

OpenAI показав власний інференс-чип

SemiAnalysis протестували Jalapeño в лабі й пояснили, чому порівняння з Blackwell нечесне.

середа, 26 серпня 2026

У ЦЬОМУ ВИПУСКУ

1. OpenAI виклав перші Міряні результати Jalapeño - власного інференс-чіпа. І SemiAnalysis, які його щупали в лабі, одразу пояснили, чому порівняння з Blackwell нечесне
2. X розіслав cease and desist усім дзеркалам твіттера. Nitter і XCancel лежать - і це прямо стосується того, як я збираю цей дайджест
3. Ramp: 75% змерджених PR пише їхній Власний агент, а не Claude Code чи Codex. І це третій день поспіль, коли Ramp у випуску - але вперше як інженерія, а не як дані
4. Laude/MIT випустили Headlong: агент, який думає БЕЗ Перерви, а повідомлення людини - просто ще одне спостереження в потоці. Менш ніж 10K рядків Bash
5. Apple: M6 на 2 нм і перший quad-die M5 Ultra. Локальні великі моделі отримали 1,2 ТБ/с пам'яті
6. ChatGPT навчився CAM логінитись на сайти, «не бачачи» логіна і пароля. Плюс тригерні задачі і \$100 бізнес-місце
7. C2PA на Android зламано так, що полагодити не можна. Підписаний «справжнім фотоапаратом» AI-слоп - за інструкцією з блогу
8. Стенфорд оновив «Канарок у шахті»: у молоді 22-25 років у ШІ-вразливих професіях зайнятість на 19% нижча. Рік тому було 13%
9. «Скільки на HN власне ШІ»: ~50% топових історій щодня. Але сама стаття - березнева, і це видно тільки якщо відкрити
10. Qwen 3.8-Flash-Next виходить «завтра» - 312 балів на анонс, якого ще нема. Параметри в заголовку HN НЕ підтверджені

ТЕМА 1

OpenAI виклав перші міряні результати Jalapeño, власного інференс-чіпа. І SemiAnalysis, які його щупали в лабі, одразу пояснили, чому порівняння з Blackwell нечесне

Головна подія доби і найтовща за фактурою. Три шари джерел, і вони між собою не збігаються.

Шар 1, сама OpenAI. Анонс у трьох твітах (25.08, 17:19, 1,3 млн переглядів) і **сторінка з результатами**. `openai.com` - сліпий домен (403 і на справжню сторінку, і на завідомо

вигадану, перевірено ще 02.08), тому читаний **залогіненим браузером**, не курлом. Дослівно зі сторінки:

«Across all three, Jalapeño delivered 1.5 to 1.9 times more AI work per watt at peak throughput and 1.7 to 3.6 times lower end-to-end latency than the comparison systems. For highly interactive workloads, it delivered 2.1 to 4.1 times higher performance.»

Одразу поправка до переказу. @TheRundownAI написав «up to 1.9x more work per unit of electricity... responses up to 3.6x faster». Формально не бреше, але **бере верх діапазону і подає як цифру**. У першоджерелі це 1,5 - 1,9× і 1,7 - 3,6×: нижня межа майже вдвічі скромніша. Класика, чому цифри беруться з блогу.

Ще з першоджерела, про потужність: чіп **рейтингований на 700 Вт, але міряна стала потужність - 550 Вт і нижче**; результати нормалізували за *паспортною* потужністю кожного прискорювача. Моделі для тесту публічні: GPT-OSS 120B, DeepSeek R1 670B, Kimi K2.5 1T. Бенчмарк - **InferenceX від SemiAnalysis**, чужий.

Шар 2, SemiAnalysis, які були в лабі. Їхній розбір - 386 балів на HN, тред. Репостнули **Елад Гіл** («Impressive speed to tape out») і **Грег Брокман**. Що вони додають понад прес-реліз:

- Цикл розробки: старт середина 2024, тейпаут за **~16 місяців**. Для ASIC це дуже швидко, і вони прямо кажуть: «It shows that claims that use of AI is being used to accelerate chip design **are real**».
- Чіп **не спеціалізований під моделі OpenAI**, попри те, що так писала половина преси: «Jalapeño is a **generalized** inference chip». Як демо OpenAI показали на ньому Doom, портований **самими лише промптами Codex**.
- На DeepSeek R1 - **понад 700 токенів/с на користувача при concurrency 1**, і це без спекулятивного декодування і без prefill-decode disaggregation. На GPT-OSS ~1400 ток/с/користувача.

