



ورقة بحثية استكشافية

التاريخ التحريري: 23 أكتوبر 2026

من مصنع الرموز إلى مُجدول الحرارة: بحث تشغيلي في اقتصاد الاستدلال بالمملكة العربية السعودية

النوع العلمي: ورقة عمل مفاهيمية وحسابية (*conceptual and computational working paper*) غير محكمة، مع تجربة برمجة خطية أُجريت فعليًا على بيانات مختلفة، ومراجعة مصادر أولية وحالات تشغيل معلنة. ليست دراسة ميدانية على منشآت Groq، ولا اختبار أداء مستقلًا لرقاقاتها، ولا تقديرًا وطنيًا لاستهلاك الماء.

ملخص بالعربية. تفحص هذه الورقة إمكان تحويل إدارة الاستدلال اللغوي في السعودية من قرارٍ حول عدد المسرعات إلى تحسين مشترك للعمل الحاسوبي ذي القيمة والزمن والكهرباء وماء التبريد. نقطة الانطلاق الميدانية إعلان Groq تعاونها مع أرامكو الرقمية في 2024 وتشغيل خدمة GroqCloud من الدمام في 2025، إلى جانب وحدة معالجة اللغة (*Language Processing Unit, LPU*) ذات الذاكرة الساكنة داخل الشريحة (*Static Random Access Memory, SRAM*) والجدولة الثابتة لانتقال البيانات، ثم ترخيص تقنياتها غير الحصري لـ NVIDIA في ديسمبر 2025 وظهور منصة Groq 3 LPX في أغسطس 2026. تُفصل وقائع التشغيل عن خطط التوسع والادعاءات المرجعية التي يصدرها الموردون. نصوغ نموذج تشغيل يميز معالجة مدخلات النموذج (*prefill*) عن التوليد المتسلسل (*decode*)، والطلبات الفورية عن العمل الدفعي ذي المواعيد النهائية، واستهلاك الماء داخل الموقع عن الماء المرتبط بكهرباء الشبكة. نُنفذ إجراءً منفصلاً على مرحلتين في نظامي طقس اصطناعيين مدة كل منهما 48 ساعة: نختار أولاً وضع التبريد الأدنى كلفة في كل ساعة، ثم نستخدم البرمجة الخطية لتوزيع العمل المرن تحت سقف قدرة واحد. يعتمد تكافؤ المرحلتين مع الاختيار المشترك على خطية التكاليف وقابليتها للفصل، وعلى افتراض أن التبديل مجاني وغير مقيد. حقق كل حل إنجازاً نحو 4394.2 وحدة من حجم العمل الاصطناعي W من دون خرق المواعيد؛ وهذا الرقم ليس عدد المهام المقبولة Y . في الجو الحار الجاف الاصطناعي، أدت زيادة وزن الماء في دالة الهدف إلى خفض استهلاكه بنحو 607 لترات، مقابل نحو 8.26 كيلوواط-ساعة إضافية قياساً بحل الطاقة الدنيا.

النتيجة برهان إجرائي على قابلية بناء مُحسَّن متعدد الموارد، وليست أثرًا مقيسًا لمنشأة أو تفوقًا مثبتًا لـ LPU. تنتهي الورقة بتصميم اختبار ميداني قابل للتكذيب، ينافس معالجات أخرى ولا يتطلب التسليم بجدوى الجدولة الحرارية سلفًا.

Abstract in English. This working paper separates three evidence layers: Groq’s publicly reported Saudi inference operations, the technical claims of its deterministic SRAM-centric Language Processing Unit, and a local optimization exercise with fully synthetic inputs. A separable two-stage procedure first selects the lower-cost cooling mode for each hour, then uses a 48-hour linear program to allocate 48 flexible jobs with release and deadline constraints while interactive work remains fixed. Equivalence to joint selection holds only under the model’s linear separable costs, free switching, and absence of cross-hour cooling-mode constraints. All policies serve the same synthetic work volume W and respect deadlines; the reported 4,394.2 units are not a task count Y . The experiment demonstrates a conditional electricity–on-site-water frontier, not observed savings in Saudi Arabia, Groq data centers, or LPU hardware. Its main research proposition is to integrate phase-aware serving metrics, physical cooling telemetry, and completed-work quality into an auditable Saudi inference measurement protocol. Vendor benchmarks and national announcements are qualified by their operational and epistemic boundaries. Site consumption, water withdrawal, electricity-related upstream water, real chip energy, network latency, and life-cycle impacts require independent facility data before any generalization.

الكلمات المفتاحية: Groq؛ وحدة معالجة اللغة LPU؛ تدفق البيانات الثابت؛ SRAM؛ HBM؛ Prefill/Decode؛ KV cache؛ البصيلة الرطبة؛ PUE؛ WUE؛ استهلاك الماء؛ الاستدلال؛ برمجة خطية؛ جدولة العمل الدفعي؛ المملكة العربية السعودية.

1. السؤال القابل للبحث وحدود الدعوى

سؤالنا المركزي: إذا كان لدى مشغل استدلال في السعودية طلب فوري ملزم بالزمن وآخر قابل للتأخير، وتتوفر بيانات موثوقة عن أثر العبء على الكهرباء وطرق التبريد والطقس، فهل يستطيع أن يخفض الاستهلاك المائي أو الكهربائي لكل مهمة مكتملة صحيحة من دون تجاوز متطلبات الزمن والجودة؟ لا تصح الإجابة العلمية بـ«نعم» حتى تُعرف استجابة المعدات والعمال الفعليين؛ ولذا نقدم هنا ثلاثة مخرجات أقل ادعاء وأكثر قابلية للتدقيق:

1. إعادة بناء حالة Groq السعودية من إعلاناتها، مع تصنيف نوع الدليل والتمييز بين حدث رُصد علنًا واتفاق أو خريطة طريق أو اختبار موزّد.

2. نموذج سببي وتقني يربط طول المدخلات، وأسلوب التوليد، والذاكرة، والمُجدول، وطاقة الخوادم، ورفض الحرارة، والطقس، واستعمال الماء.

3. تجربة تنفيذ حسابية على بيانات مختلقة تُظهر مساحة المفاضلة وإخفاقات التعميم، ثم بروتوكولًا لإحلال قياسات منشأة حقيقية مكان دوالنا المبتكرة.

لا نفترض أن الاستدلال الأرخص لكل رمز هو الاستدلال الأوفر لكل مهمة، أو أن الطاقة الأقل تتزامن حتمًا مع الماء الأقل. ولا نفترض أن استخدام رقاقة متخصصة يعني أفضلية في كل نماذج الشبكات العصبية. كما لا ننسب تجربة الـ 48 ساعة إلى Groq أو HUMAIN؛ إنها صُنعت لهذا العمل ولا تستنسخ أي بنية معلومة لهما. هذا القيد ليس هامشًا أخلاقيًا، بل شرط صلاحية الاستنتاج: لا يمكن اختبار تحسين منظومة كهربائية وحرارية بإدخال أرقام تتظاهر أنها مأخوذة من منظومة لم نصل إليها.

