



ورقة بحثية استكشافية

التاريخ التحريري: 24 أكتوبر 2026

فخ انكماش سعر الرمز: نموذج مبدئي للقيمة الاقتصادية للعمر التجاري للحوسبة المسرّعة

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

حالة العمل: نموذج منقذ فعليًا على بيانات اصطناعية، دون بيانات مالية أو تشغيلية أصلية من منشأة سعودية. نسخة: 1.0 | لم يُحكّم علميًا.

الملخص العربي

الخلفية: يتسارع تطور رقائق الذكاء الاصطناعي وبرامج الاستدلال (Inference Serving)، بينما تُستهلك استثمارات البنية الحاسوبية (AI Infrastructure) على مدى تعاقدية ومحاسبية طويلة نسبيًا. يؤدي الخلط بين تحسين الإنتاجية الفيزيائية، وانخفاض تكلفة إنتاج الرمز، وسعر بيعه، واستخدام القدرة القابلة للفوترة إلى استنتاجات متعارضة عن جدوى رأس المال. **السؤال:** تحت أي شروط تحسّن الكفاءة البرمجية العائد الاقتصادي لاستثمارات الحوسبة المسرّعة، وتحت أي شروط يعجز الارتداد في الطلب عن تعويض هبوط السعر؟ **المنهج:** مراجعة موضوعية لإفصاحات أصلية وتقنيات الاستدلال، مع تنفيذ نموذج تدفقات نقدية مخصومة (Discounted Cash Flow; DCF) لست سنوات، بثلاثة معدلات افتراضية لانخفاض السعر، وثلاث مروّات طلب، وثلاث سرعات لتحسن السعة، بإجمالي 27 تجربة تركيبية، و18 اختبار حساسية إضافيًا وسعي ثنائي لحساب عتبة مروّنة مشروطة. **النتائج المحسوبة:** تتوزع التجارب إلى 12 حالة قيمة حالية صافية (Net Present Value; NPV) موجبة و15 سالبة؛ تعطي الحالة المرجعية الافتراضية قيمة قدرها -5.58 وحدة محاسبية رغم بلوغ الاسترداد الاسمي خلال السنة الخامسة؛ وتعادل عتبة مروّنة الطلب الضرورية لعبور الصفر في هذه الحالة نحو 0.7803. إذا أبقينا التكلفة المتغيرة للوحدة ثابتة بدل تحسينها، تصبح القيمة -31.18. **التفسير:** لا تمثل هذه الأعداد انتشارًا أو احتمالًا أو عوائد شركات؛

إنما تكشف منطق تفاعل الطلب والسعة والتكلفة والخصم وفق افتراضات معلنة. **النتيجة المنهجية:** يستلزم اختبار استثمارات الحوسبة السعودية بيانات عن مهام مقبولة ذات جودة، والتكلفة الكاملة للتسليم، والعقود الفعلية، والأعمار التجارية حسب حمل العمل؛ لا يكفي عدد المعالجات أو نسبة الإشغال أو الأداء المخبري.

الكلمات المفتاحية: الاقتصاد الحاسوبي (Computational Economics)؛ وحدات المعالجة الرسومية (GPU)؛ فخ انكماش سعر الرمز (Token Deflation Trap)؛ مرونة الطلب السعرية (Demand Price Elasticity)؛ الاستدلال (Inference Serving)؛ كلفة المهمة المقبولة (Cost per Accepted Task)؛ القيمة الحالية الصافية (NPV)؛ السعودية (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. السؤال العلمي وحدوده

تعالج الورقة وحدة استثمارية في قدرة حوسبة اصطناعية يمكن تأجيرها أو بيع خدماتها، وليس الاقتصاد السعودي ككل. يمكن للوحدة أن تمثل مجموعة عتاد وخدمة استدلال ضمن حد معين من الإنفاق الرأسمالي (Capital Expenditure; Capex)، لكنها لا تمثل مركز بيانات حقيقياً أو طراز معالج محدداً. أهميته منهجية: كيف نَمِيز المسار الذي يزيد فيه التحسين التقني هامش الإيراد من مسار يزيد فيه الإنتاج لكن يضعف التعويض عن رأس المال؟

صياغة السؤال الرئيس: إذا كان السعر المحقق لكل وحدة عمل يتناقص، فكيف تتفاعل مرونة الطلب والسعة الفعلية والتكلفة المتغيرة وتوقيت التدفقات كي تحدد القيمة الحالية لاستثمار أولي ثابت؟ وينبثق منه خمسة أسئلة:

1. هل يكفي أن ينخفض ثمن الرمز ليستدل المستثمر على انخفاض تكلفة تقديم المهمة؟
2. في أي نقطة يؤدي مزيد من الطلب إلى وحدات مفوترة، وفي أي نقطة تحجبه حدود السعة أو الجودة أو زمن الاستجابة؟
3. هل يستطيع تحسين البرمجيات جعل جهاز قائم أكثر ربحية حتى في ظل سعر متراجع؟
4. ما مدى حساسية نتيجة الاستثمار للتكلفة المتغيرة ومعدل الخصم، لا للسعة الاسمية وحدها؟
5. أي بيانات سعودية عامة ومصرح بها قد تختبر هذه العلاقات دون كشف حمولة تشغيلية أو أسرار متعاقدين؟

وحدة التحليل الأصلية المقترحة للتطبيق المستقبلي هي زوج: فئة حمل عمل × دفعة عتاد زمنية (Workload Class × Hardware Vintage). وحين يُدمج هذا الزوج في إحصاء مجمل ينبغي التصريح بتوزيع الأحمال بدل اختزالها إلى متوسط منفصل عن نوع الخدمة.

2. فصل خمسة أعمار غالباً ما تُخلط

نفصل العمر الفيزيائي (Physical Life) الذي يظل فيه العتاد قابلاً للعمل، والعمر المحاسبي (Accounting Useful Life) الذي ينتج عنه الإهلاك، والعمر التعاقدي (Contractual Coverage) الذي تلتزم فيه جهات بدفع مقابل قدرة، والعمر التقني الملائم (Workload-Adjusted Technical Life) الذي تستطيع فيه المنصة تنفيذ أحمال محددة ضمن مستوى جودة وزمن، والعمر التجاري (Commercially Advantageous Life) الذي تستطيع فيه المنافسة مع البدائل بأسعار تحقق هامشاً مناسباً.

هذا الفصل ليس تقسيماً رسمياً معتمداً من هيئة سعودية ولا تقديراً عاماً لسوق الحوسبة. هو تعريف تحليلي مقترح يتيح به المقارنة بين أصول لها أعمار متغايرة. التقرير السنوي لـ «كور ويف» (CoreWeave) يتيح مثلاً على ضرورة الفصل: غيّرت الشركة العمر المقدّر لمعدات الحوسبة من خمس إلى ست سنوات اعتباراً من يناير 2023، وعزت ذلك

إلى تطور الأجهزة والبرمجيات وتصميم مراكز البيانات؛ وأفصحت عن عقود ملتزم بها بلغ متوسط مدتها المرجح نحو خمس سنوات بنهاية 2025 S01. لا يصح حساب فرق «سنة ربح» بطرح أحد الرقمين من الآخر؛ فالعقد ليس مساويًا لدفعة أصل واحدة، والمدة المرجحة ليست تاريخ انتهاء جميع الالتزامات.

