AI Compute

工作状态：已在合成数据上实际运行的模型，不含任何沙特机构的原始财务或运营数据。版本：1.0 | 未经同行评议。中文稿状态：待内容所有者审阅的编辑草稿；不声称经过独立中文母语人工审核。

中文摘要

背景：人工智能芯片和推理服务（Inference Serving）持续快速演进，计算基础设施（AI Infrastructure）投资却往往按相对较长的合同与会计周期消耗。若把物理生产率提升、令牌生产成本下降、令牌售价变化以及可计费容量利用混为一谈，就会对资本可行性得出彼此冲突的结论。问题：软件效率提升在什么条件下能改善加速计算投资的经济回报？又在什么条件下，需求反弹仍不足以弥补价格下降？方法：对一组范围明确的一手披露与推理技术资料进行定向审阅，并运行一套六年期贴现现金流模型（Discounted Cash Flow; DCF）。模型组合三种假设降价率、三种需求弹性和三种容量提升速度，共形成 27 个合成实验；另执行 18 个敏感性测试，并以二分搜索计算条件性盈亏平衡弹性阈值。计算结果：27 个实验中，12 个净现值（Net Present Value; NPV）为正，15 个为负。假设基础情景的 NPV 为 -5.58 个指数单位，尽管名义回收期在第五年内到达；使该情景跨过零点所需的需求弹性约为 0.7803。若单位可变成本保持固定而不随容量改善，NPV 降至 -31.18。解释：这些数字不是分布、概率或公司收益，也不代表沙特企业的观测结果；它们只是公开假设下需求、容量、成本与折现相互作用的数学结果。方法结论：检验沙特算力投资，需要记录质量合格的可接受任务、完整交付成本、真实合同，以及按工作负载划分的商业寿命；处理器数量、利用率和实验室性能都不足以单独回答问题。

关键词：计算经济学（Computational Economics）；图形处理器（GPU）；令牌价格通缩陷阱（Token Deflation Trap）；需求价格弹性（Demand Price Elasticity）；推理服务（Inference Serving）；可接受任务成本（Cost per Accepted Task）；净现值（NPV）；沙特阿拉伯（Saudi Arabia）。

English Abstract

Background: Accelerated AI hardware has multiple competing lifetimes: physical operability, accounting depreciation, contractual commitment, software relevance, and commercial advantage. Hardware efficiency gains do not automatically translate into increased revenue or positive investment returns. **Objective:** To identify conditional mechanisms through which declining billed inference-unit prices, demand elasticity, and effective capacity growth jointly determine a stylized investment's discounted cash flows. **Methods:** We examined a bounded set of primary technical and financial records listed in the reference register and executed a fully synthetic six-year cash-flow model with 27 factorial scenarios, 18 additional sensitivities, and a binary-search break-even elasticity calculation. **Results:** Twelve deliberately constructed scenarios yielded positive net present value and fifteen negative values. An illustrative base case yielded NPV -5.58 index units despite nominal payback in year five. The conditional break-even elasticity was approximately 0.7803. Holding unit variable cost fixed in the base case lowered NPV to -31.18 . **Interpretation:** These figures are mathematical implications of chosen assumptions, not statistical estimates, observed Saudi outcomes, estimates of CoreWeave's profitability, or predictions for 2030. The central testable conjecture is that quality-adjusted, billable inference demand and realized unit margins, rather than installed accelerator count or standalone benchmark scores, more directly determine recoverable computing capital. **Next study:** A privacy-preserving, preregistered-in-advance panel of heterogeneous workloads and hardware vintages is proposed, with production billing and accepted-task outcomes.

Keywords: AI inference economics; accelerated computing; price elasticity; economic obsolescence; NPV; throughput; memory caching; Saudi Arabia.

1. 科学问题与研究边界

本文分析的是一项可供出租容量或出售推理服务的人工智能算力投资单元，不是整个沙特经济。这个单元可以代表某个资本开支（Capital Expenditure; Capex）上限内的一组硬件与推理服务，但不对应任何真实数据中心或特定处理器型号。它的方法价值在于区分两条路径：技术改进何时会提高收入利润率，何时只会增加产出，却削弱资本补偿能力。

核心问题可表述为：当每单位工作的实际成交价格持续下降，需求弹性、有效容量、可变成本和现金流时点如何共同决定一笔固定初始投资的现值？由此衍生出五个问题：

1. 令牌价格下降，是否足以让投资者推断任务交付成本也在下降？
2. 需求增长在什么节点会变成可计费单位，又会在什么节点受到容量、质量或响应时间限制？
3. 即使价格下行，软件改进能否让现有设备更加盈利？
4. 投资结果对可变成本和折现率有多敏感，而不仅是对名义容量有多敏感？
5. 哪些经过许可的沙特公开数据，能够检验这些关系而不暴露运营负载或承包商秘密？

