



ИССЛЕДОВАТЕЛЬСКАЯ СТАТЬЯ

Редакционная дата: 24 октября 2026 года

Ловушка дефляции цены токена: предварительная модель экономической ценности коммерческой жизни ускоренных вычислений

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

Статус работы: фактически выполненная модель на синтетических данных, без первичных финансовых или операционных данных саудовского объекта. **Версия:** 1.0 | **Научное рецензирование не проводилось.**

Резюме на русском языке

Контекст: чипы искусственного интеллекта и программное обеспечение для инференса (Inference Serving) развиваются всё быстрее, тогда как инвестиции в вычислительную инфраструктуру ИИ (AI Infrastructure) амортизируются в течение сравнительно долгих договорных и бухгалтерских сроков. Смещение физического роста производительности, снижения себестоимости выпуска токена, цены его продажи и использования оплачиваемой мощности приводит к противоречивым выводам об эффективности капитала. **Вопрос:** при каких условиях программная эффективность повышает экономическую отдачу от инвестиций в ускоренные вычисления, а при каких рост спроса не компенсирует снижение цены? **Метод:** локальный обзор первичных

раскрытий и технологий инференса, а также выполненная шестилетняя модель дисконтированных денежных потоков (Discounted Cash Flow; DCF) с тремя условными темпами снижения цены, тремя значениями эластичности спроса и тремя скоростями роста мощности — всего 27 синтетических опытов; дополнительно проведено 18 тестов чувствительности и двоичный поиск условного порога эластичности. **Расчётные результаты:** 12 опытов дали положительную чистую приведённую стоимость (Net Present Value; NPV), 15 — отрицательную. В условном базовом сценарии NPV составляет −5,58 учётной единицы, хотя номинальная окупаемость достигается в пятом году; порог эластичности спроса, необходимый для перехода через ноль, равен примерно 0,7803. Если переменная стоимость единицы остаётся постоянной, а не снижается, NPV составляет −31,18. **Интерпретация:** эти числа не представляют распространённость, вероятность или доходность компаний; они лишь раскрывают логику взаимодействия спроса, мощности, затрат и дисконтирования при явно заданных допущениях. **Методологический вывод:** проверка саудовских инвестиций в вычисления требует данных о принятых задачах установленного качества, полной стоимости поставки, реальных контрактах и коммерческих сроках жизни по типам нагрузки. Числа процессоров, загрузки или лабораторной производительности недостаточно.

Ключевые слова: экономика вычислений (Computational Economics); графические процессоры (GPU); ловушка дефляции цены токена (Token Deflation Trap); ценовая эластичность спроса (Demand Price Elasticity); инференс (Inference Serving); стоимость принятой задачи (Cost per Accepted Task); чистая приведённая стоимость (NPV); Саудовская Аравия (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. Научный вопрос и его границы

Работа рассматривает **инвестиционную единицу вычислительной мощности для искусственного интеллекта, которую можно сдавать в аренду или использовать для продажи услуг**, а не экономику Саудовской Аравии в целом. Единица может представлять комплект оборудования и инференс-сервис в пределах заданных капитальных затрат (Capital Expenditure; Capex), но не соответствует реальному центру обработки данных или конкретной модели процессора. Её методологический смысл — различить траекторию, на которой техническое улучшение увеличивает маржу выручки, и траекторию, на которой выпуск растёт, но возмещение капитала ухудшается.

Главный вопрос формулируется так: **если фактическая цена единицы работы снижается, как эластичность спроса, реальная мощность, переменные затраты и сроки потоков совместно определяют приведённую стоимость фиксированной начальной инвестиции?** Из него следуют пять вопросов:

1. Достаточно ли падения цены токена, чтобы инвестор мог заключить, что снизилась стоимость выполнения задачи?
2. Когда рост спроса превращается в оплачиваемые единицы, а когда его ограничивают мощность, качество или время ответа?
3. Может ли улучшение программного обеспечения сделать существующее устройство прибыльнее при снижающейся цене?
4. Насколько результат инвестиции чувствителен к переменным затратам и ставке дисконтирования, а не только к номинальной мощности?
5. Какие общедоступные и разрешённые саудовские данные могли бы проверить эти связи, не раскрывая операционную нагрузку или тайны контрагентов?

Предлагаемая единица анализа для будущего применения — **пара «класс нагрузки × партия оборудования по времени выпуска» (Workload Class × Hardware Vintage)**. Если такие пары объединяются в совокупный показатель, необходимо указывать

распределение нагрузок, а не сводить их к среднему, оторванному от типа услуги.

2. Пять сроков жизни, которые часто смешивают

Мы различаем физический срок жизни (Physical Life), в течение которого оборудование сохраняет работоспособность; бухгалтерский срок полезного использования (Accounting Useful Life), определяющий амортизацию; договорный срок покрытия (Contractual Coverage), в течение которого стороны обязаны платить за мощность; технический срок пригодности с учётом нагрузки (Workload-Adjusted Technical Life), пока платформа выполняет заданные нагрузки при требуемых качестве и времени; и срок коммерческого преимущества (Commercially Advantageous Life), пока она способна конкурировать с альтернативами по цене, обеспечивающей приемлемую маржу.