في 2026 قدّمت «إنفيديا» (NVIDIA) منصة «روبين» (Rubin) بتصريحات عن تكلفة استدلال أقل بما يصل إلى عشرة أمثال مقارنة بـ«بلاكويل» (Blackwell) ضمن تصميم أحمال مرجعية؛ وفي سبتمبر أعلنت «أوبن إيه آي» (OpenAI) أسعارًا لمنتجات وأجهزات برمجة تطبيقات (API) أحدث أقل من منتجات سابقة في فئة المقارنة S03S04. الأول تكلفة مرجعية ترويجية من صانع العتاد؛ الثاني أسعار بيع منشورة من مزود خدمة؛ وثالث محتمل لم نرصده هو السعر المحقق فعليًا بعد عقود وخصومات واستخدام. القفز من أحدها إلى الآخر خطأ في تعريف المتغير، لا خلاف في الرأي.

3. الأدبيات التقنية: لماذا تنشأ القيمة بين الذاكرة والطاوير؟

تعتمد الإنتاجية المحققة على أكثر من الذروة الحسابية، وعلى تقنيات يمكن أن ترتفع فائدتها دون استبدال كل مكونات العتاد:

إدارة الذاكرة: خوارزمية «يُجد أنشن» (PagedAttention) ونظام «في إل إل إم» (vLLM)، في ورقة «كوون» وزملائه (Kwon et al., 2023)، تستخدم مبدأ تقسيم وإعادة مشاركة ذاكرة المفاتيح والقيم (Key-Value Cache; KV Cache) حتى يُخفف التشتت الداخلي والحجز غير المستخدم. أبلغت تجارب الورقة عن تحسن معدل خدمة (Throughput) بنحو مرتين إلى أربع مرات إزاء خطوط أساس محددة وبمستوى تأخير مماثل S05. قابلية إعادة إنتاج تلك النسبة في السعودية تتطلب مواصفات المنصة والأحمال والإصدارات، ولا تدعم الورقة تعميمًا تجاريًا مباشرًا.

الجدولة: اقترح «أوركا» (Orca, 2022) مستوى جدولة كل دورة توليد (Iteration-Level Scheduling) للحد من تعطيل طلبات قصيرة وراء طلبات طويلة وتطبيق الدفعات الانتقائية (Selective Batching) S06. وهو يذكر بأن زمن الانتظار قد يكون نتيجة قرار برمجي بقدر ما هو قدرة معالج.

تجزئة مراحل الاستدلال: تفرّق التوثيقات التقنية لمنصة «تينسور آر تي إل إل إم» (TensorRT-LLM) بين ملء السياق (Prefill) والتوليد (Decode)، وتشرح لماذا قد تناسب كل مرحلة تخصيصًا مختلفًا للموارد والتوازي؛ وتذكر تكاليف نقل مخبأ الاستدلال عند الفصل، فالحل ليس مجانيًا ولا أفضل على كل طلب S07.

نقطة الاختبار: أعلنت «إم إل كومونز» (MLCommons) في سبتمبر 2026 مجموعة «إم إل بيرف للاستدلال 6.1» (MLPerf Inference v6.1) بما فيها اختبارات مهام متعددة الخطوات S11. يمكن أن تساعد النتائج في بناء خطوط قياس موحدة للأداء، لكنها ليست قيمة حالية صافية ولا مقارنة تجارية بعد الكهرباء والتمويل وشروط البيع.

تستنتج الورقة آلية ممكنة لا إثباتاً سببياً: عندما تُحسَّن مراحل الجدولة وإدارة الذاكرة والحساب، يزداد عدد الخدمات التي يمكن إخراجها من أصل قائم. أما مقدار ما يُفَوَّتر من هذه الزيادة فيعتمد على الطلب، وسلوك العميل، وجودة النتائج، ووضع المنافسة. قد يذهب التحسن إلى مشغّل الحوسبة، أو مطور التطبيق، أو المستخدم النهائي، أو يتوزع بينهم.

4. الفرضيات المتنافسة

- **H1، أثر السعة:** عندما تكون السعة هي القيد ويظل الطلب المدفوع قادراً على امتصاص إنتاج إضافي، ترفع زيادة القدرة القابلة للتسليم صافي القيمة، بشرط ألا تغطي تكاليف التوسع والتشغيل.
 - **H2، أثر انكماش السعر:** مع طلب قليل المرونة، قد ينخفض الإيراد سريعاً رغم ارتفاع عدد الوحدات، ويصبح الاسترداد أكثر صعوبة في أفق تمويلي محدد.
 - **H3، الارتداد الحاسم:** عندما يكون الطلب شديد المرونة وتتاح سعة لاستيعابه، قد يزيد الإيراد الإجمالي بعد هبوط السعر. وهذا ليس حكماً بأن أرباح النشاط ستزيد حتماً.
 - **H4، عقد مقابل سوق فوري:** عند وجود التزامات شراء سعة متعددة السنوات، يختلف تعرض الإيراد الآجل لسعر الخدمة الفوري؛ لكن حقوق التسليم والتخلف والخصومات والتجديد تجعل الحماية مشروطة.
 - **الفرضية المنافسة R1:** قد تظهر وفرة الحوسبة انخفاضاً في وحدة الأسعار وتزامناً مع نمو ربحي كبير لأن وظائف جديدة عالية القيمة تبتلع القدرة؛ في هذه الحالة لا يظهر الفخ من بيانات الربح المحقق.
 - **الفرضية الصفرية H0:** بعد ضبط خليط أحمال العمل والكهرباء والتمويل والعقود والجودة، لا تكون لسرعة تناقص سعر الوحدة قدرة تفسيرية إضافية لعائد الاستثمار. ينبغي ألا تُسقط هذه الفرضية دون بيانات تشغيلية فعلية وتصميم استدلال سببي مناسب.
- هذه فرضيات وقضايا تأويلية؛ اختباراتنا الحسابية تدرس اتساقها داخل نموذج مكوّن من افتراضات، ولا تختبرها على مجموعة بشرية أو منشآت فعلية.

5. تصميم النموذج المنقذ ووحداته

المدخلات والترميز: لكل سنة $t = 1, \dots, 6$ ، يرمز p_0 إلى مؤشر سعر أولي مقداره 1، و q_0 إلى طلب سنوي مقداره 50 وحدة، و k_0 إلى قدرة سنوية قابلة للتسليم تساوي 50، و c_0 إلى تكلفة متغيرة أولية للوحدة 0.20، و F_0 إلى تكلفة تشغيل ثابتة سنوية 7، و I_0 إلى استثمار أولي قدره 120، و s إلى قيمة متبقية مفترضة في نهاية السنة السادسة قدرها 8، و r إلى معدل خصم 12%. جميعها وحدات مؤشر افتراضية؛ لا تعني دولارات أو ريبالات أو ميغاواط أو أجهزة أو رموزاً حقيقية.