未来应用所建议的原始分析单位是工作负载类别 × 硬件批次 (Workload Class × Hardware Vintage) 。将其汇总为总体统计时，必须披露工作负载分布，不能把它们压缩成脱离服务类型的单一平均值。

2. 五种经常被混为一谈的寿命

本文区分五种寿命：硬件仍可运行的物理寿命 (Physical Life) ；决定折旧周期的会计使用寿命 (Accounting Useful Life) ；各方承诺为容量付费的合同覆盖期 (Contractual Coverage) ；平台可在规定质量和时间内执行特定负载的工作负载调整后技术寿命 (Workload-Adjusted Technical Life) ；以及平台能够以足够利润率对抗替代方案的商业优势寿命 (Commercially Advantageous Life) 。

这套分类不是沙特机构正式采用的标准，也不是对算力市场的总体估计。它是一组分析定义，用于比较寿命各异的资产。CoreWeave 的年报说明了为什么需要拆开来看：公司自 2023 年 1 月起，把计算设备的预计使用寿命从五年调整为六年，理由包括硬件、软件与数据中心设计的进步；截至 2025 年末，已承诺合同的加权平均期限约为五年 [S01](#)。不能简单相减两个数字，再称所得差值为“一年利润”。合同不对应单一资产批次，加权期限也不是所有义务的统一到期日。

2026 年，英伟达 (NVIDIA) 发布 Rubin 平台，称在基准工作负载设计下，其推理成本相比 Blackwell 最多可降低十倍；同年 9 月，OpenAI 公布的新一代应用程序编程接口 (Application Programming Interface; API) 产品价格，也低于对照类别中的上一代产品 [S03S04](#)。前者是硬件厂商提供的推广性基准成本，后者是服务商公布的售价；还有一个未被观测的第三项，即经过合同、折扣和实际用量之后的真实成交价格。从任何一项直接跳到另一项，都是变量定义错误，不是观点分歧。

3. 技术文献：价值为何产生在内存与队列之间？

实际生产率不仅由峰值算力决定。即使不更换所有硬件，以下技术也可能提高有效产出：

内存管理：Kwon 等人 2023 年发表的 vLLM 系统与分页注意力算法 (PagedAttention) 通过切分、复用键值缓存 (Key-Value Cache; KV Cache)，减少内部碎片和闲置预留。论文实验报告称，在近似延迟下，与特定基线相比，服务吞吐量 (Throughput) 提高到约两至四倍 [S05](#)。要在沙特复现这一比例，需要明确平台配置、工作负载和软件版本；论文不能直接支持商业层面的普遍推论。

调度：Orca (2022) 提出按生成迭代进行调度 (Iteration-Level Scheduling)，减少短请求在长请求之后的阻塞，并采用选择性批处理 (Selective Batching) [S06](#)。这提醒我们，等待时间既可能来自处理器容量，也可能来自软件决策。

推理阶段拆分：TensorRT-LLM 技术文档区分上下文预填充 (Prefill) 与生成解码 (Decode)，并解释为何两阶段可能需要不同的资源和并行配置。文档也说明了阶段分离时搬运推理缓存的成本；拆分并非免费，也不会对每一个请求都更好 [S07](#)。

基准参照点：2026 年 9 月，MLCommons 发布 MLPerf Inference v6.1，其中包括多步骤任务测试 S11。这些结果可以帮助建立统一的性能测量基线，但它们不是净现值，也不是计入电力、融资和销售条款后的商业比较。

本文由此提出一种可能机制，而非因果证明：调度、内存管理和计算阶段得到优化后，现有资产可以交付更多服务。新增服务中有多少能被计费，则取决于需求、客户行为、结果质量和竞争格局。收益可能归于算力运营商、应用开发者、最终用户，也可能在三者之间分配。

4. 相互竞争的假设

- H1，容量效应：当容量构成约束，且付费需求仍能吸收新增产出时，可交付能力提升会增加净价值，前提是扩张与运营成本没有压倒收益。
- H2，价格通缩效应：需求缺乏弹性时，即使单位数量增加，收入也可能快速下降，使资本在既定融资窗口内更难回收。
- H3，决定性反弹：当需求弹性很高、又有容量承接需求时，降价后总收入可能增加。这不等于企业利润必然增长。
- H4，合同与现货市场：存在多年容量购买义务时，未来收入受即时服务价格影响的程度会不同；但交付权利、违约、折扣和续约仍使保护有条件。
- 竞争假设 R1：算力供给增加可能同时带来单位价格下降和利润强劲增长，因为新的高价值工作会吸收全部容量；此时，已实现利润数据中不会出现陷阱。
- 零假设 H0：控制工作负载组合、电力、融资、合同与质量以后，单位价格下降速度对投资回报不再具有额外解释力。在获得实际运营数据并建立恰当因果推断设计以前，不应排除这一假设。