Это разделение не является официальной классификацией саудовского ведомства или общей оценкой рынка вычислений. Это предлагаемое аналитическое определение, позволяющее сравнивать активы с разными сроками жизни. Годовой отчёт CoreWeave показывает, зачем такое разделение нужно: с января 2023 года компания увеличила расчётный срок эксплуатации вычислительного оборудования с пяти до шести лет, сославшись на развитие аппаратной и программной части и проектирование центров обработки данных; на конец 2025 года средневзвешенная продолжительность её твёрдых контрактов составляла около пяти лет S01. Вычитать одну цифру из другой и объявлять разницу «годом прибыли» нельзя: контракт не равен отдельному платежу за актив, а средневзвешенный срок не является датой завершения всех обязательств.

В 2026 году NVIDIA представила Rubin, заявив о снижении стоимости инференса до десяти раз относительно Blackwell в архитектуре эталонных нагрузок; в сентябре OpenAI объявила цены новых продуктов API, которые в выбранной категории сравнения оказались ниже цен прежних продуктов S03S04. Первое — **рекламный эталон стоимости от производителя оборудования**; второе — **опубликованные отпускные цены поставщика услуги**; третья возможная величина, которую мы не наблюдали, — **реальная цена после учёта контрактов, скидок и использования**. Переход от одной величины к другой — ошибка определения переменной, а не расхождение мнений.

3. Техническая литература: почему ценность возникает между памятью и очередью?

Достижимая производительность зависит не только от вычислительного пика. Её могут повышать методы, не требующие полной замены оборудования:

Управление памятью: алгоритм PagedAttention и система vLLM в работе Квона и соавторов (Kwon et al., 2023) используют разделение и повторное совместное использование кэша ключей и значений (Key-Value Cache; KV Cache), чтобы уменьшить внутреннюю фрагментацию и неиспользуемое резервирование. В экспериментах статьи пропускная способность (Throughput) выросла примерно в два-четыре раза относительно отдельных базовых систем при сходной задержке S05. Чтобы воспроизвести этот коэффициент в Саудовской Аравии, нужны спецификации платформы, нагрузок и версий; статья не обосновывает прямое коммерческое обобщение.

Планирование: Orca (2022) предложила планирование на уровне каждого цикла генерации (Iteration-Level Scheduling), сокращающее задержку коротких запросов за длинными и позволяющее выборочную пакетную обработку (Selective Batching) S06. Это напоминает, что ожидание зависит от программного решения не меньше, чем от мощности процессора.

Разделение стадий инференса: техническая документация TensorRT-LLM различает заполнение контекста (Prefill) и генерацию (Decode) и объясняет, почему этим стадиям могут подходить разные ресурсы и схемы параллелизма. В ней также указаны издержки переноса инференс-кэша при разделении: решение не бесплатно и не оптимально для каждого запроса S07.

Точка тестирования: в сентябре 2026 года MLCommons объявила набор MLPerf Inference 6.1, включающий тесты многошаговых задач S11. Результаты могут помочь построить единые линии измерения производительности, но сами по себе не являются чистой приведённой стоимостью или коммерческим сравнением после учёта электричества, финансирования и условий продажи.

Работа выводит возможный механизм, а не причинное доказательство: улучшение планирования, управления памятью и вычислительных стадий увеличивает число услуг, которые можно получить от существующего актива. Какая доля прироста будет оплачена, зависит от спроса, поведения клиента, качества результата и конкуренции. Выгода может достаться оператору вычислений, разработчику приложения, конечному пользователю или распределиться между ними.

4. Конкурирующие гипотезы

- **H1, эффект мощности:** если ограничением служит мощность, а платёжеспособный спрос способен поглотить дополнительный выпуск, рост поставляемой мощности увеличивает чистую стоимость при условии, что затраты на расширение и эксплуатацию не перевешивают выгоду.

- **H2, эффект дефляции цены:** при низкой эластичности спроса выручка может быстро снижаться, несмотря на рост числа единиц, а окупаемость — усложняться в заданном финансовом горизонте.
- **H3, решающий эффект отдачи:** при высокой эластичности спроса и достаточной мощности совокупная выручка после снижения цены может вырасти. Это не означает, что прибыль бизнеса обязательно увеличится.
- **H4, контракт против спотового рынка:** многолетние обязательства приобрести мощность меняют чувствительность будущей выручки к спотовой цене услуги, однако права на поставку, нарушения, скидки и продления делают защиту условной.
- **Конкурирующая гипотеза R1:** изобилие вычислений может сочетать снижение цены единицы с сильным ростом прибыли, если новые высокоценные функции поглотят мощность; тогда ловушка не проявится в данных о фактической прибыли.
- **Нулевая гипотеза H0:** после учёта состава нагрузок, электричества, финансирования, контрактов и качества скорость снижения цены единицы не даёт дополнительной объяснительной силы для доходности инвестиции. Эту гипотезу нельзя отвергать без реальных операционных данных и подходящего дизайна причинного вывода.