المتغيرات المختبرة هي δ ، معدل هبوط سعر الوحدة سنوياً؛ و ε ، القيمة المطلقة لمرونة الطلب السعرية؛ و g ، معدل تحسن السعة القابلة للبيع بفعل تحسين تشغيلي يُفترض مستمراً. طُبِّقت الصيغ الآتية برمجيّاً:

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

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

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

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

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

D_t طلب محتمل، و K_t سقف الخدمة المؤهلة، و X_t وحدات قابلة للفوترة، و CF_t تدفق نقدي تشغيلي مع قيمة متبقية في آخر سنة. افترضنا أن المعاملة أو الطلب الواحد يساوي وحدة، وأن كل وحدة مسلفة تستوفي الجودة والخدمة؛ وهذه فرضية **مريحة وغير واقعية بالضرورة** سنحتاج إلى فصلها تجريبيًا. استُخدمت دالة حد أدنى لمنع نمذجة إيراد مقابل طلب لم يجد سعة. ليس ضمن النموذج دين أو فوائد أو ضريبة أو تغيير في الأجور أو إهلاك محاسبي أو إفلاس طرف أو توقف عتاد؛ وتكلفة رأس المال ممثلة بمعدل خصم مفترض وليس سعر قرض مُستنتجًا من إفصاح أمريكي.

يستلزم الافتراض $c_t = c_0 / (1 + g)^{t-1}$ حذرًا بالغًا: افترضنا أن زيادة إنتاجية البرمجيات تُنقص التكلفة المتغيرة للوحدة بالمعدل نفسه، وهو قد يفشل إذا كان معظم التحسن متعلقًا بتشغيل أحمال أكثر دون خفض فعلي لمصروف الطاقة أو الشبكة. لذلك نفذنا مسار حساسية يثبت c_t عند c_0 مهما تغيرت السعة. تظهر نتيجة التحفظ في القسم الثامن.

6. البيانات والتنفيذ وقابلية الإعادة

لا تتضمن الدراسة سجل معاملات مصرفية أو أرقام إشغال من مشغل سعودي. تنشأ جميع القيم التجريبية من ملف `data/model_assumptions.json` القابل للقراءة؛ وينفذ `analysis/reproduce.py` الحسابات بلغة «بايثون» (Python) باستخدام مكتبتها القياسية فقط. ينتج التنفيذ خمسة سجلات فعلية:

1. `base_annual_cashflows.csv`: ستة صفوف للتدفقات السنوية والحجم والسعة والإيراد والخصم.

2. `synthetic_scenario_results.csv`: شبكة 27 حالة مستقلة وفق $3 \times 3 \times 3$.

3. `sensitivity_checks.csv`: 18 حالة إضافية تتقاطع فيها ثلاثة معدلات خصم (8%، 12%، 16%) وثلاثة معدلات تحسن سعة (0%، 25%، 50%) وحالتا تحسن التكلفة المتغيرة وثباتها.

4. `derived_checks.json`: ملخص مُستخرج للعبة والنتيجة الأساسية وعدّ الحالات.

5. ملف رسوم ورسوم بيانية بصيغتي «إس في جي» (SVG) و«بي إن جي» (PNG) بُني من الجداول الناتجة، لا من رسومات تقديرية.

توجد نسخة جدولية قابلة للفحص باسم evidence_workbook.xlsx تفصل المدخلات القابلة للتعديل عن الحسابات وصور النتائج. وهي أداة معاينة، ولا تحل محل السكرتير بوصفه السجل الحسابي المنشور للتجربة الكاملة. استخدمنا توقع سنوي يدفع في نهايته، واستثماراً في الزمن صفر، وقيمة متبقية في السنة السادسة فقط. لا مصادر مالية خارجية تغذي الأرقام الاصطناعية؛ المصادر الخارجية تعطي سبب اختبار الآليات وقيود تحويلها إلى الواقع.

7. نتائج التجربة الفعلية: لماذا لا يفوز الطلب دائماً؟

نُفذت شبكة أسعار سنوية متناقصة بمقادير 10% أو 22% أو 35%؛ ومروناً طلب ϵ مقدارها 0.5 أو 1.0 أو 1.8؛ ومعدلات زيادة سعة مقدارها 0% أو 25% أو 50%. كل تركيب يعطي ست سنوات ثم نتيجة القيمة الحالية. لم تُسحب قيم عشوائياً ولم يُقدَّر توزيع احتمالي، لذلك فإن نسبة الحالات الربحية داخل الشبكة ليست معدل نجاح متوقعاً في الواقع. إنها تعداد لأجزاء من فضاء قيم اخترناه نحن.

الحالة الأساسية: هبوط سعر 22% في السنة، ومرونة 0.7، وتحسن سعة وتكلفة للوحدة 25% سنوياً، مع خصم 12%. نتائج السنوات كما حُسبت، لا كما توقَّعها محرر النص:

السنة	مؤشر السعر	الطلب المحتمل	القدرة المتاحة	الوحدات المباعة	الإيراد الافتراضي	التدفق السنوي مع القيمة المتبقية إن وجدت
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