这些都是假设与解释命题。计算测试只检验它们在一组明确假设构成的模型内是否自洽，不涉及受试人群或真实设施。

5. 已执行模型的设计与单位

输入与符号：对每一年 $t = 1, \dots, 6$ ， p_0 表示初始价格指数 1， q_0 表示年需求 50 个单位， k_0 表示年可交付容量 50， c_0 表示初始单位可变成本 0.20， F 表示年度固定运营成本 7， I_0 表示初始投资 120， s 表示第六年末假设剩余价值 8， r 表示折现率 12%。它们全部是假设指数单位，不代表真实的美元、里亚尔、兆瓦、设备或令牌。

测试变量为： δ ，单位价格年度下降率； ε ，需求价格弹性的绝对值； g ，假设可持续的运营改进所带来的可售容量增长率。程序执行以下公式：

$$p_t = p_0(1 - \delta)^{t-1}$$
$$D_t = q_0 \left(\frac{p_t}{p_0} \right)^{-\varepsilon}, \quad K_t = k_0(1 + g)^{t-1}$$

$$X_t = \min(D_t, K_t), \quad c_t = c_0(1 + g)^{-(t-1)}$$

$$CF_t = p_t X_t - c_t X_t - F + S \mathbf{1}_{t=6}$$

$$NPV = -I_0 + \sum_{t=1}^6 \frac{CF_t}{(1 + r)^t}$$

其中， D_t 为潜在需求， K_t 为合格服务上限， x_t 为可计费单位， CF_t 为运营现金流，第六年另计剩余价值。模型假设一次交易或请求等于一个单位，且每个交付单位都满足质量和服务要求；这是一项方便但未必现实的假设，未来实证研究必须将其拆开。模型用最小值函数避免为没有容量承接的需求计入收入。模型不含债务、利息、税收、工资变化、会计折旧、交易方破产或硬件停机；资本成本由假设折现率表示，并不是从美国公司的披露中推导出的贷款价格。

对 $c_t = c_0/(1+g)^{(t-1)}$ 必须格外谨慎。我们假设软件生产率提高会按相同比率降低单位可变成本；如果大部分改进只是让设备运行更多负载，却没有真正减少电力或网络开支，这一假设就会失败。因此，另行执行的敏感性路径让 c_t 始终保持为 c_0 ，无论容量如何变化。较保守的结果见第八节。

6. 数据、执行与可复现性

本研究不包含银行交易记录，也不包含沙特运营商的利用率数字。所有实验值均来自可读文件 `data/model_assumptions.json`；`analysis/reproduce.py` 仅使用 Python 标准库执行计算。程序实际生成五类记录：

1. `base_annual_cashflows.csv`：六行年度现金流、规模、容量、收入与折现数据。
2. `synthetic_scenario_results.csv`：按 $3 \times 3 \times 3$ 组合的 27 个独立情景网格。
3. `sensitivity_checks.csv`：18 个补充情景，交叉组合三种折现率（8%、12%、16%）、三种容量改善率（0%、25%、50%），以及单位可变成本改善或保持固定两种设定。
4. `derived_checks.json`：从输出中提取的阈值、基础结果与情景计数摘要。
5. 依据结果表生成的 SVG 与 PNG 图形文件，并非手绘估计图。

另有可供检查的表格文件 `evidence_workbook.xlsx`，把可编辑输入、计算和结果图分开呈现。它是一项核查工具，不能替代脚本作为完整实验的计算记录。模型假设年末支付、零时点投资，并只在第六年计入剩余价值。合成数值不接入任何外部财务来源；外部资料只用于解释为何要检验这些机制，以及从模型走向现实的推断限制。

7. 实际实验结果：为什么需求并不总能获胜？

模型运行了三种年度降价率——10%、22% 和 35%；三种需求弹性 ε ——0.5、1.0 和 1.8；以及三种容量增长率——0%、25% 和 50%。每一个组合都运行六年，再计算现值结果。模型没有随机抽样，也没有估计概率分布，因此网格中的盈利情景比例不是现实世界的预期成功率。它只是对研究者主动选择的参数空间片段进行计数。

基础情景：价格每年下降 22%，弹性为 0.7，容量与单位成本每年改善 25%，折现率为 12%。以下各年结果来自实际计算，不是编辑者的预测：

年份	价格指数	潜在需求	可用容量	售出单位	假设收入	年度现金流（含适用时的剩余价值）
1	1.000	50.00	50.00	50.00	50.00	33.00
2	0.780	59.50	62.50	59.50	46.41	29.89
3	0.608	70.80	78.13	70.80	43.08	27.01
4	0.475	84.25	97.66	84.25	39.98	24.35
5	0.370	100.25	122.07	100.25	37.11	21.90
6	0.289	119.30	152.59	119.30	34.44	27.63*