Это гипотезы и вопросы интерпретации. Наши вычислительные тесты проверяют их согласованность внутри модели, состоящей из допущений, а не на выборке людей или реальных объектов.

5. Реализованная модель и единицы измерения

Входные данные и обозначения: для каждого года $t = 1, \dots, 6$ величина p_0 обозначает начальный индекс цены 1; q_0 — годовой спрос 50 единиц; k_0 — поставляемую годовую мощность 50; c_0 — начальные переменные затраты 0,20 на единицу; F — постоянные годовые операционные затраты 7; i_0 — начальную инвестицию 120; s — условную остаточную стоимость 8 в конце шестого года; r — ставку дисконтирования 12%. Всё это **условные индексные единицы:** они не означают реальные доллары, риалы, мегаватты, устройства или токены.

Проверяемые переменные: δ — ежегодный темп снижения цены единицы; ε — абсолютная величина ценовой эластичности спроса; g — темп роста продаваемой мощности вследствие операционных улучшений, который предполагается постоянным. Программно применены следующие формулы:

$$p_t = p_0(1 - \delta)^{t-1}$$

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

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

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

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

D_t — потенциальный спрос, K_t — предел квалифицированной услуги, x_t — оплачиваемые единицы, CF_t — операционный денежный поток с остаточной стоимостью в последнем году. Мы предположили, что одна транзакция или один запрос равен единице и что каждая поставленная единица отвечает требованиям качества и обслуживания. Это **удобное и не обязательно реалистичное** допущение, которое в будущем потребуется выделить в эмпирическом исследовании. Функция минимума не позволяет учитывать выручку от спроса, для которого не нашлось мощности. Модель не включает долг, проценты, налоги, изменение заработной платы, бухгалтерскую амортизацию, банкротство стороны или отказ оборудования; стоимость капитала выражена условной ставкой дисконтирования, а не ценой кредита, выведенной из американского раскрытия.

Допущение $c_t = c_0 / (1+g)^{(t-1)}$ требует особой осторожности: мы предполагаем, что программный рост производительности снижает переменные затраты на единицу тем же темпом. Это может быть неверно, если улучшение в основном позволяет запускать больше нагрузок, не уменьшая фактический расход энергии или сетевые затраты. Поэтому мы провели тест чувствительности, в котором c_t остаётся равной c_0 при любой динамике мощности. Более консервативный результат приведён в разделе 8.

6. Данные, исполнение и воспроизводимость

Исследование не содержит банковских транзакций или показателей загрузки саудовского оператора. Все экспериментальные значения берутся из машиночитаемого файла `data/model_assumptions.json`; расчёты выполняет скрипт `analysis/reproduce.py` на стандартной библиотеке Python. Запуск создаёт пять фактических записей:

1. `base_annual_cashflows.csv`: шесть строк годовых потоков, объёма, мощности, выручки и дисконтирования.
2. `synthetic_scenario_results.csv`: сетка из 27 независимых случаев по схеме $3 \times 3 \times 3$.
3. `sensitivity_checks.csv`: 18 дополнительных случаев, сочетающих три ставки дисконтирования (8%, 12%, 16%), три темпа роста мощности (0%, 25%, 50%) и два режима переменных затрат — снижение и постоянство.
4. `derived_checks.json`: производная сводка порога, базового результата и числа случаев.

5. Графические файлы SVG и PNG, построенные по итоговым таблицам, а не по оценочным рисункам.

Для проверки доступна электронная таблица `evidence_workbook.xlsx`, где изменяемые входные данные отделены от расчётов и снимков результатов. Это средство просмотра, а не замена скрипту как опубликованной вычислительной записи полного эксперимента. Мы использовали годовые потоки в конце года, инвестицию в момент времени ноль и остаточную стоимость только в шестом году. Синтетические числа не получают финансовых данных из внешних источников; внешние источники объясняют, почему механизмы стоит проверять, и ограничивают перенос выводов в реальность.

7. Фактические результаты эксперимента: почему спрос побеждает не всегда?

Выполнена сетка с ежегодным снижением цены на 10%, 22% или 35%; эластичностью спроса ϵ 0,5, 1,0 или 1,8; и приростом мощности 0%, 25% или 50%. Каждая комбинация даёт шесть годовых периодов и итоговую чистую приведённую стоимость. Значения не выбирались случайно, вероятностное распределение не оценивалось; поэтому доля прибыльных случаев в сетке **не является ожидаемой частотой успеха в реальности**. Это лишь подсчёт участков выбранного нами пространства параметров.

Базовый сценарий: цена снижается на 22% в год, эластичность равна 0,7, мощность растёт на 25% в год и затраты на единицу снижаются тем же темпом, ставка дисконтирования — 12%. Годовые результаты приведены в расчётном виде, а не как прогноз редактора:

Год	Индекс цены	Потенциальный спрос	Доступная мощность	Проданные единицы	Условная выручка	Годовой поток с остаточной стоимостью, если она есть
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*

Шестой год включает условную остаточную стоимость в восемь единиц. Все столбцы содержат синтетические индексные числа, округлённые до двух или трёх знаков. Точный расчёт находится в CSV.