2. سلّم الأدلة: ماذا حدث فعليًا في الحالة السعودية؟

نقترح في هذه الورقة خمسة أوصاف للأدلة: إعلان نية، عرض تطبيق، إعلان تشغيل خدمة، قياس موزّد أو معيار على منصة معلنة، قياس مستقل قابل للتكرار داخل موقع بعينه. لا يقع التسلسل على خط مستقيم؛ فقد تملك شركة اختبارًا مستقلًا لمسرّع في أمريكا قبل أن توثق تشغيله في مكان آخر. ويوضح هذا التصنيف كيف يضلّل جمع العناوين الصحفية تحت كلمة «إنجاز» واحدة.

تاريخ	الواقعة	الوصف الذي تدعمه المصادر	ما لا تسمح باستنتاجه
12 سبتمبر 2024	Groq وأرامكو الرقمية: شراكة مركز استدلال؛ مشهد حوار صوتي؛ سوق nawat	إعلان تعاون وعرض تفاعلي؛ تقديرات لحجم مستقبلي <u>S01</u>	أن جميع طاقات المركز المخططة قد صارت مركبة أو ذات إيراد
ديسمبر 2024 وفبراير 2025	تركيب عنقود في ثمانية أيام؛ إعلان خدمة GroqCloud من الدمام <u>S02</u>	واقعة تشغيل معلنة من المشغل، ذات فترة ومكان محددين	قدرة الكهرباء، استهلاك الماء، عدد الخوادم المفصل، الإيراد المحلي
فبراير 2025	التزام سعودي بقيمة 1.5 مليار دولار لتوسع البنية <u>S02</u>	إعلان التزام مالي لتوسع البنية	مبلغ الإنفاق الفعلي إلى اليوم أو عدد الرفوف التي اشترت
مايو 2025	Groq موفر استدلال معلن لـ <u>S03</u> HUMAIN	علاقة توريد معلنة	الحصة الفعلية من طلبات HUMAIN ونوع كل عملية شراء
24 ديسمبر 2025	ترخيص Groq غير الحصري لتقنياتها لـ NVIDIA؛ انتقال المؤسس وفريق هندسي <u>S06</u>	اتفاق تقني معلن؛ Groq مستقلة قانونيًا حسب بياناتها	بيع الشركة ككيان بالكامل أو نهاية خدماتها
24 أغسطس 2026	دخول Groq 3 LPX NVIDIA الإنتاج الكامل؛ Nebius أول متبنٍ	حالة منتج وتصنيع وتبنٍ معلنة <u>S07</u>	أن هذه النسخة تعمل بالفعل في مركز الدمام
أغسطس 2026	Groq تعلن عملها مع Dell لنشر LPX وتخطط لتبنيه <u>S12</u>	ترتيب توريد وهدف نشر؛ صياغة مستقبلية	تركيب هذا المنتج في السعودية بتاريخ التوثيق

التناقض الظاهر في المكان. ذكرت بيانات فبراير ومايو 2025 «الدمام»، بينما تسرد صفحة التشغيل الحالية على groq.com نقطة **SA-1 — Riyadh**. لا يقدم المصدر الحديث تفسيرًا تشغيليًا لهذا الاختلاف، بل يسرد 13 مركزًا عالميًا ويذكر بعض المواقع. نسجل المصدرين دون تركيب قصة انتقال أو افتعال مركزين؛ يلزم إثبات مباشر من الجهة المشغلة أو وثائق منشأة قبل أي رسم جغرافي للعنقود. S02؛ S03؛ S08

ما معنى أن تبني Groq رقاقتها الخاصة؟ أعلن المورد أن LPU وما يتصل بها صُمم وصُنع وُجِّع في أمريكا الشمالية؛ لذلك فعبارة «مصنع استدلال في السعودية» تُشير إلى **موضع الحوسبة والخدمة**، لا إلى ملكية تقنية التصميم ولا إلى امتلاك مصنع السيليكون. قد تُشغل السعودية نموذجًا عربيًا على رقاقة أمريكية، في مركز سعودي، عبر واجهة تجارية عالمية. يتطلب قياس القيمة المحلية معرفة ماذا أنجز في كل طبقة دون القفز من واحدة إلى الأخرى.

3. لماذا لا تتناسب سرعة الرموز طرديًا مع FLOPS؟

الأداة التحليلية الأولى من هندسة الحواسيب هي **الكثافة الحسابية** $I = F/B$ ، حيث F عدد العمليات و B البايتات اللازمة لإحضار البيانات. إذا بلغت قدرة الحساب C عملية في الثانية وعرض حزمة الذاكرة M بايت في الثانية، فإن السقف النظري المبسط للسرعة هو:

$$P_{\max} \leq \min(C, I M)$$

ليس هذا قانونًا كافيًا لحساب أداء نموذج حقيقي؛ إنه تقريب **Roofline** يستبعد جدولة النواة، واستدعاءات الشبكة والعمليات الصغيرة، والتشتت بين المعالجات، وتكرار البيانات وأزمنة التزامن. لكن التمييز الذي يصنعه أساسي: حين يكون $IM < C$ تصبح الزيادة في قدرة الضرب الحسابي وحدها غير مؤثرة حتى يتحسن تحريك الذاكرة أو إعادة استعمال البيانات.

في **Prefill** يمكن معالجة كثير من مواضع السياق على التوازي؛ تتشكل عمليات كثيفة على شكل مصفوفات. وفي **Decode** لنموذج مُولّد ذاتي الارتداد، يتوقف الرمز التالي على الرمز السابق. تعدد المستخدمين يسمح بتجميع طلبات وتشكيل دفعات، لكنه لا يُزيل التبعية داخل تسلسل مستخدم واحد. قد تتيح الذاكرة الأعلى عرضًا زيادة معدل توليد الرموز الفردي، إلا أن وضع جميع معاملات النموذج وحالات KV فيها مسألة توزيع وإدارة سعة.

للتدقيق في الوحدات، لنفترض نموذجًا محايدًا فيه **70 مليار مُعامل**. الأحجام الدنيا الخام للأوزان، قبل أي كائن إضافي، هي:

تمثيل الوزن	الحساب	حجم الأوزان فقط
16 بت	70 مليار $\times$ 2 بايت	140 GB عشريًا
8 بت	70 مليار $\times$ 1 بايت	70 GB عشريًا
4 بت	70 مليار $\times$ 0.5 بايت	35 GB عشريًا

هذا **حساب تعريف لا Benchmark**. لا يخبرنا عدد LPUs المطلوبة لخدمة نموذج، لأن التوزيع بين الشرائح يعتمد على تجزئة الأوزان، والKV، وإدارة الذاكرة الأخرى، وسرعات الربط والتكميم، والمترجم البرمجي. وكون رُق NVIDIA Groq 3 LPX يعلن GB SRAM 128 إجمالية مع LPU 256 لا يثبت أن نموذجًا 70B بدقة 16 بت يجب أن يُرفض، بل يبرهن أن **السعة والتوزيع والكثافة** لا تختفي خلف رقم عرض حزمة الذاكرة. كذلك لا تجوز مقارنة رقم الشركة <80 TB/s لذاكرة LPU القديمة ببعض عروض HBM بوصفها مفاضلة عادلة في إنتاج مهمة مطابقة على الجهازين؛ يجب تثبيت النموذج والprecision والدفع وحجم السياق وحالة S04. KV. S05؛ S07؛ S08