第六年包含 8 个单位的假设剩余价值。所有栏目都是合成指数值，四舍五入至两位或三位小数；精确计算见 CSV。

需求增加到两倍以上，年度收入却从 50 降至约 34.44。六年收入合计约 251.02 个单位，但收入并不等于利润：可变成本、固定成本和投资时点共同留下约 -5.5835 的净现值。扣除初始成本后的累计现金流，在不折现时于第五年转为非负；但由于较晚回流的资金权重更低，净现值仍为负。不能把 *Payback Period* 与 *NPV* 混为一谈。

下表固定年降价率为 22%、容量改善率为 25%，只改变弹性：

假设需求弹性	六年 NPV（假设单位）	第六年收入	分析结果
0.5	-17.98	26.87	数量增长不足以抵消价格通缩。
1.0	+8.91	44.05	更高需求吸收了可售容量。
1.8	+8.91	44.05	结果相同，因为交付受容量上限约束。

1.0 与 1.8 两行相等，并非随意的近似。在这组假设中，第一年之后的每一年，潜在需求都超过可计费容量，于是 $x_t = K_t$ 。购买意愿继续上升，也不会改变实际产出。因此，必须区分潜在市场增长、可交付规模增长和实际付费规模增长。即使市场存在更多同价需求，只要不能扩容或在客户接受范围内提价，未获得服务的需求就无法补偿收入。

合成模型中的 NPV 临界面

净现值随价格下降和需求弹性假设而改变正负

		10%	22%	35%
需求价格弹性绝对值 (假设)	0.5	+9.8 盈余	-18.0 缺口	-48.0 缺口
	1.0	+28.4 盈余	+8.9 盈余	-46.9 缺口
	1.8	+66.4 盈余	+8.9 盈余	-46.9 缺口

计费单位价格年降幅 (假设)

所有单元格固定：可交付产能年增 25%；贴现率 12%；前期投资为 120 指数单位。

六年期合成指数单位；并非 SAR、USD 或观测到的 GPU 回报。

假设情景中的净现值网络

图 1：在年度容量改善率固定为 25% 时，不同弹性与降价组合的实际实验结果矩阵。数字不是概率分布，也不是对沙特机构的估计。

完整实验的 27 个情景中，12 个净现值为正，15 个为负，三个轴上的符号混合。当容量在六年内始终固定为 50 个单位 ($g=0$) 时，所选设定中的结果全部为负，因为需求虽然增长，实际交付最多仍是 50，而价格继续下降。另一端有一个高弹性、高容量情景 ($\delta=.22$ 、 $\epsilon=1.8$ 、 $g=.50$)，净现值达到约 +113.75。但它依赖软件生产率连续多年高速增长的强假设，不能据此按相同数字作出真实投资论证。

8. 敏感性测试结果，以及结论何时改变

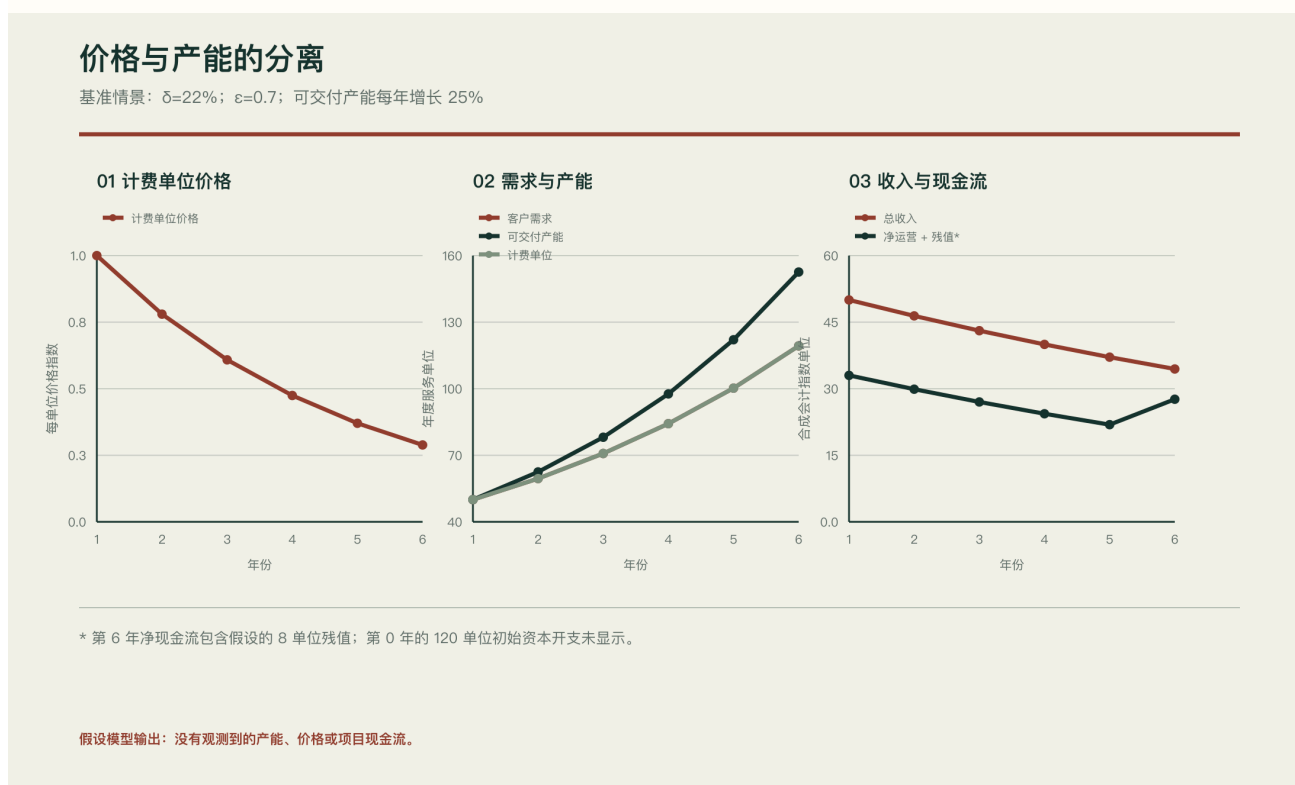
实际测试把折现率设为 8%、12% 和 16%，容量改善率设为 0%、25% 和 50%，同时分别让单位可变成本按假设比率下降或保持固定。在价格下降 22%、弹性 0.7、容量改善 25% 的情景中，结果如下：