Спрос более чем удваивается, хотя годовая выручка снижается с 50 примерно до 34,44. Общая выручка за шесть лет составляет около **251,02 единицы**, однако выручка не равна прибыли: переменные и постоянные затраты вместе с временной стоимостью начальной инвестиции оставляют **чистую приведённую стоимость около –5,5835**. Накопленные недисконтированные потоки после вычета первоначальных затрат становятся неотрицательными в пятом году, но приведённая стоимость остаётся отрицательной, поскольку позднее возвращение денег весит меньше. Нельзя смешивать *Payback Period* и *NPV*.

В следующем выбранном сравнении ежегодное снижение цены на 22% и рост мощности на 25% остаются постоянными, меняется только эластичность:

Предполагаемая эластичность спроса	NPV за шесть лет в условных единицах	Выручка шестого года	Аналитический результат
0.5	–17.98	26.87	Рост объёма недостаточен, чтобы возместить снижение цены.
1.0	+8.91	44.05	Более высокий спрос заполняет продаваемую мощность.
1.8	+8.91	44.05	Результат тот же, поскольку поставку ограничивает мощность.

Равенство строк 1,0 и 1,8 не является произвольным округлением: при этих допущениях во все годы после первого доступный спрос превышал оплачиваемую мощность, поэтому $x_t = k_t$; увеличение готовности покупать не меняет фактический выпуск. Следует различать рост **потенциального рынка**, рост **поставляемого объёма** и рост **фактически оплаченного объёма**. Если при той же цене услуга недоступна, интенсивность интереса не компенсирует дефицит, пока нельзя расширяться или поднять цену в пределах готовности клиента.

Граница NPV в синтетической модели

Знак чистой приведённой стоимости меняется при разных допущениях о снижении цены и эластичности спроса

		10%	22%	35%
Абсолютная ценовая эластичность спроса (допущение)	0.5	+9.8 профицит	-18.0 дефицит	-48.0 дефицит
	1.0	+28.4 профицит	+8.9 профицит	-46.9 дефицит
	1.8	+66.4 профицит	+8.9 профицит	-46.9 дефицит

Годовое снижение цены оплачиваемой единицы (допущение)

Во всех ячейках: рост поставляемой мощности 25% в год; ставка дисконтирования 12%; первоначальные вложения 120 индексных единиц.

Синтетические индексные единицы за шесть лет; не SAR, USD и не наблюдаемая доходность GPU.

Сетка чистой приведённой стоимости в условных сценариях

Рисунок 1: матрица результатов фактически выполненного эксперимента с разными значениями эластичности и цены при постоянном предполагаемом росте мощности на 25% в год. Числа не являются вероятностными распределениями или оценками саудовских организаций.

Полный эксперимент дал **12 сценариев с положительной и 15 с отрицательной стоимостью** из 27; знак менялся по трём осям. Когда мощность на все шесть лет зафиксирована на уровне 50 единиц ($g=0$), все результаты при выбранных настройках становятся отрицательными: спрос растёт, но поставить больше 50 нельзя, а цена продолжает снижаться. На другом краю сетки один сценарий высокой эластичности и роста мощности ($\delta=.22$, $\epsilon=1.8$, $g=.50$) дал около **+113,75**, однако этот результат обусловлен чрезвычайно быстрым предполагаемым ростом программной производительности несколько лет подряд и не служит аргументом для реальной инвестиции на этих числах.

8. Результаты тестов чувствительности и условия смены вывода

Фактически проверены ставки дисконтирования 8%, 12% и 16%, рост мощности 0%, 25% и 50%, а также два режима переменных затрат на единицу — снижение по принятому темпу и постоянный уровень. Для случая с падением цены на 22%, эластичностью 0,7 и ростом мощности на 25% получены следующие результаты:

Ставка дисконтирования	Переменные затраты снижаются вместе с ростом мощности	Переменные затраты остаются постоянными
8%	+7.84	-22.46
12%	-5.58	-31.18
16%	-16.82	-38.62

Интерпретация важнее запоминания значений: если убрать допущение о снижении переменных затрат на единицу, одного роста мощности недостаточно для предполагаемой доходности, потому что выручку определяет купленный спрос, а не объём работы, который программа способна запустить. В некоторых комбинациях увеличение σ с 25% до 50% вообще перестаёт влиять на выручку после достижения требуемого предложения, пока не меняются цена и спрос. Это не отрицание ценности программного обеспечения, а вопрос о том, **какое именно ограничение мы сняли**: вычислительное, затратное или клиентское?

Мы провели итеративный двоичный поиск (Binary Search) по ε в диапазоне от 0 до 2,5, оставив прочие параметры базового сценария неизменными, чтобы найти наименьшую эластичность, при которой $NPV \geq 0$. Она составила примерно **0,7803**. Это не оценка эластичности спроса в Саудовской Аравии или мире, а **порог внутри конкретной модели**. Изменение первоначальной инвестиции, цены, качества поставки, свойств спроса или горизонта дисконтирования изменит и его. Порог можно пересчитать по `data/derived_checks.json` и скрипту, а не по имени известного экономического эффекта.