4. «الاحتمية» في Groq: ماذا يضمن المترجم البرمجي (Compiler) وماذا لا يضمن؟

تصف Groq بنية *programmable assembly line*: يقسم المُترجم البيانات والتعليمات ويوزع أعمالها على وحدات التنفيذ وروابط النقل. تصبح كثير من حركات الرسائل بين المعالجات **مخططًا لها على أزمنة معلومة**. هذه حتمية داخلية على مستوى عمليات البرنامج، وليست وعدًا بثبات إجمالي زمن استجابة كل مستخدم. يفيد التفريق في هندسة الخدمة؛ فالتحكم داخل الشريحة يقلل مصدرًا من مصادر الارتجال أثناء التنفيذ، بينما تظل هناك متغيرات خارجية: طول السؤال، حالة طابور الخدمة، تزامن نماذج عديدة، تحويل البيانات، تحويل الصوت، الإنترنت، وحدود طلبات S04. API. S05؛ S09

يمكن تصور الزمن الكلي للمهمة، مع التبسيط الواعي، على النحو التالي:

$$T_{task} = T_{net} + T_{queue} + T_{prefill} + \sum_{k=1}^N T_{decode,k} + T_{tools} + T_{post}$$

هذه صيغة **جمع تشغيلي** توضح عناصر الزمن، لا افتراضًا بأن كل عنصر يُقاس بالطريقة نفسها؛ قد تتداخل أجزاء منه في البث المتدفق أو التنفيذ المتوازي. ومن الضروري التفريق بين **TTFT** و **inter-token latency** و **end-to-end**؛ فالأخير يتضمن الإخراج الكامل، بينما قد تصل أول كلمة سريعًا ثم تطول بقية الجملة. وتوثق Groq في واجهات Prometheus مقاييس للطابور و TTFT ونهاية الطلب يمكن، من حيث المبدأ، بناء تجارب حولها. لكننا لم نستخدم واجهة Groq حية، ولم نحصل على مفاتيح أو بيانات لمستفيد سعودي، ولا نزعم استخراج تلك القياسات.

S10

في أنظمة الوكلاء (*agents*) يضيف تكرار التفاعل عبئًا آخر: الطلب القصير ذو خمس دورات توليد متتابة قد يتأثر بوقت الرمز بين الدورات أكثر من طلب طويل ينتظر دفعة واحدة. يجب أن توزن الميزة بتأثيرها في **وقت إنجاز المهمة المعتمدة**، لا في معدل التوليد ضمن اختبار أحادي المستخدم. وهذا أيضًا يفسر قرار NVIDIA المعماري: مزج Rubin GPU و LPX لتحسين المعالجة عبر أطوار متباينة، وفق وصفها. لا ينبغي اختزال هذا الإعلان إلى «الـ GPU فشل والـ LPU انتصر» أو العكس؛ البنية الهجينة نفسها دليل على أن تحسين مرحلة قد يحتاج معماريات متعددة.

S07; S12

5. تعيين المتغير الذي ينبغي تحسينه

نقترح وحدة خرج Y أدق من «عدد الرموز»:

$$Y = \sum_{j=1}^J \mathbf{1}\{\text{quality}_j \geq q_j\} \mathbf{1}\{\text{finish}_j \leq d_j\} v_j$$

حيث j مهمة، و q_j عتبة جودة مقبولة مُسبقًا، و d_j موعد نهائي، و v_j وزن قيمة معن قبل التنفيذ. إذا وضعنا $v_j = 1$ ، أصبح Y **عدد المهام** المقبولة جودةً والمكتملة في موعدها. هذا هو المقياس المفاهيمي المقترح للدراسة الميدانية، وليس مصدر الرقم 4394.2 في التجربة.

أما التنفيذ الاصطناعي فيستخدم حجم عمل منفصلًا: $W = \sum_{t=0}^{47} (U_t + \sum_j x_{j,t})$ ، حيث تمثل U_t وحدات العمل الفوري وتمثل $x_{j,t}$ وحدات الدفعات المنجزة في الساعة t . كلاهما على مقياس اصطناعي مطّيع وفق تعريف المدخلات، مع اختلاف معاملي الكهرباء. تُنجز كل سياسة الحجم نفسه $W \approx 4394.2$ وجميع الوظائف الدفعية الـ 48. يتيح هذا التعريف مقارنة السياسات داخل التجربة، لكنه لا يثبت تساوي قيمة وحدة تفاعلية ووحدة دفعية في خدمة حقيقية. لذلك يجب أن تعرّف الدراسة الميدانية v_j وآلية التحقق من الجودة مسبقًا، لأن التكميم أو نقل المهمة بين نماذج قد يغيران جودة الإجابة. نقترح في العمل الحقيقي قراءة المؤشرين معًا:

$$\eta_E = Y/E_{site}, \quad \eta_W = Y/W_{site},$$

ولا يكفي جعلهما صورتين مقلوبتين لـ PUE و WUE. فمؤشر PUE مقامه E_{IT} ، و WUE وفق معيار الموقع مقامه E_{IT} أيضًا، بينما الهدف الاقتصادي/الهندسي هنا مقامه **مهمة ناجحة في الوقت**. عند انخفاض استخدام الخوادم قد تبدو بعض نسب المنشأة مستقرة لكن عدد المهام المقبولة لكل ساعة يتغير جذريًا.

هناك أربعة مسارات سببية ينبغي اختبارها مستقلة: (أ) اختيار النموذج والتكميم والتجميع يولد كمية عمل وتحريك ذاكرة؛ (ب) الحوسبة تتحول إلى حرارة؛ (ج) نوع التبريد والطقس يحددان الكهرباء الإضافية وماء الموقع؛ (د) مصدر الكهرباء يضيف ماءً خارج الموقع. ومن الخطأ جمع ماء الموقع وماء الكهرباء ثم تسميته WUE بالمعنى القياسي من دون شرح الحدود. توصي أدبيات الماء المعاصرة بتحديد موقع الاستخدام وطريقة التبريد وعوامل استهلاك الماء الكهربائي والعبء الفعلي؛ وقد وجدت مراجعة Lei وزملائه عام 2025 تباينًا بالغًا في استهلاك الماء على مستوى العبء عبر ظروف وتقنيات مختلفة. S11؛ S13

6. فيزياء منع الحَرَج: لماذا تحتاج الحرارة إلى اسمين؟

في مركز البيانات تنتقل الحرارة عبر طبقات: من تقاطع الترانزستورات إلى الحزمة الحرارية، ثم إلى سطح تبديد، فمبرد أو هواء، ثم إلى مبادل، ثم إلى وسيط خارجي. يمكن أن تُستخدم صفيحة تبريد سائل وتظل دائرة السائل داخل الرف شبه مغلقة؛ لذلك فعبارة «تبريد سائل» لا تعني وحدها وجود استهلاك تبخيري مستمر. وفي الحالة التبخيرية ينتقل قدر من الطاقة إلى بخار الماء، وتُستهلك مياه تعويض مع اختلاف عمليات التخلص من الأملاح والتصريف والرشح والمفقود. S11

المتغيران الجويان الأساسيان ليسا مترادفين:

- **درجة البصيلة الجافة** T_{db} : الطاقة الحرارية المحسوسة في الهواء التي تُعرض عادة في تقارير الطقس.
- **درجة البصيلة الرطبة** T_{wb} : حرارة مقياس تحت تبخر ماء من سطح رطب، وترتبط بقدرة الهواء على استقبال بخار إضافي؛ الاقتراب منها يحد التبريد التبخيري النظري.