折现率	假设可变成本随容量改善	假设可变成本固定
8%	+7.84	-22.46
12%	-5.58	-31.18
16%	-16.82	-38.62

理解结果比背下数字更重要。去掉单位可变成成本下降的假设后，仅有容量提升不足以产生原先预期的回报，因为收入由需求实际购买的数量决定，不由软件理论上能运行的数量决定。在部分组合中，当供给已达到需求上限后，把 g 从 25% 提高到 50% 根本不会影响收入，只要售价和需求规模没有变化。这不是否定软件，而是在问：我们真正移除了哪一种约束？是计算约束、成本约束，还是客户缺席的约束？

我们在 0 至 2.5 区间内对 ε 进行迭代二分搜索（Binary Search），固定基础情景的其他数值，以求出使 $NPV \geq 0$ 的最小弹性，结果约为 0.7803。这不是对沙特或全球市场需求弹性的估计，而是特定模型内部的阈值。改变初始投资、价格、交付质量、需求特征或折现期限，阈值也会改变。它可以通过检查 `data/derived_checks.json` 并运行脚本重新计算，不能仅凭某个知名经济概念的名称得出。

测试限制： 本研究没有模拟电价随机上涨、网络拥塞、合同流失或数据中心地理差异。没有统计样本，因此不存在 p 值、置信区间或统计显著性主张。18 个敏感性情景也不代表完整风险空间；测试只覆盖三组明确、已披露参数的交叉组合。



假设实验中价格、需求、容量与收入的路径

图 2：基础情景结果。容量改善、价格老化与现金流时点是不同现象；财务纵轴仅使用合成指数单位。

9. 检验与沙特的关联，但不把不同公司的数据强行拼接

有三类一手事实与实证问题有关，但它们并不构成同一条测量序列：

A. 美国商业披露，不是沙特测量：CoreWeave 的 2025 年报告显示，设备会计寿命与合同期限并不相同。其 2026 年财务报表披露，截至 2026 年 6 月 30 日，剩余履约义务（RPO）达到 1037 亿美元；公司预计从该日期起的前 24 个月确认 41%，第 25 至 48 个月确认 39%，余下部分最晚至第 78 个月确认

S01S02。同一披露还列出一张综合到期表，其中未来债务本金支付合计 355.51 亿美元 S02。不能把两个数字相加或相除，再据此构造担保比率、危机结论或过剩判断。前者是按收入确认规则尚未履行的合同服务，后者是跨多个时点到期的融资义务；二者不是相互对冲的净现金流。长期合同可能保护运营商免受即时价格波动，但本研究没有分析单份合同条款或履约记录。

B. 商业相关方发布的沙特运行公告：2026 年 8 月 31 日，AMD、思科（Cisco）与 HUMAIN 宣布，采用 AMD Instinct MI355X 处理器和 Cisco Silicon One 网络的系统已经在沙特投入生产，为 HUMAIN 在境内的客户提供服务 S09。同一公告还称，计划从 2027 年开始部署最高 250 兆瓦容量，并以 2030 年达到一吉瓦为目标。这两项是规划容量，不是 2026 年的已发布运行结果。公告来自具有直接利益的参与方，因此本文明确归因于发布者，不把它当成对需求、成本或利润的独立审计。

C. 面向用户的沙特服务：沙特通讯社（Saudi Press Agency; SPA）于 2025 年 8 月 25 日宣布推出基于 ALLaM 34B 的 HUMAIN Chat，提供阿拉伯语界面与语音工具 S10。这能确认产品按发布机构说法已上线，也提出了一个问题：算力投资如何转化为有用需求？但它不能支持关于利润、收入、令牌数量或用量统计的推断。它可以作为社会使用的情境案例，不能作为未公开的数值观测写入模型。