Границы этих тестов: мы не моделировали случайные распределения роста цены электричества, сетевой перегрузки, потери контрактов или географических различий центров обработки данных. Статистических выборок нет, поэтому нет значений p , доверительных интервалов или заявлений о статистической значимости. Нельзя утверждать и то, что 18 случаев покрывают всё пространство рисков. Тесты включали три пересекающихся набора явно заданных параметров.

Разрыв между ценой и мощностью

Базовый сценарий: $\delta=22\%$; $\epsilon=0.7$; поставляемая мощность растёт на 25% в год



* Поток 6-го года включает предполагаемую остаточную стоимость 8 единиц; первоначальные вложения 120 единиц в нулевом году не показаны.

ГИПОТЕТИЧЕСКИЕ РЕЗУЛЬТАТЫ: без измеренной мощности, цен или денежных потоков проекта.

Траектории цены, спроса, мощности и выручки в условном эксперименте

Рисунок 2: результаты базового сценария. Рост мощности, старение цены и время денежного потока — разные явления; единица финансовой оси является лишь составным индексом.

9. Проверка применимости к Саудовской Аравии без смешения данных разных компаний

Есть три категории первичных фактов, имеющих отношение к эмпирической проверке, но они не образуют единого ряда измерений:

А. Американское коммерческое раскрытие, а не саудовское измерение: годовой отчёт CoreWeave за 2025 год показывает различие между бухгалтерским сроком оборудования и сроком контрактов. В отчётности за 2026 год объём оставшихся обязательств к исполнению (RPO) составил **103,7 млрд долларов на 30 июня 2026 года**; предполагалось признать 41% суммы в первые 24 месяца после этой даты, 39% — в месяцы 25–48, а остаток — до 78-го месяца S01S02. В раскрытии также приведена общая сумма будущих платежей по основной сумме долга — **35,551 млрд долларов** согласно полному графику погашения S02. Складывать или делить эти числа, чтобы получить коэффициент защиты или сделать вывод о кризисе либо излишке, нельзя. Первое относится к договорным услугам, которые ещё не оказаны и подчиняются правилам признания выручки; второе — к финансовым обязательствам с разными сроками. Это

не два противоположных чистых потока. Долгосрочные контракты могут защищать оператора от спотовой цены, но мы не анализировали их отдельные условия или историю исполнения.

Б. Объявление о вводе в эксплуатацию в Саудовской Аравии от коммерческих участников: 31 августа 2026 года AMD, Cisco и HUMAIN объявили, что системы на процессорах AMD Instinct MI355X и сетях Cisco Silicon One в Саудовской Аравии введены в промышленную эксплуатацию для обслуживания клиентов HUMAIN внутри страны и за её пределами S09. В том же объявлении говорится о намерении развернуть мощность **до 250 мегаватт** начиная с 2027 года и достичь **одного гигаватта** к 2030 году. Это планируемые мощности, а не опубликованный результат эксплуатации за 2026 год. Поскольку сообщение исходит от заинтересованных сторон, мы прямо им его приписываем и не считаем независимым аудитом спроса, затрат или маржи.

В. Саудовская услуга для пользователей: 25 августа 2025 года Saudi Press Agency (SPA) объявила о запуске HUMAIN Chat на основе ALLaM 34B с арабоязычным интерфейсом и голосовыми инструментами S10. Согласно источнику, это подтверждает запуск продукта и ставит вопрос о превращении инвестиции в мощность в полезный спрос, но не позволяет заключать о прибыли, выручке, числе токенов или статистике использования. Случай можно использовать как контекст общественного применения, но нельзя вводить в нашу модель как неопубликованное числовое наблюдение.

Саудовский вопрос состоит не в буквальном повторении истории другой страны, а в том, способны ли вычислительные объекты продавать **прикладной результат** на внутреннем и трансграничном рынках при заданных уровнях обслуживания и меняющихся затратах. Экономика обучения фундаментальной модели (Foundation Model Training) может отличаться от экономики обслуживания инференса (Inference Serving), поэтому одна цена токена не годится для оценки активов, предназначенных для крупномасштабного обучения.

10. Различие между «ценой», «затратами» и «ценностью»: предлагаемая система измерений

Чтобы превратить эту работу в эмпирическое исследование, необходимо развести три учётные единицы, которые смешиваются в публичном обсуждении:

1. **Затраты поставщика (Provider Cost):** фактические расходы на выполнение единицы, включая энергию, сеть, программные компоненты, поддержку, неиспользованную мощность и влияние амортизации и финансирования при подходящем распределении. Это не только предельные операционные затраты.

2. **Фактическая цена продажи (Realized Selling Price):** сумма, которую поставщик действительно получает после учёта контрактов, скидок, кэширования и разных схем выставления счетов. Она не обязательно равна прейскурантной цене (List Price), опубликованной в интернете.