يمكن أن تختلف مدينتان لهما T_{db} نفسها في T_{wb} اختلافًا كبيرًا بسبب الرطوبة. ويتغير أداء المبادلات أيضًا بحسب اقتراب درجة حرارة المائع من درجة الهواء والقيود التصميمية. ولهذا ليست قاعدة «شغل مهام الذكاء الاصطناعي ليلاً» قاعدة صحيحة على إطلاقها. لا تعني حرارة ليلٍ أقل دائمًا أفضليةً للتبخير، ولا تضمن أسعار كهرباء أقل، ولا تُخبرنا شيئًا عن الماء المستهلك لتوليد تلك الكهرباء.

المقايضة المثالية في انتقال الطور توضح أبعاد الحجم. يستلزم تبخير كيلوغرام ماء عند قرابة 25°C طاقة كامنة نحو 2442 kJ/kg حسب قياسات قياسية، و 1 kWh يساوي 3600 kJ ؛ وبالتالي $1.474 \approx 3600/2442$ كيلوغرام ماء متبخّر لكل kWh حرارة يُزال بالتبخّر الكامل المثالي. عند كثافة تقارب كيلوغرامًا للتر يساوي ذلك نحو 1.47 L/kWh من الحرارة المبخرة. هذا ليس حدًا مطلقًا للماء في نظامٍ متعدد آليات التبريد، ولا يساوي سحب الماء الكلي من الشبكة، ولا كمية الماء على مستوى خدمة سحابية. S14

تعريفات مترابطة ولكن مختلفة:

الرمز	التعريف	الوحدة	الحد المفهومي
E_{IT}	طاقة المعدات الحاسوبية والذاكرة والشبكات المؤهلة ضمن حد IT	kWh	قد تتأثر بالتكثيف والخمول ونوع المعالج
E_{site}	طاقة المنشأة بما فيها IT والتبريد والبنية المساعدة	kWh	لا تساوي كهرباء توليد الشبكة قبل الفواقد بالضرورة
W_{site}	ماء مستهلك داخل الموقع ضمن تعريف التجربة	liter	منفصل عن السحب وعن ماء توليد الكهرباء
(PUE)	E_{site}/E_{IT}	بلا وحدة	ينبغي مقارنته على فترة وحدود قياس متسقة
(WUE)	W_{site}/E_{IT}	L/kWh	هنا قياس موقع؛ لا يشمل upstream
W_{elec}	استهلاك ماء سلاسل توليد الطاقة المنسوب للاستهلاك	liter	يحتاج عامل شبكة بالمكان/الوقت لا يتوفر هنا
Y	مهام مقبولة الجودة ومكتملة ضمن المهلة	عدد أو قيمة موزونة	يجب تعريفه مسبقاً في الدراسة الميدانية
W	حجم العمل الاصطناعي في التجربة	وحدات مطبوعة	ليس عدد مهام ولا مقياساً تشغيلياً مثبتاً

الهدف متعدد الموارد على الحدود الأوسع يصبح مثلاً $E_{site} + \lambda\{W_{site} + W_{elec}\}$ ، لكننا لم نحسب W_{elec} لعدم توافر عامل كهرباء سعودي هامشي محدد بالساعة وموقع التوليد ونوع الماء؛ ولذلك لم نقدمه صفرًا على أنه حقيقة، بل أبقيناه خارج حدود النموذج.

7. تصميم التجربة الاصطناعية: ثماني وأربعون ساعة بلا اسم مدينة

للتمييز بين البرهان الحسابي والبرهان الميداني أنشأنا مجموعتين من الطقس على 48 ساعة، مع دورة يومية مصطنعة متكررة مرتين. يتبع كل من T_{db} و T_{wb} منحني جيبيًا بمرحلة واحدة، وفق معاملات مختلفة لنظام اسمه hot_dry وآخر اسمه coastal_humid. الأسماء أوصاف حسية لسيناريوين، وليست محطات رصد سعودية. ترمز الساعات إلى فترات حسابية؛ ليس للساعة الصفر يوم ميلادي ولا قراءة NCM مقابلة. توثق بوابة المركز الوطني للأرصاد خدمة أرصاد تاريخية وحالية وتوقعات عبر واجهة تتطلب اعتمادًا؛ لم نمتلك بياناتها الموثقة لكل ساعة ضمن هذه المهمة. ولا تبرر جداول المعدلات المناخية السنوية ادعاء ساعة حقيقية في موقع مركز بيانات. [S18](#)

ثم اخترعنا **48 مهمة دفعية** $j = 0 \dots 47$ ، تدخل إحداها كل ساعة، ويجب إتمامها خلال حد يصل إلى تسع ساعات بعد الإتاحة. وأضفنا حملًا تفاعليًا U_t على مدار 48 ساعة؛ لا يسمح النموذج بنقله إلى وقت آخر، فيحفظ استعجاله حتمًا. حددنا الطاقة الاصطناعية لكل وحدة تفاعلية بـ 0.8 kWh، ولكل وحدة دفعية بـ 1.0 kWh. أضفنا 18 kWh من كهرباء IT ثابتة في كل ساعة تمثل الخمول/الاستعداد، وسقفًا مقداره 150 kWh IT في الساعة. ليس لأي «وحدة عمل» في هذا الإعداد معنى مباشر بعدد الرموز أو شريحة Groq أو قدرة مركز حقيقي. إنها عدادات مجردة لتحليل الاختيار الزمني.

منحنى الكهرباء الاصطناعي لكل طريقة تبريد هو نسبة $c_{m,t}$ من طاقة IT عند الساعة t ، حيث $m \in \{dry, evap\}$. ونظيره المائي $w_{m,t}$ بوحدة IT L/kWh. في الشيفرة المعطاة، يرتفع عبء التبريد الجاف عند حرارة هواء مرتفعة، بينما يكون التبخيري أقل كهرباء في بعض الساعات وأعلى استهلاك ماء. ويحتوي الخيار الجاف على حمل مائي اصطناعي صغير غير صفري 0.035 IT L/kWh لتمثيل استهلاكات مساندة؛ فلا يجوز وصفه في الورقة بأنه عملية بلا ماء إطلاقاً.

عنصر إدخال	مقدار مختلق	سبب إدراجه	ما لا يدل عليه
الأفق	48 ساعة	يسمح بالجدولة داخل أيام افتراضية	ليس موسمية سنة كاملة
قدرتنا القصوى	150 kWh IT لكل ساعة	يفرض ازدحاماً معقولاً	لا يمثل قدرة رف تجاري
كهرباء استعداد	18 kWh IT / ساعة	يمنع ادعاء إلغاء الاستهلاك بمجرد نقل العمل	ليست قياساً لخمول Groq
كهرباء وحدة التفاعل	0.8 kWh	توحيد العمل الملزم بالزمن	لا مساواة بوحدة token
كهرباء وحدة الدفعة	1.0 kWh	يجعل نقل العمل قابلاً للحساب	لا مساواة بمقياس GPU/LPUs
مرونة الدفعات	حتى 9 ساعات	يقيد نقل الحمل	لا تعميم على كل Batch API
λ	0 إلى 0.12 kWh/L	رسم مفاضلة رياضية	ليس سعر ماء ولا وزناً رسمياً

الأوامر الحرفية التي تنتج هذه المدخلات في `analysis/reproduce.py`. الاحتفاظ بها في الشيفرة، لا بالاكتهاء بجدول التقريب، مهم لأن أي شخص يمكنه رؤية الدوال القطعية (max) ونقاط تحول التبريد وتعديلها ثم إعادة تشغيل النتائج.