因此，沙特面对的问题不是照抄某个国家已经执行的故事，而是计算设施能否在本地与跨境市场中，按照服务等级要求和动态成本出售应用结果。基础模型训练（Foundation Model Training）与推理服务（Inference Serving）的经济结构可能不同，因此不能只用令牌价格评估主要用于大规模训练的资产。

10. 区分“价格”“成本”与“价值”：建议的测量登记表

要把本文推进为实证研究，首先需要厘清公共讨论中经常混在一起的三种计量单位：

1. **服务商成本（Provider Cost）：**单位工作的真实运营成本，包括能源、网络、软件组件、支持、闲置容量，以及在恰当分摊层级上的折旧与融资影响。它不等同于边际运营成本。
2. **实际成交价格（Realized Selling Price）：**服务商在合同、折扣、缓存与不同计费方式之后实际取得的价格。它不一定等于网页上的公开标价（List Price）。
3. **客户价值（Value to Customer）：**已完成任务为客户带来的效用，可以通过节省时间、生产率、质量或相关收入估计，但必须独立测量，不能自动等同于价格。

请求并不均质，因此本文建议在每百万令牌成本之外，同时报告“每项可接受任务成本”（Cost per Accepted Task）。可接受任务必须事先按准确性、安全性和时限标准定义。测量还应特别记录任务重试比例、上下文长度差异、输入缓存命中率（Cache Hit Rate）、输出令牌数、提前停止、资源不可用，以及按有据可查的方法分摊到负载层级的用电量。

对工作负载类型 w 和时期 t ，可作如下描述性表示：

$$CAC_{w,t} = \frac{E_{w,t} + N_{w,t} + O_{w,t} + A_{w,t}}{J_{w,t}^{accepted}}$$

其中， E 是估算的能源与冷却成本， N 是网络与数据成本， O 是分摊的运营与支持成本， A 是资本使用份额， J 是该时期通过验收标准的任务数量。这个公式只是建议，尚未成为标准，实施时需要成本分摊政策，并证明没有重复计算开支。若分母中没有可接受任务，就不能展示虚假的“零成本”，也不能除以零。要完成整个项目分析，还须加入未分摊开支与递延收入，不能把这项局部指标当成净利润。

还可增加两项指标：约定时限内可接受任务的边际利润（SLA-Compliant Contribution Margin），以及由已承诺合同产生的服务占比与跟随时市场价格的收入占比（Spot-Market Revenue）。每项指标都应说明统计期、服务定义、实际价格、币种与费用分摊方法。

11. 一项能够证伪假设的后续沙特研究设计

建议设计，尚未执行：如果数据可得，可在自愿且有记录的合作基础上，选择代表多种使用模式的运营商与开发者，对汇总后的企业工作类别和连续硬件批次开展 24 至 36 个月的月度观察。第一阶段的预设分析应为描述性研究。只有在出现能够明确隔离的变化时，例如只为一部分工作负载升级推理软件、同时保留未同步变化的对照组，才进入因果设计。除非真正实施随机分配，否则不能把升级随意称为“随机实验”。

未来数据表的最低要求：匿名硬件批次标识；模型类别与版本；输入和输出长度区间及缓存率；可接受任务比例；首令牌时间与服务等级内的请求完成时间；可计费工作单位；实际成交均价及分布；按已披露方法分摊的电力与冷却成本；合同价格占比；折旧表和选定经济折现率；工作负载起止日期。数据表不应包含关键设施地址、机房位置、用户日志、个人合同或访问密钥。任何可能暴露敏感运营用量规律的记录，都应延迟发布或汇总处理。

比较步骤：

1. 在查看结果前定义三至五种负载类别：短交互推理、长上下文、非即时批处理、补充训练，以及在获准采集时纳入多阶段任务。这些只是可以按数据条件调整的示例。
2. 为每一类负载建立专用标准质量测试，使吞吐量提升若以结果正确性为代价，能够在分母中显现。
3. 按时间切分样本；用较早时期的数据校准模型，用之后的样本外时期（Out-of-Sample）评估，并阻止未来价格或预测当时尚不存在的硬件世代泄漏进模型。
4. 同时估计实际成交价格曲线与付费规模，并加入质量变化、容量可用性与合同变化的解释变量，不要只从单一公开标价估计弹性。
5. 在条件允许时，在同一硬件批次与固定负载内部比较软件升级效果，同时跟踪客户组合、时间和能源支出的变化。若没有合理对照组，只能把结果表述为相关关系。
6. 把资本到期义务与合同条件差异纳入假设有据可查的财务模型，并将结果与真实支付规模比较，避免从单一技术指标推断利润。