3. **Ценность для клиента (Value to Customer):** польза выполненной задачи, которую можно оценивать сэкономленным временем, производительностью, качеством или связанной выручкой. Она требует отдельного измерения и не равна цене автоматически.

Поскольку запросы неоднородны, предлагается вместе со стоимостью миллиона токенов сообщать **стоимость принятой задачи** (Cost per Accepted Task). Принятая задача должна быть заранее определена требованиями к точности, безопасности и времени. Особого внимания требуют доля повторно выполненных задач, различия длины контекста, коэффициент попаданий входных данных в кэш (Cache Hit Rate), число выходных токенов, ранняя остановка, недоступность ресурсов и расход электричества, рассчитанный на уровне нагрузки по документированной методике.

Для типа нагрузки w и периода t описательная формула имеет вид:

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

Здесь E — расчётные затраты на энергию и охлаждение; N — затраты на сеть и данные; O — распределённые затраты на эксплуатацию и поддержку; A — доля использования капитала; J — число задач, прошедших критерий принятия в периоде. Формула **предложена и не утверждена**. Она требует политики распределения затрат и подтверждения отсутствия двойного учёта. Если в знаменателе нет принятой задачи, нельзя показывать фиктивную «нулевую» стоимость или делить на ноль. В анализ полного проекта следует добавить нераспределённые расходы и отложенную выручку, а не выдавать частичный показатель за чистую прибыль.

Ещё два показателя — маржа принятой задачи, выполненной в согласованных временных пределах (SLA-Compliant Contribution Margin), и доля услуг, создающих поток по твёрдому контракту, относительно выручки, зависящей от спотового рынка (Spot-Market Revenue). Для каждого показателя необходимо указывать период, определение услуги, фактические цены, валюту и метод распределения расходов.

11. Дизайн следующего саудовского исследования, способного опровергнуть гипотезу

Предлагаемый, но не выполненный дизайн: ежемесячные наблюдения за агрегированными классами корпоративных нагрузок и последовательными партиями оборудования в течение 24–36 месяцев, если данные станут доступны, при документированном добровольном сотрудничестве операторов и разработчиков с разными типами использования. Первый предварительный анализ должен быть описательным. Причинный дизайн возможен лишь при ясно изолируемом моменте изменения — например, когда программное обеспечение инференса обновляется на части набора нагрузок, а сопоставимая группа одновременно не меняется. Обновление нельзя произвольно называть «рандомизированным экспериментом», если реального случайного распределения не было.

Минимальный состав будущей таблицы данных: обезличенный идентификатор партии оборудования; класс и версия модели; диапазоны входа, выхода и доли кэширования; доля принятых задач; время до первого токена и время выполнения запроса в рамках уровня обслуживания; оплачиваемые единицы работы; средняя фактическая цена и её распределение; электричество и охлаждение, отнесённые по раскрытой методике; доля договорной цены; график амортизации и выбранная экономическая ставка дисконтирования; даты начала и завершения нагрузки. Таблица не должна содержать адреса критических объектов, расположение серверных, пользовательские журналы, договоры физических лиц или ключи доступа. Данные, способные раскрыть чувствительные операционные схемы использования, следует задерживать или агрегировать.

Этапы сравнения:

1. До просмотра результатов определить от трёх до пяти классов нагрузки: короткий интерактивный инференс, длинные контексты, пакеты неоперативной обработки, дообучение и многошаговые задачи, если их сбор разрешён. Это примеры, которые можно скорректировать в зависимости от доступных данных.
2. Для каждого класса провести отдельный стандартный тест качества, чтобы рост производительности за счёт правильности результата не исчезал из знаменателя.
3. Разделить выборку по времени: прежний период использовать для калибровки модели, последующий период вне выборки (Out-of-Sample) — для оценки; исключить утечку будущих цен и поколений оборудования, недоступных на момент прогноза.
4. Оценить кривую фактической цены и оплаченного объёма с переменными, показывающими изменение качества, доступность мощности и контракты, вместо оценки эластичности по одной прейскурантной цене.

5. По возможности сравнивать эффект программных обновлений внутри одной партии оборудования и постоянной нагрузки, отслеживая изменения состава клиентов, времени и расходов на энергию; при отсутствии разумной контрольной группы представлять результат только как корреляцию.
6. Включить сроки погашения капитальных обязательств и различия договорных условий в финансовую модель с документированными допущениями, сопоставляя результаты с объёмом фактических платежей, чтобы не выводить прибыль из одного технического показателя.

Предлагаемый причинный показатель, если данные позволят: разница в изменении стоимости принятой задачи между нагрузками, получившими программное улучшение, и сопоставимыми нагрузками, не получившими его в тот же период, при допущениях параллельных предшествующих трендов (Parallel Trends) и отсутствия иных дифференцированных изменений. Однако центры обработки данных используют общие ресурсы, поэтому эффект обновления может распространяться на другие нагрузки через очереди или общую память; допущение об отсутствии взаимного влияния единиц (No Interference) может оказаться неверным. В качестве альтернативы можно сравнивать целые пулы ресурсов с явным описанием распространения эффекта или ограничиться описанием механизма.