8. المسألة الرياضية وقابلية حلها

ليكن $x_{j,t} \geq 0$ مقدار وحدة الدفعة j المنجزة في الساعة t . لا يسمح بالمتغير إلا إذا كانت الساعة داخل فترة الإطلاق r_j والمهلة d_j . وليكن b_j عمل الوظيفة المطلوبة. يجب أن تنجز كل وظيفة بالكامل:

$$\sum_{t=r_j}^{d_j} x_{j,t} = b_j, \quad x_{j,t} \geq 0.$$

قيد الطاقة لكل ساعة، ضمن صياغتنا ذات القدرة الثابتة أثناء الساعة، هو:

$$18 + 0.8 U_t + \sum_{j:r_j \leq t \leq d_j} x_{j,t} \leq 150.$$

لأننا سمحنا بتجزئة العمل الدفعي إلى وحدات، فإن البرمجة الخطية حل مناسب لهذه اللعبة. في الواقع قد توجد مهام غير قابلة للتجزئة، وتحويل KV وفتح نسخ نماذج وأحجام دفعات صحيحة فقط، وحينها قد نحتاج برمجة أعداد صحيحة مختلطة (MILP) أو حلًا أنيًا متدرجًا؛ لا ندعي أن نموذجًا كسريًا يحل تلك التفاصيل.

تُختار طريقة التبريد في كل ساعة قبل بناء تكاليف البرمجة الخطية، بأصغر قيمة تساوي: $c_{m,t} + \lambda w_{m,t}$. بعد تثبيت m_t ، تصبح تكلفة وحدة عمل دفعية في تلك الساعة $1 + c_{m_t,t} + \lambda w_{m_t,t}$ ، بحيث نحل:

$$\min_x \sum_{t=0}^{47} \sum_j x_{j,t} [1 + c_{m_t,t} + \lambda w_{m_t,t}]$$

تظل طاقة التفاعل والاستعداد جزءًا من التقييم النهائي لجميع السياسات، ولا يُسقطها هدف الجدولة لأنهما ثابتان بالنسبة إلى x عند تثبيت طريقة التبريد. تحت التكاليف الخطية القابلة للفصل في هذا النموذج، يكافئ اختيار التبريد بأصغر كلفة لكل ساعة ثم حل برنامج توزيع العمل الاختيار المشترك المقابل. لا يمتد هذا التكافؤ إلى منشأة واقعية من دون تحقق. فطريقة التبريد هنا بلا تكلفة تبديل ولا حد أدنى لزمان التشغيل، والتبديل بين الأوضاع حر وغير مقيد؛ وكلها افتراضات ينبغي إزالتها في دراسة صناعية. وفوق ذلك، تختلف أنظمة التبريد الفعلية في إمكان اختيار وضع جاف أو تبخيري بالكامل داخل الساعة؛ فمبنى مصمم بلا تبريد تبخيري لا يمكن للبرنامج إضافة برج ماء بضغط زر.

نستخدم محلل SciPy HiGHS / linprog وننتج قيود المساواة لكل وظيفة وقيود القدرة لكل ساعة. تُبنى سياستان مرجعيتان من التشغيل عند أول فرصة (FIFO) مع وضع تبخيري أو جاف دائم، وسياستان تكميليتان: FIFO مع اختيار أقل تبريد كهربائي في كل ساعة، وإجراء محسن من مرحلتين يختار وضع التبريد لكل ساعة أولاً ثم ينقل العمل الدفعي داخل نافذته. في الحساسية نكرر الأمثلة عند ثمانية أوزان للماء بين الصفر و0.12؛ تشمل حزمة قابلية التكرار كل النقاط، لا المختار للعرض فقط. النموذج لا يتضمن توصيف GPU أو LPU داخل دوال الطاقة؛ لذلك لا يصح قراءة أية نتيجة على أنها مقارنة رقاقية.

9. نتائج التنفيذ والفحص الداخلي

نقدنا الملف البرمجي محليًا باستخدام SciPy؛ وثائق التشغيل ونسخة المدخلات والنتائج متاحة. الجدول أدناه يختار سبع سياسات من كل نظام؛ وكل النتائج للفترة ذاتها وبالعبء نفسه. FIFO-energy يعني بدء الدفعة حين تتاح مع اختيار طريقة التبريد الأقل كهرباء لكل ساعة، و λ LP يعني حل البرمجة الخطية مع وزن ماء مختار. تتغير الطاقة والماء تبعًا للوظائف المختلفة؛ لا نمناها أي مرتبة إحصائية خارج هذا النموذج. أعادت مراجعة حسابية مستقلة إنتاج القيم التجميعية أدناه بالدقة المعروضة. لكن للحل أمثليات متعادلة، كما أن إصدارات الاعتماديات غير مثبتة بالكامل؛ لذلك قد يوزع تشغيل آخر بعض وحدات الدفعات على ساعات مختلفة مع بقاء الهدف والحجم الكلي والمواعيد والسعة ومجاميع الكهرباء والماء على حالها. لا ندعي تطابق الجدول الساعي بايتيًا بين بيئات الحل.

الطقس الاصطناعي	السياسة	الكهرباء الكلية kWh	ماء الموقع L	ساعات التبريد الجاف
حار جاف	FIFO-تبخيري	5223.07	6818.11	0
حار جاف	FIFO-جاف	5490.52	165.20	48
حار جاف	FIFO-أقل طاقة	5192.71	4829.34	18
حار جاف	LP, $\lambda=0$	5179.88	4166.87	18
حار جاف	LP, $\lambda=0.02$	5188.14	3559.52	22
حار جاف	LP, $\lambda=0.05$	5243.32	2222.32	30
حار جاف	LP, $\lambda=0.12$	5407.56	165.20	48
ساحلي رطب	FIFO-تبخيري	5277.24	7076.90	0
ساحلي رطب	FIFO-جاف	5504.52	165.20	48
ساحلي رطب	FIFO-أقل طاقة	5261.37	5451.87	14
ساحلي رطب	LP, $\lambda=0$	5245.08	4893.67	14
ساحلي رطب	LP, $\lambda=0.02$	5263.52	3640.80	22
ساحلي رطب	LP, $\lambda=0.05$	5313.90	2248.73	30
ساحلي رطب	LP, $\lambda=0.12$	5438.69	165.20	48

اختبار الثبات الداخلي. كل سياسة أنجزت تقريباً 4394.2 وحدة من حجم العمل الاصطناعي W ، منها حمل فوري ثابت و48 وظيفة دفعية، مع صفر تجاوز لمواعيد إتاحة أو إنجاز الوظائف وصفر تجاوز لسقف IT. هذا الحجم ليس عدد المهام المقبولة Y . تتحقق الشيفرة من كل مجموع تكاليف، ومن مساواة العمل المجزأ للعمل المطلوب، ومن وقوع كل جزء داخل نافذة الوظيفة، وتمنع قبول نتيجة محلل غير ناجحة. إلا أن هذه التأكيدات تختبر سلامة تنفيذ النموذج كما كتبناه؛ لا تختبر واقعية معامل حراري أو سلوك عميل أو جهاز Groq.