数据允许时建议的因果估计量：比较同一时期内，接受软件改进的工作负载与未接受改进的相似负载，其可接受任务成本变化之差；该估计依赖干预前平行趋势（Parallel Trends）和不存在其他差异性变化的假设。然而，算力中心共享资源，升级影响可能通过队列或共享内存扩散到其他负载，因此“单元之间不

存在干扰”（No Interference）的假设可能不成立。研究可以改在整个资源池层级进行比较并明确影响扩散，也可以只描述机制。

重要反事实检查：推理软件升级可能与新客户进入、更小模型发布或低价新合同同时发生。若不固定这些变化，就可能把组合变化或议价能力造成的效果误归于软件。价格弹性研究还要处理未满足需求：只看完成的预订，可能误以为需求不强，却漏掉容量满载时被拒绝的请求。

12. “令牌价格通缩陷阱”可能在哪些地方失效？

一个醒目的名称不能自动证明自身正确。至少有六项边界会削弱论点，甚至反转结果：

第一，处理器不同于最终产品。服务价格的一部分来自软件、安全、支持、访问权和用户界面；这些要素的价值可能在处理器成本下降时上升。公司内部令牌成本下降，不必然造成总收入下降。

第二，工作负载可以迁移。不再适合训练前沿模型（Frontier Model）的处理器，可能转去运行较小模型或更便宜的推理任务。设备离开性能前沿后，价值不一定归零。当前模型有意采用限制性假设，没有允许负载以不同定价转移给新用户。

第三，任务类型会变化。新一代应用可能需要多工具推理或超长上下文，因此即使单位令牌成本下降，整项任务成本也会上升。得到一个结果所需的令牌数量发生变化后，每百万令牌价格就不再代表服务的经济质量。

第四，合同。较高的合同收入占比，可能在还款期内把现金流与即时价格隔离；价格影响则推迟到客户续约时出现。这种时差需要进入模型。为清楚展示机制，基础情景假设所有单位价格都随时间变化，没有加入这层缓冲。

第五，非计算资源。电力许可、变电站延期、冷却用水或网络设备都可能限制算力部署，使软件产生的理论容量无法变成服务。相反，内存和并行算法也可能减少新增设备采购。本研究没有实际估算任何一项约束。

第六，需求充足与替代约束。如果低价创造了新的有用工作，市场又具有弹性，收入和利润都可能增长。但短期容量仍受限时，价格和合同可能先变化，而不是数量变化。模型只隔离一项约束，没有估计多家竞争企业之间的一般均衡。

要分开这些因素，分析者不能只回答一次“价格每年下降多少”。还要持续观察版本、任务、合同期限、受抑制需求与可变成本，并测试其他模型，包括不采用恒定弹性的需求模型，以及部分工作负载价格并不持续下降的模型。

13. 新投资之前建议进行的区分性测试

本文建议开展一项非侵入式投资决策试验，无需设施位置或敏感流量细节。获准运营的机构选择两类工作负载：一类以响应时间为主要约束，另一类以批处理规模为主要约束。随后在获准资源上做三组比较：同一硬件上运行旧版与新版软件；在条件允许时，同一软件运行旧设备与新设备；在许可边界内，将部分负载在两者之间迁移。模型、质量区间、响应时限与实际成交价格数据均需固定或控制。测量期要足够长，能够观察需求变化，而不只是短时峰值。

如果软件升级降低了每项可接受任务成本，并在没有降低售价时提高运营商总利润，就支持软件挽救路径。如果售价下降快过总成本，虽然请求数量增加，资本回收却变弱，就支持陷阱路径。如果付费任务强劲增长、利润也上升，就能在该负载类别上反驳“价格通缩必然带来亏损”的说法。如果控制变量后利润没有变化，则应缩小主张范围，而不是把结果当成研究者的个人失败。

可以把以下决策规则作为待检验框架，而不是把它宣称为现成标准：在为新一批算力融资以前，比较三种方案的条件净现值：（A）购买新设备；（B）改进现有服务软件；（C）临时购买外部容量，用来测量真实需求。每个方案都应计入融资成本、风险、能源消耗、开发周期与合格服务收入。任何方案都可能因负载和合同条件而获胜。这只是 M12 文章范围内的实用比较框架，不是普遍“治理”方案，也不是具体投资建议。

14. 可以向决策层说什么，不可以说什么

可以说：“推理技术成本下降与软件改进正在改变长寿命资产的经济结构，值得检验付费需求弹性与资本回收时间。”也可以说：“2026 年沙特已有多种硬件投入运行的公开声明，并有后续扩张目标，但我们审阅的材料没有披露详细利润率。”

不能说：“每一颗处理器都会在一年内失去价值”；“合同使亏损不可能发生”；“27 个实验中 12 个盈利，意味着盈利概率是 44%”；或者“合成 NPV 衡量了沙特项目”。同样不能把美国公司已确认价值的未来合同收入等同于当前现金，也不能把 2030 年规划兆瓦数等同于 2026 年已使用容量或已赚取利润。