Важная встречная проверка: одновременно с обновлением инференс-программы могут появиться другие клиенты, меньшая версия модели или новый контракт по более низкой цене. Если не зафиксировать эти изменения, программе можно ошибочно приписать причинный эффект, вызванный сменой состава или переговорной силы. Исследование ценовой эластичности должно учитывать неудовлетворённый спрос: наблюдатель, считающий только завершённые бронирования, может счесть спрос слабым, хотя другие заявки отклоняются из-за полной загрузки.

12. Где «ловушка дефляции цены токена» может не сработать?

Привлекательное название не должно превращаться в допущение истинности. Есть шесть ограничений, ослабляющих аргумент или меняющих вывод на противоположный:

Во-первых, процессор не равен конечному продукту. В цену услуги входят программное обеспечение, безопасность, поддержка, права доступа и пользовательский интерфейс; ценность этих элементов может расти при снижении стоимости процессора. Падение внутренней цены токена не обязательно ведёт к падению общей выручки компании.

Во-вторых, работу можно переносить. Процессоры, уже не подходящие для обучения передовой модели (Frontier Model), могут обслуживать меньшие модели или более дешёвые инференс-задачи. Отставшее от вершины устройство не обязательно стоит ноль. Модель намеренно не позволяет переносить нагрузки к новым пользователям с иным ценообразованием; это ограничивающее допущение.

В-третьих, меняется тип задачи. Приложению нового поколения может понадобиться инференс с несколькими инструментами или длинная контекстная память, поэтому стоимость задачи вырастет даже при снижении цены токена. Если меняется число токенов, нужных для результата, цена миллиона токенов перестаёт представлять экономическое качество услуги.

В-четвёртых, контракты. Высокая доля договорной выручки может отделить период окупаемости от спотовой цены, тогда как эффект цены проявится позже, при продлении. Возникает временной разрыв, который нужно моделировать. В базовом сценарии для демонстрации механизма мы предположили переменную цену всех единиц без этой защиты.

В-пятых, невычислительные ресурсы. Развёртывание может ограничивать разрешение на подключение энергии, задержка подстанции, вода для охлаждения или сетевое оборудование, и тогда теоретическая программная мощность не станет услугой. В обратном направлении алгоритмы памяти и параллелизма могут уменьшить потребность в новом оборудовании. Мы не оценивали ни одно из этих ограничений эмпирически.

В-шестых, обилие спроса и альтернативные ограничения. Если снижение цены создаёт новые полезные задачи, а рынок эластичен, выручка и маржа могут расширяться. Но если в краткосрочном периоде мощность ограничена, вместо количества могут меняться цены и контракты. Наша модель изолирует одно ограничение и не оценивает общее равновесие между конкурирующими компаниями.

Чтобы разделить эти факторы, недостаточно однажды ответить на вопрос «на сколько процентов в год падают цены?». Необходимо наблюдать версии, задачи, сроки контрактов, скрытый спрос и переменные затраты, а также проверять альтернативные модели, включая спрос без постоянной эластичности и цены, которые не снижаются на отдельных нагрузках.

13. Предлагаемый различающий тест перед новой инвестицией

Исследование предлагает неинвазивный опыт для инвестиционного решения, не требующий расположения объекта или чувствительных данных о потоках.

Уполномоченный оператор выбирает два класса нагрузки: один, где критично время ответа, и другой, где важен размер пакета. На разрешённых ресурсах проводятся три сравнения: старая и новая версия программы на одном оборудовании; старое и новое оборудование на одной программе, если это возможно; перенос части нагрузок между ними в разрешённых пределах. Фиксируются модель ИИ, диапазон качества, время ответа и данные фактических цен. Периоды должны быть достаточно длинными, чтобы наблюдать изменение спроса, а не только краткий пик.

Если программное обновление снижает стоимость принятой задачи и повышает общую маржу оператора без снижения отпускной цены, наблюдение поддерживает траекторию программного спасения. Если цена продажи падает быстрее полной стоимости и окупаемость ухудшается несмотря на рост числа запросов, оно поддерживает траекторию ловушки. Если объём оплачиваемых задач быстро растёт вместе с маржой, то для этого класса опровергается гипотеза о неизбежной убыточности ценовой дефляции. Если после учёта факторов маржа не меняется, это ограничивает масштаб утверждения, а не означает личную неудачу исследователя.

Можно проверить **правило решения**, не выдавая его за действующий стандарт: прежде чем финансировать новую партию вычислений, сравнить условную приведённую стоимость трёх вариантов — (а) купить новое оборудование; (б) улучшить программу существующей услуги; (в) временно арендовать внешнюю мощность для измерения реального спроса. Каждый вариант включает стоимость финансирования, риски, энергопотребление, срок разработки и выручку от принятых услуг. Победить может любой — в зависимости от нагрузки и контрактов. Это практическая рамка сравнения внутри статьи M12, а не общая «система управления» и не рекомендация конкретной инвестиции.