نقطة مهمة في قراءة الجدول: عند الانتقال من $LP \lambda=0$ إلى $LP \lambda=0.02$ في الجو الحار الجاف، ارتفعت الكهرباء kWh 8.26 بينما انخفض الماء L 607.35 في 48 ساعة. في الجو الساحلي الرطب ارتفعت الكهرباء kWh 18.43 وانخفض ماء الموقع L 1252.87. هذه أرقام دقيقة لبرنامج مختلق، ولا ينبغي تقديم النسب المئوية الناتجة عنها على أنها عوائد اقتصادية أو بيئية لخطة سعودية. توفر مقارنة العالمين درساً منهجياً: معاملات تبريد مختلفة يمكن أن تنقل نقطة مبادلة كيلوواط-ساعة إضافية بترتات ماء موفرة.

حافة المفاضلة المحسوبة على مجموعتين مختلفتين

الشكل 1. كل نقطة ناتجة من محلل فعلي، والمعامل λ مُعلن. لا توجد نقطة في الشكل تعود إلى قياس ميداني أو شبكة كهرباء واقعية.

10. لماذا تنجح اللعبة الحسابية وتظل ناقصة أمام الآلة؟

نظريًا يظهر تحريك الدفعات كأنه استبدال مجاني للأوقات. عمليًا يمكن أن تنهار فائدة التحريك في أربعة مواضع:

أولاً: الطاقة الساكنة والإطفاء. إذا بقيت الرفوف مستهلكة أكثر قدرتها الكهربائية أثناء السكون، لن يخفض نقل المهمة جزءًا كبيرًا من الطاقة. إذا اتبع المشغل سياسة إطفاء عميق فلن يكون الاستيقاظ فوريًا وربما تحتاج النماذج إلى دقائق التحميل وإعادة امتلاء الذاكرة. ميزة نموذجنا أنه يفرض حمل خمول ثابتًا فلا يختزل طاقة المنشأة إلى مجموع طاقة كل مهمة فقط؛ عيبه أنه لا يختبر إجراءات الإطفاء والاستيقاظ والاستعداد للطفرات الفورية.

ثانيًا: الاحتفاظ بالموديلات في SRAM/HBM. الدفعة التي تنتقل من وقت لآخر قد تصطدم بحد سعة الذاكرة. طرازٌ ثري بالأوزان وKV يتطلب توزيعًا بين شرائح؛ وتكلفة تفريغ حالة واستعادة أخرى قد تأكل مكسب التبريد. يجب قياس *model residency, KV cache occupancy, swap traffic, compilation cache* و *interconnect load* جنبًا إلى جنب مع الطاقة لا بعد انتهاء التجربة.

ثالثًا: متطلبات الخدمة ومزج النموذج. لحظة ازدحام Batch بمفردها ليست مهمة، لكن قد يكون المستخدم تفاعليًا في طابور آخر على الأجهزة نفسها. يجب أن يبقى ذيل زمن TTFT ووقت تكملة الطلب ضمن مهلة كل فئة، لا الاكتفاء بالمعدل المتوسط. والتباين في طول السياق والمخرجات يمكن أن يبدل تمامًا الطلب على Prefill وDecode. لذا يتعين على الاختبار الحقيقي أخذ عينات من أطوال وتوزيع موجات الوصول، لا افتراض وحدات متجانسة.

رابعًا: الديناميكا الحرارية متعددة الدوائر. تتحدد فاعلية التبريد بالمساحة السطحية، ونوع مبادل الحرارة، وتكلفة ضخ المائع، ودرجة مدخل الخادم، وكمية المطر والغبار ومعالجة مياه التعويض، لا بترموتر الهواء فقط. ليس من العلمية افتراض انتقال نظام فعلي بين وضع جاف وتبخيري في كل ساعة دون مواصفات مصنعه. ستحتوي الدراسة الميدانية على منحنيات عمل حقيقية، *part-load performance, cooling approach temperature, fan curves, pump curves*، وتكلفة تغيير الوضع.

هذا هو الحد المعرفي بين تجربة إمكان (existence proof) وتجربة أثر (effect estimate). الأولى تثبت أن تحويل المرونة إلى تحسين متعدد الموارد ممكن حسابيًا عند اختيار معاملات تؤيده. الثانية تتطلب إثبات أن هذه المعاملات تمثل الآلات والطقس والأحمال الواقعية، وتُقارن مع نظام التشغيل القائم بتوزيع عادل للمهام والقياسات.

11. أبحاث سابقة: ما الذي أنجز قبل هذه الورقة؟

لم تبدأ هذه المسألة العلمية مع هذه الورقة. ففي مراجعة بيركلي لعام 2025 درس الباحثون العلاقة بين كفاءة الخادم و طاقة الكهرباء وماء الشبكة ونوع التبريد والتغير المناخي. وتعرض وزارة الطاقة الأمريكية PUE و WUE كنسب مفيدة مع تحذيرات حدود القياس؛ أي إن التفريق بين موقع المنشأة واستهلاك الموارد كان معروفًا قبل هذه الدراسة.

S11؛ S13

وفي ورقة **Balancing Bits and Drops: Stress-Adjusted Water Management for Data Centers** لعام 2026 بحث مؤلفوها أثر ضغط الماء المكاني والزمني، وقدموا تنظيمًا لأعباء التشغيل يأخذ الماء في الاعتبار، مع إمكانية النظر إلى التبريد الجاف والماء خارج الموقع. هذا اعتراض مباشر على أي ادعاء بأننا «ابتكرنا أول مُجدول للماء». الإضافة الاختبارية المقترحة هنا أضيق: دراسة الاستدلال متعدد المراحل على مورد سعودي، مع فصل أنواع الوظائف وبيانات الرقابة والتبريد والساعات. S15

أما **PowerSlider** الذي نُشر كمسودة في أغسطس 2026 فتناول قيود القدرة اللحظية وجعل تفاوت Prefill/Decode وأزمة جودة الخدمة جزءًا من حل متكرر قصير المدى؛ ذكر مؤلفوه بقاء Goodput تجريبيًا عند 78.3% لخفض قدرة قدره 30%، مقابل 47.6% في أقوى خط أساس لديهم ضمن البيئة التي اختبروها. لا يجوز نقل تلك المقادير إلى LPU أو Groq أو السعودية، كما أن هدف الورقة الكهربائية ليس مطابقة هدف ماء التبريد عندنا. لكنها تثبت أن **جدولة مرحلية واعية بالقدرة** ليست مجرد موضوع محاثة نظرية؛ لها نموذج تنفيذي وأعباء وتجارب منشورة يمكن تحديدها. S16

وأخيرًا، أوراق ومطالبات الموردين أنفسهم تمنحنا مقاييس خاضعة للاختبار: وصفت NVIDIA على منصة 3 Groq LPX نتيجة مرجعية لمخرجات Gemma 4 31B وذكرت أداءً عاليًا في أعمال الوكلاء. أنصف النتائج بإحالتها إلى منصة ونموذج وسياق وشروط Benchmark، لا بتحويلها إلى متوسط جميع خدمات السعودية. كما أن اختلاف القياسات بين رمز لمستخدم واحد ورموز إجمالية ووقت المهمة كلها يفرض مقابلة عادلة حين تُختار المقارنات. S07

12. تصميم تجربة ميدانية سعودية تُفشل الفرضية إن أخطأت

هذا الجزء **مقترح ولم يُنفذ**. يختلف عن تنفيذ Python المبين أعلاه. نحتاج إذنًا من مُشغل مركز حوسبة، وقراءة قياس كهربائي ومائي مصدق، وقدرة على نسخ الأعباء دون تعريض بيانات مستخدمين أو إنتاج حقيقي للخطر. لا يفرض التصميم كشف بنية خاصة للعامة؛ يكفي نشر تجميعات إحصائية منضبطة تتيح فهم الطريقة ومراجعة المعادلات.