当前条件性判断：价格下降不是技术失败，对用户而言还可能带来巨大的社会与经济收益。只有在付费需求、真实成本和时间表无法跟上价格变化时，它才构成资本风险。另一方面，优秀软件、周密合同与负载重新分配，可以提高现有资产的可行性。现有公开数据不足以在具有代表性的沙特市场中区分这些情形，因此研究建议仍是围绕具体决策开展定向测量，而不是预判胜利或亏损。

15. 可复现性声明与数据伦理

在文章目录运行一条命令即可本地复现计算实验：

```
python3 analysis/reproduce.py
```

过程只需标准 Python; analysis/requirements.txt 对此作了记录。data/model_assumptions.json 包含全部选定数值, 程序生成 data/base_annual_cashflows.csv、data/synthetic_scenario_results.csv、data/sensitivity_checks.csv 和 data/derived_checks.json。任何假设变化都会生成不同结果, 发布者必须说明新模型版本及参数。生成图形需要额外绘图库, 但验证数字方程不需要。若表格运行环境支持公式重算, evidence_workbook.xlsx 可以独立核查基础情景。

本研究没有收集个人数据、敏感运营信息或私有依赖图。没有声称获得机构批准, 也没有声称经过同行评议、完成沙特实证验证或接触未公开合同。所用真实资料全部来自公开披露、论文或公告, 并明确指出商业公告表达的是发布方立场, 不是第三方独立审计。合成数据专为说明方程内部可能存在的因果关系而生成。

16. 一手与参考来源

S01. CoreWeave, Inc. *2025 Annual Report on Form 10-K*. U.S. Securities and Exchange Commission.

Accounting Policies: Property and Equipment; Committed Contracts; Take-or-Pay 等部分。

<https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/crwv-20251231.htm>。财务数据截至 2025 年 12 月 31 日; 报告于 2026 年公开。

S02. CoreWeave, Inc. *Quarterly Report on Form 10-Q, period ended June 30, 2026*. SEC. Note 2, Remaining Performance Obligations; Debt maturity table.

<https://www.sec.gov/Archives/edgar/data/1769628/000176962826000366/crwv-20260630.htm>。观测日期为 2026 年 6 月 30 日。

S03. NVIDIA. *NVIDIA Kicks Off the Next Generation of AI With Rubin*. 5 January 2026. 关于该代设计与推理令牌估计成本的行业公告, 不是经审计财务比较。 <https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer>。

S04. OpenAI. *Introducing GPT-6 Sol and Luna*. September 2026; page updated 29 September 2026. API pricing, caching 等部分; 比较两代不同产品的价格。 <https://openai.com/index/introducing-gpt-6-sol-and-luna/>。

S05. Kwon, Woosuk, et al. *Efficient Memory Management for Large Language Model Serving with PagedAttention*. Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 2023). DOI:

10.1145/3600006.3613165. <https://doi.org/10.1145/3600006.3613165>。测试结果受论文所用硬件与负载条件限制。

S06. Yu, Gyeong-In, et al. *Orca: A Distributed Serving System for Transformer-Based Generative Models*.

USENIX OSDI 2022. <https://www.usenix.org/conference/osdi22/presentation/yu>。涉及队列处理与迭代调度。

S07. NVIDIA TensorRT-LLM. *Disaggregated Serving*. Technical documentation, current living version at cutoff. <https://nvidia.github.io/TensorRT-LLM/features/disagg-serving.html>。平台工程文档；不能据此认定阶段拆分始终是更优方案。

S08. William Stanley Jevons. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines*. 1865. 历史版本: <https://oll.libertyfund.org/titles/jevons-the-coal-question>。用于说明反弹假设的思想史，不是当代算力需求弹性的证明。

S09. AMD, Cisco, HUMAIN. *AMD, Cisco and HUMAIN Expand Saudi Arabia's AI Infrastructure as AMD Instinct Systems Go Live*. Joint company announcement, 31 August 2026. <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2026/m08/amd-cisco-and-humain-expand-saudi-arabia-ai-infrastructure.html>。一手来源本身具有商业利益；本文区分了已宣布投入运行的能力与规划容量。

S10. Saudi Press Agency. *HUMAIN Launches HUMAIN Chat, Powered by ALLaM 34B*. 25 August 2025. <https://www.spa.gov.sa/ar/N2385004>。产品发布公告，未提供公开收入数据。

S11. MLCommons. *MLCommons Sets Participation Record with New MLPerf Inference v6.1*. 16 September 2026. <https://mlcommons.org/2026/09/mlperf-inference-v6-1-results/>。衡量硬件与系统性能，不是资本回报研究。

S12. Safwan AlModhayan. *Reproducible Synthetic Pricing – Demand – Capacity Model*. `analysis/reproduce.py`、`data/model_assumptions.json`，以及本材料包内的结果表与测试。原创计算来源，不是市场观测来源。