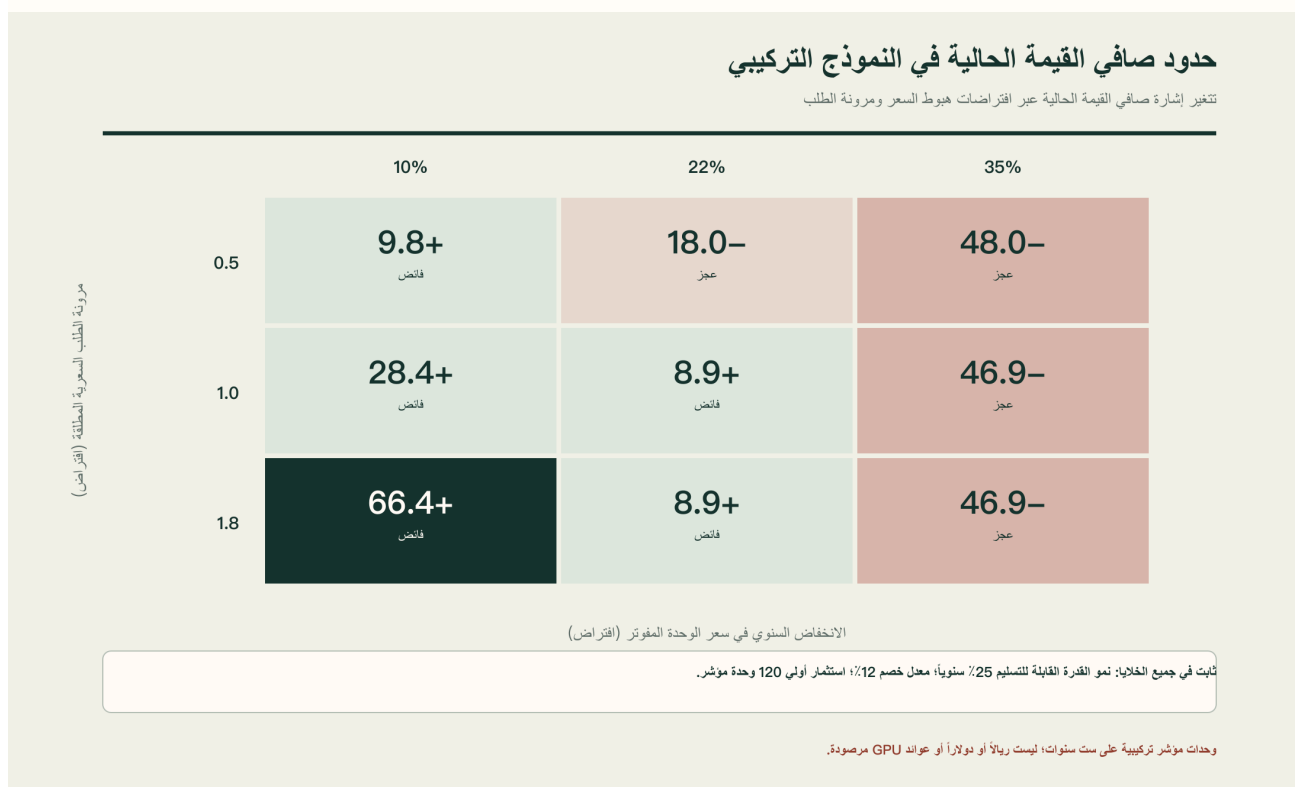
السنة السادسة تضم قيمة متبقية مفترضة مقدارها ثمانى وحدات. جميع الأعمدة أرقام مؤشر صناعية؛ مقربة إلى منزلتين أو ثلاث. الحساب الدقيق موجود في CSV.

نلاحظ أن الطلب يرتفع أكثر من الضعف رغم هبوط الإيراد السنوي من 50 إلى نحو 34.44. إجمالي الإيرادات خلال ست سنوات حوالي 251.02 وحدة، لكن الإيراد ليس ربحاً: التكاليف المتغيرة والثابتة والاستثمار الزمني تترك قيمة **حالية صافية تقارب -5.5835**. تبلغ التدفقات المتراكمة بعد طرح التكلفة الأولية قيمة غير سالبة في السنة الخامسة **دون خصم**، بينما تظل القيمة الحالية سالبة لأن العودة المتأخرة للمال أقل وزناً. من المهم ألا نخلط *Payback* *Period* بنتيجة *NPV*.

المقارنة المختارة الآتية تثبت السعر السنوي الهابط بنسبة 22% وتحسن السعة 25%، وتغير المرونة فقط:

مرونة الطلب المفترضة	NPV في ست سنوات بوحدة افتراضية	إيراد السنة السادسة	النتيجة التحليلية
0.5	17.98-	26.87	زيادة الحجم غير كافية لتعويض انكماش السعر.
1.0	8.91+	44.05	طلب أعلى يغطي القدرة المبعة.
1.8	8.91+	44.05	النتيجة نفسها لأن سقف السعة يقيّد التسليم.

المساواة بين صفّي 1.0 و1.8 ليست تقريبًا اعتباطيًا: في كل السنوات اللاحقة للبداية عند هذه الافتراضات، تجاوز الطلب المتاح قدرة الفوترة، فأصبح $x_t = K_t$ ؛ لا يغير رفع الرغبة الشرائية المخرجات الواقعية. لذلك يجب التفريق بين نمو السوق الممكن ونمو الحجم القابل للتسليم ونمو الحجم المدفوع فعليًا. وإذا كانت هناك خدمة غير متاحة عند السعر نفسه، لن تُعوّضها كثافة الحماسة إلا إذا كان بالإمكان التوسع أو رفع السعر ضمن قبول العميل.



شبكة القيمة الحالية الصافية في السيناريوهات الافتراضية

الشكل 1: مصفوفة نتائج تجربة فعلية لمرونة وأسعار متعددة مع ثبات فرضية التحسن السنوي للسعة عند 25%. الأرقام ليست توزيعات احتمالية ولا تقديرات مؤسسات سعودية.

التجربة الكاملة أعطت 12 سيناريو بقيمة موجبة و15 سالبة من أصل 27، مع إشارات مختلطة حسب ثلاثة محاور. فعندما تُثبّت القدرة على 50 وحدة طوال ست سنوات ($g=0$)، تصبح النتائج كلها سالبة في الإعدادات المختارة، لأن الطلب يزيد لكن لا يمكن تسليم أكثر من خمسين والسعر يواصل الهبوط. في الطرف المقابل، بلغت القيمة في سيناريو واحد شديد المرونة والسعة ($\delta=0.22$ و $\varepsilon=1.8$ و $g=0.50$) نحو 113.75+، لكنها نتيجة مشروطة بافتراض مفرط في سرعة نمو إنتاجية البرمجيات عبر سنوات متتابعة، لا حجة للاستثمار الفعلي بهذه الأرقام.

8. نتائج اختبارات الحساسية ومتى تتغير الخلاصة

جرى بالفعل اختبار تغير معدل الخصم بين 8% و12% و16%، وتحسن السعة بين 0% و25% و50%، مع تكلفة متغيرة للوحدة إما تنخفض بالنسبة المفترضة أو تبقى ثابتة. على حالة السعر 22% والمرونة 0.7، أظهرت النتائج التالية عند تحسن سعة 25%:

معدل الخصم	تفترض التكلفة المتغيرة تتحسن مع السعة	تفترض التكلفة المتغيرة ثابتة
8%	7.84+	22.46-
12%	5.58-	31.18-
16%	16.82-	38.62-