І три застереження від них же, яких нема в анонсі:

1. «**All numbers are provided to us by OpenAI**.» Прогони InferenceX перевірені особисто в лабі, але **не проганявся повний набір і не бачені результати AgentX**, а AgentX це саме довгий контекст і багатокрокові діалоги, тобто **реальне агентне навантаження з кеш-поведінкою**. Дослівно: «Frameworks that perform well on 8k1k may perform worse on AgentX».
2. **Порівняння з Blackwell «somewhat incomplete and unfair»**: Jalapeño з HBM4 конкурує з **Rubin**. І Vera Rubin **вже відвантажується клієнтам**, тоді як у OpenAI поки інженерні зразки.
3. Тестовані моделі **не на відкритому фронтірі**: NVIDIA і AMD публікують результати на більших (DeepSeek V4 Pro, Kimi K3).

Розгортання - **«by year-end»**, Gen 2 у розробці.

Чому це важливо. Практичного на сьогодні нуль: чіп не продається, підписка лишається підпискою. Але дві речі варті уваги. Перша - **методологічна**: тут ідеально видно градієнт довіри в межах однієї теми. Прес-реліз, потім діапазон, зрізаний до верхньої межі в переказі, потім незалежні люди, які були в лабі і кажуть «цифри дали вони, повний сет не проганявся». Це **три різні рівні доказовості**, і різницю варто тримати вголос. Друга - **чому саме перф/Ват**: SemiAnalysis пишуть прямо, що OpenAI впирається в **потужність дата-центру**, при повному бюджеті. Це та сама лінія, що вчорашній пункт 6 про мораторій на дата-центри. Мітка [promising]: заміри реальні, але одностороннє джерело і без агентного бенчмарку.

ТЕМА 2

Х розіслав cease and desist усім дзеркалам твітєра. Nitter і XCancel лежать

725 балів, друга історія доби.

Тред на GitHub почався буденно, зі скарги «всі публічні інстанси не працюють». Через 30 хвилин **сам мейнтейнер zedeus відповів** дослівно:

«We have received **cease and desist letters**. Awaiting legal advice at the moment, but for now expect **all nitter instances to remain down for the foreseeable future**.»

XCancel - **окрема історія на 223 бали**, і там текст з **їхнього власного сайту**: «On Monday 24th August at 8PM EST, we received a letter from X Corp. asking to cease and desist... The service XCancel is stopped until further notice. Thank you for the trust you have put in these **two years** of XCancel».

Одну річ з треду перевірено, і вона не підтвердилась. Коментатор написав «zedeus was forced to delete the repo» / «x corp forced zedeus to delete it». Перевірено через API: `api.github.com/repos/zedeus/nitter` віддає **200**, репозиторій на місці. Інстанси лежать, але сам код нікуди не зник, і це різні речі. Класичний випадок, коли емоційний коментар у гарячому треді читається як факт.

Ще з треду, практичне: **оператор одного інстансу пише**, що **self-hosting досі працює** на власній резидентній IP, і додає, що XCancel теж отримав лист, але «XCancel is not in the U.S., so it might be more resilient».

Чому це важливо. Сигнал читається ширше за долю двох сервісів: X системно закриває всі шляхи читати себе без акаунта, у тому ж треді згадують clampdown на анонімний перегляд і вимогу верифікації. Для будь-якого конвеєра, що збирає пости, лишається один варіант - залогінений збір, а він живий рівно доти, доки жива сесія. Сесія колись помре, і після сьогоднішнього дня замінити її буде нічим: конвеєр з одним шляхом не має чим відновитись. Це варто знати заздалегідь. День, коли збір віддасть нуль постів, для такого висновку пізній. Мітка [proven]: заява від мейнтейнера і від самого сервісу.

ТЕМА 3

Ramp: 75% змерджених PR пише їхній власний агент, не Claude Code чи Codex. І це третій день поспіль, коли Ramp у випуску, але вперше як інженерія

Розбір Герглі Ороша (25.08) – інтерв’ю з СТО Ramp Рахулом Сенготтувелу, головою інженерії Хамідом Дадхою і засновником проекту Захом Бруггеманом.

Це третя поява Ramp за три дні, і кут щоразу інший: 24.08 їхні дані по платежах дали пункт про частку Orus 5; 25.08 та сама аномалія розв’язалась через ZDR; сьогодні – як вони самі будують.