العينة والطبقات. يُجمع أثر تشغيل حقيقي لأسابيع تتضمن تنوعًا في درجات الحرارة والرطوبة وطول الطلبات. يُقسم العمل إلى ثلاثة أصناف مسبقة: محاثة حية بزمان حدي، مهام وكيل متعددة الجولات بزمان معقول لإنهائها، وعمل دفعي ذو مواعيد نهائية حقيقية. يُحفظ توزيع المدخلات والـ outputs والعلامات الزمنية وإصدارات النماذج في كل

سياسة تجريبية. تُستبعد النصوص الخاصة من قاعدة التجربة أو تُحوّل إلى قياسات طول مجهولة بعد تعريف صريح لما يمكن مقارنته.

أذرع المقارنة. (أ) المُجدول الإنتاجي كما هو؛ (ب) جدولة واعية بزمن الدفعات وحمل GPU/LPU فقط؛ (ج) جدولة واعية بالزمن والكهرباء؛ (د) جدولة واعية بالزمن والكهرباء وماء الموقع، ثم (هـ) ذراع حسابي يضيف عامل ماء توليد الكهرباء إذا توفر مصدر موثوق. يجب أن يخضع كل ذراع لنفس سقوف الجودة والخدمة والطاقة الممكنة؛ وأن تُصمم آلية حظر إذا بدأت المقارنة تضر طلبًا فعليًا. ليست ثمة أفضلية مفترضة للذراع الأخير: إذا لم يُحسّن العمل المفيد لكل وحدة مورد ضمن فواصل الثقة فترفض دعواه.

طبقة العتاد. تُقارن بنية عامة GPU مع LPU أو منصة هجينة حين تتاح النماذج نفسها بدقة وجودة قابلتين للمقارنة. يُوثّق شكل الدقة (*precision*)، نوع التكميم، طول السياق، سعة الـ KV ومواقع تحميل النماذج، عدد المستخدمين المتزامنين، الشبكة، وتسلسل الأدوات. لا يصح أخذ واجهة Web Demo على مسرّع ومقارنتها ببيئة إنتاج مختلفة في شبكة أخرى ثم عزو الفارق إلى الفيزياء وحدها. وإذا عجز نموذج عن العمل على رقاقة معينة بشروط الجودة نفسها، تكون تلك نتيجة أهلية مهمة، لا حالة تُحذف من التقرير.

طبقة البيئة. يُجمع في الزمن ذاته T_{db} و T_{wb} وحرارة تغذية السوائل وعودتها وقدرة المراوح والمضخات، وعداد ماء التعويض والسحب والاستهلاك منفصلين، ومصدر الماء ونوعه إذا كان متاحًا. وتُقرأ كهرباء IT و PDU و UPS والتبريد عند حدود عداد معلنة بوضوح. يُنشر مدى دقة أجهزة القياس وفترات أخذ العينات والفروق بين ساعة الكهرباء وساعة الطلب كي لا نختلق ارتباطًا زمنيًا من قياسات متأخرة.

مقاييس النجاح. يتصدرها عدد المهام المكتملة صحيحة ضمن المهلة، وطاقة المنشأة لكل ألف مهمة صالحة، وماء الموقع المستهلك لكل ألف مهمة صالحة، مع PUE/WUE بالموازاة فقط، و TTFT وزمن الإنجاز عند p50/p95/p99. ويجب الإبلاغ عن معدل فشل الجودة وإعادة التشغيل، لأنّ تسريع النظام بإخراج جواب غير صالح ليس إنتاجًا أفضل. تُؤخذ المدة والتكلفة التشغيلية اللازمة لتحضير وتجميع نماذج متعددة في الحسبان، فلا يُختزل مجال عمل يوم كامل إلى لحظة أسرع Token.

حدود التعرف السببي. لا تكفي مقارنة أسبوع معتدل الحرارة بأسبوع حار. تغيّر خليط الطلبات والنسخ البرمجية والطقس والقدرة المتاحة والطاقة السعرية في الوقت نفسه. الأفضل توزيع عبء اختباري مرخّص بالتناوب العشوائي داخل نوافذ متقاربة الظروف، مع حجب انتقال آثار الذراع السابق كوقت تشغيل الرف أو تهيئة النموذج، أو استعمال تصميم crossover ذي فترات تطهير. وعلى الباحث أن ينشر التوزيع المُسبق للنتائج التي سيحكم بها، والنتائج السلبية وتباين الشروط، قبل إصدار استنتاج على مستوى المملكة أو تكاليف مشاريع بالمليارات.

13. الاقتصاد التالي: من قيمة المُعالج إلى قيمة تحويل الكهرباء إلى وقتٍ نافذ

تستطيع مدينة ذات منشآت حوسبة أن تباع قدرةً على الورق، وأن تباع عدد رموز إنتاج خام، وأن تباع مهامًا بجودة وزمن ملزمين. تختلف الثلاثة في مقدار المهارة الصناعية المنقولة. عند النوع الأول يكون المنتج سعة مجهزة؛ عند الثاني يكون منتجًا محاسبيًا أصغر؛ أما عند الثالث فتحتاج الشركة إلى هندسة أطوال السياق، وملاءمة النماذج مع المعالج، وعرض الشبكة، والمُجدول، وعدادات الطاقة والماء، وتجربة تضمن أن العمل مفيد. ومن الممكن أن ترتفع القيمة المنتجة حين يتحسن توزيع الحمل من دون زيادة عدد المسرعات، كما يمكن أن تمحو تكلفة التنفيذ والتعقيد تلك الزيادة. ليس ذلك بيان عوائد؛ إنه افتراض ميكرو-اقتصادي قابل للاختبار.

في السعودية تظهر فرصة عمل تشغيلية مجاورة لخدمة Groq الأصلية: تطوير محوّلات عبء الاستدلال بين معماريات متباينة وخوارزميات جدولة ذات قياس مرتبط بالظروف المحلية، ومكتبات Benchmark تربط نوع الموديل وسياقه بجودته في العربية وغيره. يمكن بيع هذه المعرفة في منصات تقيس **accepted tasks per joule** و **accepted tasks per site liter**، وتفصل مهمة تحتاج أجزاء الثانية عن مهمة تستفيد من مرونة ساعات. لكن لا تُستنتج المبيعات أو الميزة الدفاعية من كون السوق العالمي يستثمر بكثافة؛ ينبغي إثبات قيمة القرار الناتجة عن البرمجيات أمام حلول NVIDIA و Groq الأصلية وأعمال الجامعات ومزودي التبريد. وإلا سنعيد تسمية أدوات موجودة لا أكثر.