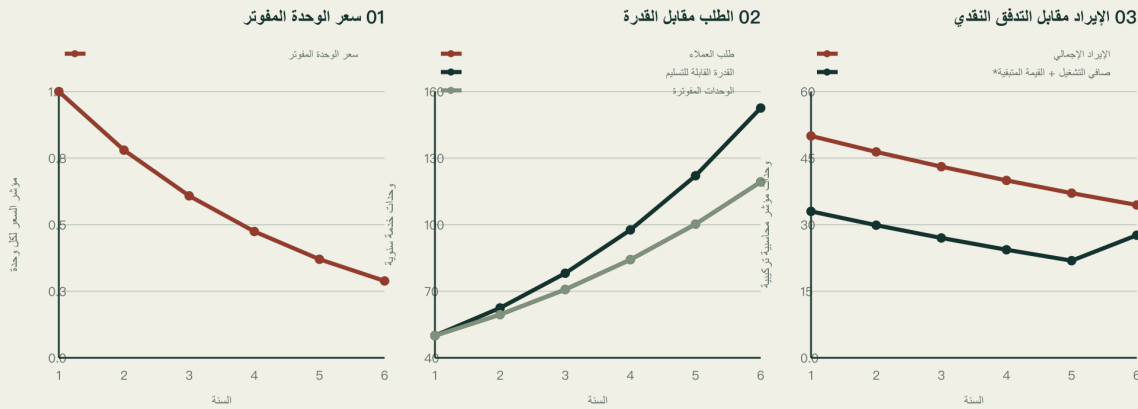
تفسير النتائج أهم من حفظ القيم: إن حذف افتراض انخفاض التكلفة المتغيرة للوحدة يجعل تحسينات السعة وحدها غير كافية لتوليد العائد المفترض، لأن الإيراد يتحدد بما يشتريه الطلب لا بما يستطيع البرنامج تشغيله. عند بعض تركيباتنا، يصبح رفع α من 25% إلى 50% غير مؤثر في الإيراد أصلاً بعد بلوغ العرض المطلوب، ما دام سعر البيع وحجم الطلب لم يتغيرا. ليس هذا رفضاً للبرمجيات؛ إنه سؤال عن أي قيد أزلناه فعلاً: قيد الحوسبة، أم قيد التكاليف، أم قيد وجود زبون؟

استخدمنا بحثاً ثنائياً (Binary Search) تكرارياً على ϵ ضمن الفترة من 0 إلى 2.5، مع تثبيت بقية قيم الحالة الأساسية، لإيجاد أصغر مرونة تجعل $NPV \geq 0$ ، فكانت **0.7803 تقريباً**. هذه ليست مرونة مقدرة للطلب السعودي ولا للسوق العالمية. إنها عتبة داخل نموذج معين؛ تغيير الاستثمار الأولي، أو السعر، أو جودة التسليم، أو خصائص الطلب، أو فترة الخصم، يغيرها. يمكن إعادة حسابها بفحص `data/derived_checks.json` وتشغيل السكريبت، لا بالاستناد إلى اسم اقتصادي معروف.

حدود هذه الاختبارات: لم نفحص توزيعات عشوائية لارتفاع الكهرباء، أو ازدحام الشبكة، أو فقد عقود، أو اختلاف جغرافيا مراكز البيانات. لم نستخدم عينات إحصائية؛ لذلك لا توجد قيم p أو فواصل ثقة أو دعوى دلالة إحصائية. ولن نقول إن اختبار 18 حالة يمثل كل فضاء المخاطر. ما شملته الاختبارات ثلاث مجموعات متقاطعة من معاملات محددة ومعلنة.

فجوة السعر والقدرة

الحالة الأساسية: $\delta=22\%$ ؛ $\varepsilon=0.7$ ؛ نمو القدرة القابلة للتسليم 25% سنوياً



* يتضمن تدفق السنة 6 قيمة متبقية مقترضة قدرها 8 وحدات؛ لا يظهر الاستثمار الأولي 120 وحدة في السنة صفر.

مخرجات نموذج افتراضي: لا توجد قدرة أو أسعار أو تدفقات نقدية لمشروع مرصودة.

مسارات السعر والطلب والسعة والإيراد في التجربة الافتراضية

الشكل 2: نتائج الحالة الأساسية. تحسن القدرة وتقدم السعر وتوقيت التدفق ظواهر مختلفة؛ وحدة المحور المالي مجرد مؤشر تركيبي.

9. اختبار الصلة بالسعودية دون إسقاط بيانات الشركات بعضها على بعض

توجد ثلاث فئات من الوقائع الأصلية ذات صلة تجريبية، وليس بينها سلسلة قياس واحدة:

أ. إفصاح تجاري أمريكي، لا قياساً سعودياً: يظهر تقرير «كور ويف» (CoreWeave) لعام 2025 اختلاف العمر المحاسبي للمعدات وعمر العقود، وتعرض قوائمها لعام 2026 رقم التزامات أداء متبقية (RPO) بلغ **103.7 مليار دولار في 30 يونيو 2026**، يُتوقع الاعتراف بـ 41% منه خلال أول 24 شهراً من ذلك التاريخ، و 39% خلال الأشهر 25-48، والباقي حتى الشهر 78 S01S02. الإفصاح يذكر أيضاً مجموع مدفوعات أصل دين مستقبلية **35.551 مليار دولار** حسب جدول استحقاقات شامل S02. لا يسمح جمع الرقمين أو قسمتهما ببناء نسبة ضمان أو استنتاج أزمة أو فائض؛ الأول خدمات تعاقدية لم تُقدّم بعد وفق قواعد الاعتراف بالإيراد، والثاني استحقاقات تمويلية متعددة الأوقات، وليس تدفقين صافيين متقابلين. قد تحمي العقود طويلة الأجل المشغل من السعر الفوري؛ لكن لم نحل شروطها الفردية أو سجلات تنفيذها.

ب. إعلان تشغيل سعودي من أطراف تجارية: أعلنت «إيه إم دي» (AMD) و«سيسكو» (Cisco) و«هيومين» (HUMAIN) يوم 31 أغسطس 2026 أن أنظمة معالجات «إم آي 355 إكس» (AMD Instinct MI355X) وشبكات (Cisco Silicon One) في السعودية دخلت الإنتاج لخدمة عملاء «هيومين» (HUMAIN) داخل المملكة

وخارجها S09. وقد أفصح الإعلان ذاته عن نية نشر سعة تصل إلى 250 ميغاواط بدءًا من 2027، وعن هدف بلوغ غيغاواط واحد بحلول 2030؛ هاتان قدرتان مخطّطتان، وليستا نتيجة تشغيل منشورة لعام 2026. وبما أن الإعلان صادر عن الأطراف ذات المصلحة، نُسب لها بوضوح، ولم يُعدّ تقرير تدقيق مستقل في الطلب أو التكلفة أو الهامش.

ج. خدمة سعودية موجهة للمستخدم: أعلنت «وكالة الأنباء السعودية» (Saudi Press Agency; SPA) يوم 25 أغسطس 2025 طرح «هيومين شات» (HUMAIN Chat) المعتمد على «عَلَام 34 مليار» (ALLaM 34B)، مع واجهة عربية وأدوات صوتية S10. هذا يثبت إطلاق منتج حسب الجهة المعلنة، ويفتح سؤال كيف يتحول الاستثمار في القدرة إلى طلب مفيد، لكنه لا يسمح باستنتاج ربح أو إيراد أو عدد رموز أو إحصاء استخدام. يجوز استخدامه حالة سياقية للاستخدام المجتمعي؛ ولا يجوز إدخاله في نموذجنا بوصفه ملاحظة عددية لم تُنشر.