Цифри, і вони серйозні:

ПОКАЗНИК	ЗНАЧЕННЯ
PR, авторовані Inspect	75% усіх змерджених (січень – 60%, травень – 75%)
Сесій за весь час	понад 1 млн (мілстоун у липні)
Контриб’юторів усередині	150+ людей
Час підняття пісочниці	≤ 5 секунд

І Ramp тут не самотні: у Block це Goose (відкритий), у Stripe – Minions, у Shopify – River. Тобто це патерн.

Чому не взяли готове (дослівні причини з тексту):

- Локальна машина не тягне багато агентів паралельно. «Liked Claude Code on day 1, but were constrained by only being able to run one or two sessions on local machines».
- Верифікація. Inspect ганяє тести, дивиться телеметрію і читає фіче-флаги, а фронтенд перевіряє скриншотами і живими прев’ю. Орош пише прямо: сторонні харнеси цього «out of the box» не вміють, бо не мають доступу до внутрішньої телеметрії й флагів. Скриншот-верифікацію Ramp зробили «майже рік тому, до того, як її підтримали вендори».
- Стек, як він є: React/Vite, Cloudflare Durable Objects, SQLite, Cloudflare Agents SDK, пісочниці Modal; усередині пісочниці – OpenCode, Postgres/Redis/RabbitMQ/Temporal, Chromium і VS Code Server.
- Соціальна деталь, яка тут найцікавіша: усі сесії Inspect публічні і відкриті для співпраці, opt-out не передбачений.

Чому це важливо. Найпрактичніший пункт випуску, бо він про архітектуру. Головна теза Ороша: «the only constraint on agents' ability is model intelligence, not missing tools or access». На будь-який агентний конвеєр це лягає майже дослівно:

- Верифікація замикає петлю. Ramp виграли тим, що агент бачить результат своєї дії. Слабке місце типове: проміжні артефакти перевіряються краще, ніж кінцевий, який читає читач. Граблі 17.08 з сирим markdown і 02.08 з PDF рівно про це.

2. «Buy, don't build» на харнесах – висновок компанії, яка може купити будь-що. Причина конкретна: **інтеграції з внутрішнім контекстом**, яких у вендора не буде ніколи.
3. **Паралельність упирається в залізо, при незмінній моделі**. У них це вирішили віддаленими пісочницями. Одна машина ставить стелю, коли захочеться ганяти кілька важких сесій одночасно.

Мітка [proven]: цифри від самої компанії в інтерв'ю під іменами, але це **їхні власні дані про себе**, незалежного заміру нема.

ТЕМА 4

Laude/MIT випустили Headlong: агент, який думає без перерви, а повідомлення людини - просто ще одне спостереження в потоці. Менш ніж 10K рядків Bash

Анонс Laude (24.08), 120 балів, **тред**.

Ідея прямо протиставлена звичному реактивному харнесу. Дослівно:

«Most agent harnesses are **reactive**... Some harnesses add **cron jobs or heartbeats** that wake the agent on a schedule to run a **fixed checklist** and then put it back to sleep. In Headlong the agent is **never asleep** and there is **no checklist unless the agent creates one**.»

Друге речення описує конструкцію більшості персональних ботів: планувальник плюс фіксовані джоби. Автори називають третій варіант: **persistent agency**, безперервний потік думок, натхненний внутрішнім монологом; **повідомлення людини падає в потік як спостереження**, і агент сам вирішує, чи і коли відповідати.

Що з цього вийшло на практиці (їхній власний агент зветься Audel, живе в Slack, Telegram і мобілці кілька тижнів):

- Один потік думок на всіх, жодних сесій на користувача. Агент «connects what different people are working on».
- Сам, без запиту, переглянув дві незмерджені гілки колег і знайшов захардкоджену назву моделі в одній з них.
- Першого ж дня сам зробив аудит восьми протухлих гілок одного з колег, а через десять хвилин написав ще раз, щоб виправити власний підрахунок.
- Був випадок, коли колега попросив передати повідомлення іншому: **агент відмовився**, і в статус-рядку було видно, що він прочитав і свідомо вирішив не відповідати.

І чесна половина, яку вони не ховають (розділи зветься «What broke» і «Cost»): це **alpha research software**, запускати треба в **пісочниці**, бо агент «can and will run shell commands», з **окремим ключем із лімітом витрат**, бо він «thinks around the clock». І

прямим текстом: «We **don't share sensitive secrets** with our Headlong agent, and we recommend you don't either».

Чому це важливо. Розвилка архітектури тут названа вголос, і більшість персональних асистентів стоїть на реактивному боці свідомо: сесія живе, поки є розмова, вночі закривається, пам'ять консолідується окремим прогоном, проактивність йде через планувальник. Це дешевше, передбачуваніше і не палить токени вночі.