14. Что можно и чего нельзя говорить руководству

Можно сказать: «Снижение технической стоимости инференса и улучшение программного обеспечения меняют экономику долгоживущих активов и требуют проверки эластичности оплачиваемого спроса и времени возврата капитала». Можно сказать и так: «В Саудовской Аравии в 2026 году объявлено о работе разнородного оборудования и последующих целях расширения, однако подробные показатели маржи в изученных материалах не опубликованы».

Нельзя говорить: «Любой графический процессор обесценится меньше чем за год»; «контракты исключают убыток»; «12 случаев из 27 означают вероятность прибыли 44%»; «синтетический NPV измеряет саудовские проекты». Нельзя также приравнять заявленную стоимость будущей договорной выручки американской компании к текущей ликвидности или планируемые к 2030 году мегаватты — к использованной мощности либо заработанной прибыли 2026 года.

Текущий условный вывод: снижение цены не является технической неудачей и может принести пользователям огромную общественную и экономическую пользу. Оно становится капитальным риском, когда оплачиваемый спрос, фактические затраты и сроки не успевают за ценами. С другой стороны, хорошее программное обеспечение, продуманные контракты и перераспределение нагрузок могут повысить жизнеспособность существующих активов. Доступные публичные данные не позволяют различить эти случаи на репрезентативном саудовском рынке, поэтому рекомендацией исследования остаётся **измерение, направленное на конкретное решение**, а не предсказание победы или поражения.

15. Заявление о воспроизводимости и этике данных

Вычислительный эксперимент локально воспроизводится одной командой из каталога статьи:

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

Для запуска нужна только стандартная среда Python; это зафиксировано в `analysis/requirements.txt`. Файл `data/model_assumptions.json` содержит все выбранные значения; в результате создаются `data/base_annual_cashflows.csv`, `data/synthetic_scenario_results.csv`, `data/sensitivity_checks.csv` и `data/derived_checks.json`. Любое изменение допущений порождает другие результаты; публикующий их обязан назвать версию модели и новые значения. Для создания графиков нужна дополнительная библиотека визуализации, но она не требуется для проверки числовых формул. Файл `evidence_workbook.xlsx` предоставляет независимую табличную проверку базового сценария, если среда электронных таблиц поддерживает пересчёт формул.

Личные данные, чувствительная операционная информация и частные карты зависимостей не собирались. Не заявляются институциональные одобрения, научное рецензирование, эмпирическое подтверждение в Саудовской Аравии или доступ к неопубликованным контрактам. Все реальные данные взяты из публичных раскрытий,

статей или объявлений; коммерческие заявления представляют позиции их сторон, а не независимый сторонний аудит. Синтетические данные созданы специально, чтобы показать возможные причинные связи **внутри уравнений**.

16. Первичные и справочные источники

S01. CoreWeave, Inc. *2025 Annual Report on Form 10-K*. U.S. Securities and Exchange Commission. Разделы Accounting Policies: Property and Equipment; Committed Contracts; Take-or-Pay. <https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/crww-20251231.htm> . Финансовая отчётность на 31 декабря 2025 года; отчёт стал доступен в 2026 году.

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/crww-20260630.htm> . Дата наблюдения — 30 июня 2026 года.

S03. NVIDIA. *NVIDIA Kicks Off the Next Generation of AI With Rubin*. 5 January 2026. Отраслевое заявление об архитектуре поколения и расчётной стоимости инференс-токена, а не аудированное финансовое сравнение. <https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer> .

S04. OpenAI. *Introducing GPT-6 Sol and Luna*. September 2026; page updated 29 September 2026. Разделы API pricing, caching; сравнение цен продуктов двух поколений. <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> . Результаты тестов зависят от оборудования и нагрузок статьи.

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> . Обработка очередей и планирование циклов.

S07. NVIDIA TensorRT-LLM. *Disaggregated Serving*. Technical documentation, current living version at cutoff. <https://nvidia.github.io/TensorRT-LLM/features/disagg-serving.html> . Документация инженеров платформы не означает, что разделение всегда является лучшим выбором.

- S08.** William Stanley Jevons. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines*. 1865. Историческая копия: <https://oll.libertyfund.org/titles/jevons-the-coal-question> . История гипотезы эффекта отдачи, а не современное доказательство эластичности спроса на вычисления.
- 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> . Первичный источник исходит от заинтересованных сторон; введённая в эксплуатацию и планируемая мощность разделены.
- S10.** Saudi Press Agency. *HUMAIN Launches HUMAIN Chat, Powered by ALLaM 34B*. 25 August 2025. <https://www.spa.gov.sa/ar/N2385004> . Объявление о пользовательском продукте без опубликованных данных о выручке.
- 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/> . Измерение производительности оборудования и систем, а не исследование доходности капитала.
- S12.** Сафван Аль-Мудайян. *Reproducible Synthetic Pricing–Demand–Capacity Model*. `analysis/reproduce.py` ВМЕСТЕ С `data/model_assumptions.json`, таблицами результатов и тестов в этом пакете. Оригинальный вычислительный источник, а не источник рыночных наблюдений.