السؤال السعودي إذن ليس أي قصة نفّذتها دولة أخرى حرفيًا، وإنما ما إذا كانت مرافق الحوسبة قادرة على بيع نتيجة تطبيقية في أسواق محلية وعابرة للحدود بموجب مستويات خدمة وتكاليف متحولة. وقد تختلف اقتصاديات تدريب نموذج تأسيس (Foundation Model Training) عن اقتصاديات تشغيل خدمة استدلال (Inference Serving)، لذلك يُمنع استعمال أسعار الرمز وحدها لتقييم أصول مخصصة للتدريب على نطاق واسع.

10. التمييز بين «السعر» و«التكلفة» و«القيمة»: سجل قياس مقترح

يحتاج تحويل هذه الورقة إلى بحث إمبريقي إلى تصحيح ثلاث وحدات حسابية تتداخل في النقاش العام:

1. **تكلفة البائع (Provider Cost):** كلفة التشغيل الفعلية للوحدة، بما في ذلك الطاقة، والشبكة، والمكونات البرمجية، والدعم، والقدرة غير المستخدمة، وأثر الإهلاك والتمويل عند مستوى مناسب من التخصيص. ليست مماثلة للتكلفة التشغيلية الهامشية فقط.
2. **السعر المحقق (Realized Selling Price):** ما يتحصل عليه البائع فعليًا بعد العقود والخصومات والتخزين المؤقت والفوترة المختلفة. ليس سعر القائمة (List Price) المنشور على الويب بالضرورة.
3. **قيمة العميل (Value to Customer):** منفعة المهمة المنجزة، ويمكن أن تقدر بزمّن توفير أو إنتاجية أو جودة أو إيراد مرتبط، لكنها تحتاج قياسًا مستقلًا ولا تساوي السعر آليًا.

ولأن الطلبات غير متجانسة، يُقترح تقرير «التكلفة لكل مهمة مقبولة» (Cost per Accepted Task) إلى جانب تكلفة مليون رمز، على أساس أن المهمة المقبولة تحدّد مسبقًا بضوابط دقة وسلامة ومدة. ويراعى على وجه الخصوص نسبة مهام أعيد تنفيذها، وتفاوت طول السياقات، ونسبة (Cache Hit Rate) للمدخلات، وعدد الرموز الخارجة، والتوقف المبكر، وعدم توفر الموارد، واستهلاك الكهرباء المحسوب على مستوى العبء بمعدل موثق.

يمكن تمثيلها وصفيًا عند نوع حمل w وفترة t :

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

حيث E تكلفة الطاقة والتبريد المقدرة، و N تكلفة الشبكة والبيانات، و O تكلفة التشغيل والدعم المنسوبة، و A نصيب استخدام رأس المال؛ و T عدد المهام التي اجتازت معيار القبول خلال الفترة. الصيغة **مقترحة وغير معتمدة**، وتتطلب سياسة تخصيص تكاليف وإثبات عدم مضاعفة المصروفات؛ وعندما لا توجد مهمة مقبولة في المقام فلا يصلح عرض تكلفة وهمية «صفر» أو القسمة على صفر. ويجب إضافة النفقات غير المخصصة والإيرادات المؤجلة إلى تحليل مشروع كامل بدل تقديم المؤشر الجزئي بوصفه الربح الصافي.

ثمة معياران إضافيان: هامش المهمة المقبولة في حدود زمن متفق عليه (SLA-Compliant Contribution Margin)، وحصة الخدمات التي تولد تدفقًا بعقد ملتزم به مقابل إيرادات يتبع سعر السوق الفوري (Spot-Market Revenue). وفي كل مؤشر يصرَّح بالفترة الزمنية وتعريف الخدمة والأسعار الفعلية والعملية ومنهج تخصيص الإنفاق.

11. تصميم دراسة سعودية تالية قابلة لتكذيب الفرضية

التصميم المقترح، غير المنفذ: رصد شهري على مستوى فئات عمل مؤسسية مجمعة ودفعات عتاد متعاقبة لمدة 24-36 شهرًا إن أُتيحت البيانات، بتعاون طوعي موثَّق من مشغلين ومطورين يمثلون أنماط استخدام متعددة. يكون التحليل المسبق الأول توصيفيًا، يلي ذلك تصميم سببي فقط إذا توفرت لحظة تغيير قابلة للعزل بوضوح، مثل ترقية برنامج استدلال على جزء من مجموعة أحمال، مع مجموعة مقارنة لم تتغير بالتزامن. لا نقدّم الترقية اعتباريًا كـ«تجربة عشوائية» ما لم يحدث تخصيص عشوائي حقيقي.

الحد الأدنى لجدول البيانات المستقبلي: معرّف دفعة عتاد مجهول؛ فئة نموذج وإصداره؛ نطاقات مدخلات ومخرجات ومعدل التخزين المؤقت؛ نسب المهام المقبولة؛ زمن الرمز الأول وزمن إنجاز الطلب ضمن مستوى خدمة؛ وحدات عمل صالحة للفوترة؛ متوسط سعر محقق وتوزيعه؛ كهرباء وتبريد منسوبان بمنهج مفصّل عنه؛ حصة سعر تعاقدية؛ جدول الإهلاك ومعدل خصم اقتصادي مختار؛ تواريخ بدء وانتهاء الحمل. لا يتضمن الجدول عناوين منشآت حرجة، أو مواقع غرف معدات، أو سجلات مستخدمين، أو عقود أفراد، أو مفاتيح وصول. ويؤخر أو يجمع ما قد يكشف أنماط استخدام تشغيلية حساسة.

خطوات المقارنة:

1. تعريف ثلاثة إلى خمسة أصناف حمل قبل مشاهدة النتائج: استدلال قصير متفاعل، سياقات طويلة، دفعات معالجة غير فورية، تدريب تكميلي، ومهام متعددة المراحل إن كان جمعها مصرحًا. هذه مجرد أمثلة قابلة للتعديل حسب البيانات المتاحة.
2. اختبار جودة معياري خاص بكل صنف، بحيث لا تحسن الإنتاجية على حساب صحة الناتج دون أن يظهر ذلك في المقام.
3. تقسيم العينة زمنيًا؛ استخدم بيانات فترة سابقة لمعايرة النموذج، وفترة لاحقة خارج العينة (Out-of-Sample) للتقييم، وامنع تسرب الأسعار المستقبلية أو أجيال أجهزة لم تكن متاحة وقت التنبؤ.

4. تقدير منحني سعر محقق وحجم مدفوع مع متغيرات توضح تغير الجودة وتوفر القدرة والعقود، بدل تقدير المرونة من بيانات سعر قائمة واحد فقط.

5. اختبار أثر الترقيات البرمجية بالمقارنة ضمن دفعة العتاد وحمل ثابت حيث يمكن، مع رصد تغير مزيج العملاء والتوقيت ومصرفات الطاقة؛ وإذا لم توجد مجموعة مقارنة معقولة، تُعرض النتيجة ارتباطية فقط.