الاتجاه المستقبلي بين 2027 و 2035 ليس توقعًا رقميًا؛ إنه شجرة اختبارات: إذا ظهرت بيانات تبريد معيارية قابلة للاستخدام وتعاملت المنصات مع تعدد المعالجات بتكلفة نقل محدودة، فقد تستحق الجدولة الفيزيائية أن تصبح جزءًا من **Control plane** الخاص بخدمة الاستدلال؛ وإذا ظلّت الأحمال الفورية هي الكتلة العظمى أو استهلك نقل الحالة أكثر مما يوفره التوقيت، فالأفضل الاستثمار في خفض طاقة الرمز عن طريق **Compiler** والتكميم والكفاءة الكهربائية. وإذا تعارض توفير ماء الموقع مع ماء الكهرباء من خارج الموقع، فلا يصح إعلان فوز قبل تحديد حدود المحاسبة. هذه الاستنتاجات مشروطة ويمكن تعديلها بنتيجة تجريبية، لا تُعلن كحقيقة آتية في 2035.

14. قابلية التكرار والاستقلال

تتضمن حزمة قابلية التكرار الشيفرة وبيانات المدخلات والنتائج القابلة للمراجعة:

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

الاعتماد على **NumPy** و **SciPy** و **Matplotlib**، وليس على مفاتيح API أو استدعاء داخلي لمنتج خاص. ملفات `data/workload_synthetic.csv` و `data/weather_synthetic.csv` هي المدخلات المنتجة اصطناعيًا، و `data/policy_hourly.csv` و `data/policy_summary.csv` و `data/policy_sensitivity.csv` و `data/run_manifest.json` سجل إعدادات التشغيل. مصنف

evidence_workbook.xlsx يقدم جدول نتائج قابلاً للفحص وصيغاً حسابية ومراجع إلى المدخلات. الرسوم تحفظ بصيغتي PNG و SVG، وفهرسها في figure_register.csv يضع علامة صريحة على طابعها الاصطناعي.

لا توجد عينة احتمالية من المناخ أو الشركات في هذه الورقة، ولا تحليل فاصل ثقة يمكن تبريره حول نتيجة صناعية. محلل HiGHS يحل مسألة خطية مكتوبة؛ صحة الحل بالنسبة إلى الفرضيات لا تجعله تقديرًا سلوكيًا أو وطنيًا. لم نختبر صفوفًا حقيقية من مراكز السعودية، ولم نؤكد معدل الاستهلاك لكل رمز على شريحة Groq. يستطيع القارئ إعادة تشغيل الشيفرة ومن ثم استبدال المنحنيات بعلاقات مقاسة بشكل مرخص واختبار الفروق. والتمييز بين التكرار الحسابي (computational reproducibility) وإعادة القياس المستقلة (empirical replication) شرط أساسي: أعيد إنتاج المجاميع بالدقة المعروضة، مع إمكان اختلاف التوزيع الساعي بين حلول متعادلة في الأمثلة؛ أما إعادة القياس الميداني المستقلة فهي مشروع علمي لاحق لم يُنجز.

15. أشياء لا نعرفها ويجب ألا يتظاهر الرسم بأنه يعرفها

لا تكشف المصادر العامة هنا عن استهلاك مركز Groq في الدمام للكهرباء والماء ولا عن تصاميم رفض الحرارة فيه، أو نوع وكمية رقائق كل موقع، أو عدد Tokens المنتجة من داخل السعودية وحدها، أو أي قدرة ماء إضافية قُصدت مقابل التزام الاستثمار. لم نجد إثباتًا يربط منصة NVIDIA Groq 3 LPX الجديدة بمركز سعودي بعينه، وإن كانت Groq أعلنت أنها من الأوائل الذين يخططون لتبنيها. لم تُبرهن وجود فارق أداء مستقل بين Groq LPU و GPU عند شروط مهمة موحدة من داخل المملكة. ولأن تعبير «موفر للاستدلال لـ HUMAN» علاقة تجارية معلنة وليس سجلًا لفواتير حقيقية، لا نستنبط منه حجم استخدام أو إيراد.

كذلك لم نجلب بيانات ساعة من NCM؛ التسمية «ساحلي» في التجربة مجرد قويس جوي مصطنع. ولم نقس WUE لأي مركز، ولا هامش ماء تحلية أو ماء كهرباء الشبكة. وقد تكون للمركز وسيلة تبريد أو قيود تشغيل لا تسمح بالاختيار الذي نفترضه. كل هذه الحدود ليست عيوب صياغة يمكن إصلاحها بحشو توصيات؛ إنها الحدود الدقيقة لما يستطيع هذا الدليل إثباته.

الخلاصة القابلة للاعتراض: تملك السعودية إشارة تشغيل حقيقية في استدلال Groq، وأمامها فرصة اختبار برنامج يحسن العمل الصحيح المنجز في الوقت لكل وحدة كهرباء وماء. ولكن النجاح العلمي والاقتصادي لا يبدأ من افتراض أن LPU أفضل في كل شيء، ولا من وعد يساوي التحول إلى التبريد الجاف بالاستدامة. يبدأ من سؤاليين تشغيليين: ما الذي يحد سرعة المهمة حقًا، الذاكرة أم الحساب أم الشبكة أم الطابور؟ وما الطاقة والماء التي احتاجتها هذه المهمة في هذا المكان وهذه الساعة وفق عدادات قابلة للتدقيق؟

المراجع ومصادر الاقتفاء

المواد أدناه هي مصادر أصلية للمؤسسات أو الأوراق العلمية، ولا تُعامل الإعلانات التسويقية كتحقق مستقل. النسخة المفهرسة sources.md تضم التواريخ والقيود والروابط، وsource_register.csv يحتوي مواضع الإثبات.

- [.S01 Groq / Aramco Digital, partnership announcement, 12 Sep 2024](#)
- [.S02 Groq, Dammam operations and Saudi expansion commitment, 10 Feb 2025](#)
- [.S03 Groq, HUMAIN inference provider, 27 May 2025](#)
- [.S04 Groq, The Groq LPU Explained, 7 Mar 2025](#)
- [.S05 Groq, GroqChip Processor product brief, 2024](#)
- [.S06 Groq, non-exclusive NVIDIA inference technology license, 24 Dec 2025](#)
- [.S07 NVIDIA, Groq 3 LPX full production, 24 Aug 2026](#)
- [.S08 Groq, infrastructure and 13-center listing](#)
- [.S09 Groq, production service tiers and batch APIs](#)
- [.S10 Groq, Prometheus request latency metrics](#)
- [S11 U.S. Department of Energy, Cooling Water Efficiency Opportunities, 9 Jan 2019](#)
- [.S12 Groq, NVIDIA Groq 3 LPX and Vera Rubin future adoption, 2026](#)
- [S13 Lei et al., *The water use of data center workloads*, 2025, DOI 10.1016/j.resconrec.2025.108310](#)
- [S14 U.S. National Bureau of Standards/NIST, water latent-heat measurements around 25°C](#)
- [.S15 Talukder et al., *Balancing Bits and Drops*, arXiv:2607.22617, 2026](#)
- [.S16 Li et al., *PowerSlider: Exploiting Phase Asymmetry*, arXiv:2608.21719, 2026](#)
- S17 صفوان المضيان، analysis/reproduce.py؛ تجربة منفذة فعليًا على مدخلات اصطناعية متعمدة، وجميع الافتراضات محلية.
- [S18 Saudi National Center for Meteorology, weather-data API, availability documented but user credentials required; the API was not queried for this study](#)