6. إدخال استحقاقات رأس المال وفروق شروط العقود في نموذج مالي توثق افتراضاته وتقارن نتائجه بحجم المدفوعات الفعلية حتى لا يُستنتج الربح من مقياس تقني منفرد.

المقدر السببي المقترح إن سمحت البيانات: الفرق في تغير تكلفة المهمة المقبولة بين أحمال حصلت على تحسين برمجي وأحمال مماثلة لم تحصل عليه في الفترة نفسها، وفق فروض الاتجاهات السابقة المتوازية (Parallel Trends) وعدم حدوث تغييرات أخرى تفاضلية. لكن مراكز الحوسبة مشتركة الموارد، ما يعني احتمال انتشار أثر الترقية إلى الأحمال الأخرى عبر الطوابير أو ذاكرة مشتركة؛ لذلك قد لا يصح افتراض انعدام التداخل بين الوحدات (No Interference). يمكن بدلاً من ذلك اختيار مقارنات على مستوى تجمع موارد كامل مع تحديد انتشار الأثر أو الاكتفاء بوصف الآلية.

فحص مضاد مهم: قد يصاحب ترقية برنامج استدلال دخول عملاء مختلفين، أو إصدار نموذج أصغر، أو عقد جديد بسعر أقل. إن لم نثبت هذه التغيرات فقد ننسب إلى البرنامج أثراً سببياً ناتجاً عن تبدل المزيج أو القوة التفاوضية. وتحتاج دراسة مرونة السعر إلى معالجة الطلب غير الملبى: من يقيس الحجوزات المكتملة فقط قد يظن الطلب ضعيفاً، بينما تُرفض طلبات أخرى عند امتلاء السعة.

12. أين يمكن أن يفشل «فخ انكماش سعر الرمز»؟

ينبغي ألا يتحول اسم جذاب إلى افتراض صحة. توجد ستة حدود تضعف الحجة أو تعكس نتيجتها:

أولاً: تغاير المعالج عن المنتج النهائي. جزء من سعر الخدمة ثمن البرمجيات والأمان والدعم وحقوق الوصول وواجهة المستخدم؛ وقد تزيد قيمة هذه العناصر مع انخفاض تكلفة المعالج. انخفاض سعر الرمز داخل الشركة لا يستلزم انخفاض إجمالي إيراداتها.

ثانياً: قابلية تحويل العمل. قد تعود معالجات لم تعد تناسب تدريب نموذج حدودي (Frontier Model) لتنفيذ نماذج أصغر أو مهام استدلال أرخص. قيمة الجهاز بعد تخلفه عن القمة لا تساوي صفرًا بالضرورة. لا يتيح النموذج نقل الأعباء إلى مستخدمين جدد بتسعير مختلف، وهي فرضية مقيدة عن قصد.

ثالثاً: تطور نوع المهمة. قد يتطلب تطبيق من جيل جديد استدلالاً متعدد الأدوات أو ذاكرة سياق طويلة، فترتفع تكلفة المهمة رغم انخفاض تكلفة الرمز. إذا تغير عدد الرموز المطلوبة للنتيجة، يتوقف ثمن المليون رمز عن تمثيل جودة الخدمة الاقتصادية.

رابعًا: العقود. حصة مرتفعة من الإيراد التعاقدية قد تعزل فترة السداد عن السعر الفوري، بينما يظهر أثر السعر لاحقًا عند تجديد العملاء. يخلق ذلك فرقًا زمنيًا يجب نمذجته؛ فرضنا في السيناريو الأساسي سعرًا متغيرًا لكل الوحدات لعرض الآلية دون هذا العازل.

خامسًا: الموارد غير الحاسوبية. قد يقيد نشر الحوسبة ترخيص كهرباء أو تأخر محطة محولات أو ماء تبريد أو معدات شبكة، فلا تتحول قدرة البرمجيات النظرية إلى خدمة. في الاتجاه المعاكس، قد تخفف خوارزميات الذاكرة والتوازي الحاجة إلى شراء أجهزة جديدة. لم نقدّر أيًا من هذه القيود فعليًا.

سادسًا: وفرة الطلب والقيود البديلة. إذا ولدت الأسعار الأرخص أعمالًا نافعة جديدة وكان السوق مرتنًا، يمكن أن تتوسع الإيرادات والهوامش. لكن إن بقيت القدرة محدودة في الأجل القصير، قد تتغير الأسعار والعقود بدل الكميات. نموذجنا يعزل قيدًا واحدًا، ولا يقدر توازنًا عامًا بين شركات متنافسة.

ولفصل هذه العوامل، لا يكفي أن يجيب محلل مرة واحدة عن سؤال «كم تتراجع الأسعار سنويًا؟». يجب أن يراقب الإصدارات والمهام والمدد التعاقدية والطلب المكتوم والتكلفة المتغيرة، وأن يختبر نماذج بديلة، ومنها نموذج طلب لا يتبع مرونة ثابتة، ونموذج أسعار غير تناقصية على بعض الأحمال.

13. الاختبار المميز المقترح قبل أي استثمار جديد

يقترح البحث تجربة قرار استثمارية غير هجومية، لا تحتاج موقع منشأة أو تفصيل تدفق حساس. تختار جهة مشغلة ومخولة فئتين من أعباء العمل، واحدة يغلب فيها الوقت الحرج للاستجابة، والثانية يغلب فيها حجم الدفعات. ثم تجري ثلاث مقارنات على الموارد المصرح بها: إصدار برنامج قديم وحديث على العتاد نفسه؛ جهاز قديم وحديث على البرنامج نفسه إن أمكن؛ وتحويل جزء من الأحمال بينهما ضمن القيود المسموحة. يُضبط نموذج الذكاء ونطاق الجودة ووقت الاستجابة وبيانات التسعير المحقق. وثُقاس فترات طويلة بما يكفي لملاحظة تغير الطلب وليس اختبار الذروة القصيرة فقط.

إذا أدت ترقية البرمجيات إلى انخفاض تكلفة المهمة المقبولة وارتفاع الهامش الإجمالي لدى المشغل من دون سعر بيع أدنى، تدعم هذه الملاحظة مسار الإنقاذ البرمجي. وإذا انخفض سعر البيع أسرع من التكلفة الكلية، وأدى ذلك إلى ضعف الاسترداد رغم نمو عدد الطلبات، تدعم مسار الفخ. وإذا زاد حجم المهام المدفوعة بقوة وارتفع الهامش، تبطل على تلك الفئة فرضية أن الانكماش السعري ضرورة للخسارة. وإذا لم يتغير الهامش بعد الضبط، فهذا تقييد لحجم الادعاء، وليس فشلًا شخصيًا للباحث.

قاعدة قرار يمكن اختبارها بدل الادعاء بأنها معيار قائم: قبل تمويل دفعة حوسبة جديدة، قارن قيمتها الحالية الشرطية عبر ثلاثة بدائل: (أ) شراء الجهاز الجديد؛ (ب) تحسين برنامج الخدمة الحالي؛ (ج) شراء سعة خارجية مؤقتة لقياس الطلب الحقيقي. يتضمن كل بديل كلفة التمويل والمخاطر واستهلاك الطاقة وفترة التطوير وإيراد الخدمات المقبولة. قد يفوز أي بديل بحسب الأحمال والعقود. هذا إطار مقارنة عملي داخل نطاق مقال M12، وليس «حكمة» عامة أو توصية باستثمار معين.

14. ما يجوز قوله للقيادة وما لا يجوز

يمكن قول: «تراجع تكلفة الاستدلال التقنية وتحسن البرمجيات يغيران اقتصاد الأصول طويلة العمر، ويستحقان اختبار مرونة الطلب المدفوع والزمن اللازم لاسترداد رأس المال». ويمكن أيضًا قول: «ثمة تشغيل معلن لعتاد متنوع في السعودية عام 2026 وأهداف توسع لاحقة، لكن الهوامش التفصيلية غير منشورة في المواد التي راجعناها».

ولا يجوز القول: «كل وحدة معالجة ستصبح عديمة القيمة في أقل من سنة»، أو «العقود تجعل الخسارة مستحيلة»، أو «12 من 27 تجربة تعني احتمال ربح 44%»، أو «قيمة NPV التركيبية تقيس مشاريع سعودية». كذلك لا يجوز القول إن إيراداتًا تعاقديةً مستقبليًا أمريكيًا مثبت القيمة يعادل سيولة حالية، أو إن الميغاواط المخطط بحلول 2030 يعادل قدرة مستخدمة أو ربحًا مكتسبًا عام 2026.

الحكم المشروط الحالي: إن تناقص السعر ليس إخفاقًا تقنيًا، وقد يكون نجاحًا اجتماعيًا واقتصاديًا ضخمًا للمستخدمين. يصبح مخاطرة رأسمالية في الظروف التي يفشل فيها الطلب المدفوع والتكلفة الفعلية والجدول الزمني في مجارة الأسعار. من جهة أخرى، يمكن للبرمجيات الجيدة والعقود المحكمة وإعادة تخصيص الأعباء أن ترفع جدوى أصول موجودة. لا تميز البيانات العامة المتاحة بين هذه الحالات في سوق سعودي ممثل، لذلك تظل توصية البحث القياس الموجه لقرار محدد لا ترجيح نصر أو خسارة.

15. بيان القابلية لإعادة الإنتاج وأخلاقيات البيانات

تتكرر التجربة الحسابية محليًا بأمر واحد من دليل المقال:

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

تحتاج العملية «بايثون» (Python) القياسية فقط؛ ملف `analysis/requirements.txt` يوثق ذلك. يتضمن `data/base_annual_cashflows.csv` ويتولد `data/model_assumptions.json` جميع القيم المختارة، و `data/sensitivity_checks.csv` و `data/synthetic_scenario_results.csv` و `data/derived_checks.json`. أي تغيير في الافتراضات يولد نتائج أخرى، وعلى من ينشرها ذكر نسخة النموذج وقيمته الجديدة. يتطلب إنشاء الرسوم مكتبة عرض إضافية، لكنه لا يلزم للتحقق من المعادلات الرقمية. ملف `evidence_workbook.xlsx` يوفر تدقيقًا جدوليًا مستقلًا للحالة الأساسية إذا توفرت بيئة تشغيل الجداول ودعم إعادة حساب الصيغ.

لم يتم جمع بيانات شخصية أو معلومات تشغيلية حساسة أو خرائط اعتماديات خاصة. لا توجد موافقات مؤسسية مزعومة، ولا ادعاء تحكيم أو إثبات تجريبي سعودي أو اطلاع على عقود غير منشورة. البيانات الحقيقية المستخدمة جميعها إفصاحات أو أوراق أو إعلانات عامة، مع الإشارة إلى أن الإعلان التجاري يعبر عن أطرافه لا عن تدقيق ثالث مستقل. البيانات التركيبية منشأة خصيصًا لشرح العلاقات السببية الممكنة داخل المعادلات.

16. المصادر الأصلية والمرجعية

- S01.** CoreWeave, Inc. *2025 Annual Report on Form 10-K*. U.S. Securities and Exchange Commission. Accounting Policies: Property and Equipment. Contracts. Take-or-Pay. <https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/crwv-20251231.htm>. نسخة البيانات المالية 31 ديسمبر 2025؛ أُتيح التقرير في 2026.
- S02.** CoreWeave, Inc. *Quarterly Report on Form 10-Q, period ended June 30, 2026*. Debt maturity table. SEC. Note 2, Remaining Performance Obligations. <https://www.sec.gov/Archives/edgar/data/1769628/000176962826000366/crwv-20260630.htm>. تاريخ الرصد 30 يونيو 2026.
- S03.** NVIDIA. *NVIDIA Kicks Off the Next Generation of AI With Rubin*. 5 January 2026. مصدر بيان صناعي عن تصميم الجيل وتكلفة رمز استدلال تقديرية، وليس مقارنة مالية مدققة. <https://nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer>.
- S04.** OpenAI. *Introducing GPT-6 Sol and Luna*. September 2026; page updated 29 September 2026. أقسام API pricing, caching؛ مقارنة أسعار منتجات من جيلين مختلفين. <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>. نتائج اختبارية مشروطة بأجهزة وأحمال الورقة.
- 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>. معالجة طابور وجدولة الدورات.
- S07.** NVIDIA TensorRT-LLM. *Disaggregated Serving*. Technical documentation, current living version at cutoff. <https://nvidia.github.io/TensorRT-LLM/features/disagg-serving.html>. وثائق مهندسي المنصة؛ لا تجعل اختيار الفصل أفضل دائماً.
- S08.** William Stanley Jevons. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines*. 1865. نسخة تاريخية: <https://oll.libertyfund.org/titles/jevons-the-coal-question>. استدعاء لتاريخ فرضية الارتداد، لا

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>. المصدر الأصلي نفسه ذو

مصلحة؛ فصلنا بين التشغيل المعلن والسعة المخططة.

S10. Saudi Press Agency. *HUMAIN Launches HUMAIN Chat, Powered by ALLaM 34B.* 25 August 2025. <https://www.spa.gov.sa/ar/N2385004>. إعلان منتج للاستخدام، بلا بيانات إيراد

منشورة.

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>. قياس أداء عتاد ونظم، وليس دراسة عائد رأسمالي.

S12. صفوان المضيان. *Reproducible Synthetic Pricing–Demand–Capacity Model.* analysis/reproduce.py مع data/model_assumptions.json، وجداول النتائج والاختبارات ضمن هذه الحزمة. مصدر حسابي أصلي، لا مصدر رصد للسوق.