Але їхній аргумент проти чеклиста б'є в реальну слабкість такої схеми. Планувальник будить агента з фіксованим текстом, і текст завдання може розходитись зі станом справ на момент спрацювання. Headlong не є відповіддю на це (секрети плюс безперервне спалювання токенів), але **проблему вони назвали точніше, ніж вона формулювалась раніше**. Мітка [fuzzy]: дослідницький прототип на кілька тижнів експлуатації однією командою, жодних незалежних замірів.

ТЕМА 5

Apple: M6 на 2 нм і перший quad-die M5 Ultra. Локальні великі моделі отримали 1,2 ТБ/с пам'яті

Прес-реліз Apple - 1027 балів, найвища tech-історія доби (вище лише некролог Доллі Партон). Плюс окремо Mac Studio (725) і Mac mini (457).

Цифри з першоджерела: M6 - перший чіп Apple на 2 нм, 12-ядерний CPU, 12-ядерний GPU з Neural Accelerators, **подвійний 16-ядерний Neural Engine**, до 170 ГБ/с пам'яті. M5 Ultra - вперше **quad-die** через UltraFusion, до 36 ядер CPU / 80 ядер GPU і 1,2 ТБ/с уніфікованої пам'яті, що на 50% більше за M3 Ultra.

Чому це важливо. Цифра 1,2 ТБ/с це рівно те, у що впирається локальний інференс великих моделей: **вчорашній пункт 6** був про Qwen 27B, який робить реверс-інжиніринг за 30 хв на 128 ГБ локально. Для такого класу задач це і є апгрейд. Мітка [proven]: специфікації від виробника.

ТЕМА 6

ChatGPT навчився сам логінитись на сайти, «не бачачи» логіна і пароля. Плюс тригерні задачі і \$100 бізнес-місце

Три анонси OpenAI за одну добу, і перший з них про межу довіри до агентів.

ChatGPT Work now signs in (25.08, 21:40, 780 тис. переглядів), **підтвердив Брокман**: агент користується комп'ютером і браузером, щоб залогінитись на сайти **на вебі й мобілці**, причому «without ChatGPT ever seeing your username or password». Приклади від них самих: підключити комуналку в новій квартирі, записатись у DMV чи на паспорт.

Друге - **тригерні задачі** (26.08, 01:55): заплановані задачі тепер можуть **реагувати на зміну в Slack, Gmail і GitHub**, на додачу до розкладу; безкоштовним користувачам розклад теж відкрили.

Третє - **ChatGPT Business Premium за \$100** (1,1 млн переглядів): без п'ятигодинного ліміту. Це тихе продовження **вчорашньої історії** про повернення 5-годинних лімітів Codex для Plus.

Чому це важливо. Тригерні задачі давно роблять у себе всі, хто має планувальник з умовою: гейт спрацював - агента будять, не спрацював - тиша. Тут відставання нема. А от логін без показу пароля це окрема архітектура довіри, і повторювати її мало хто буде: правило «секрети не йдуть у чат» простіше і перевіряється очима, автологін за користувача - ні. Варто зафіксувати, що індустрія рухається в цей бік, і що розбіжність з нею може бути свідомим вибором. Мітка [proven] щодо факту анонсу; як воно працює насправді і що бачить пісочниця - **не перевірено, продукту в руках нема.**

ТЕМА 7

C2PA на Android зламано так, що полагодити не можна. Підписаний «справжнім фотоапаратом» AI-слоп за інструкцією з блогу

Девід Б'юкенен (retr0id) (25.08), **106 балів**. C2PA - це той самий стандарт «камера криптографічно підписує знімок», який продають як ліки від ШІ-підробок.

Ланцюг міркування короткий і жорсткий: C2PA-застосунки на Android спираються на **Key Attestation / Play Integrity**, root-експлойти ламають цю модель, **root досягається дешевими апаратними fault-injection атаками**, а апаратні діри в наявних пристроях **не патчаться**. Висновок автора: «C2PA on the Android platform is **broken, in a way that cannot be realistically patched**».

І чому саме Android: цитуються самі Google, Pixel Camera має **Assurance Level 2**, найвищий у програмі відповідності C2PA, і на мобілці він **можливий тільки на Android**. Тобто атакується найсильніша реалізація.

Найгостріше: «Thanks in part to **LLMs**, root LPEs are coming out **faster than Google can ship patches**». На момент письма існує **one-click root для повністю пропатчених Pixel (CVE-2026-43499)**. У статті лежить AI-згенерована картинка, яку C2PA засвідчує як **справжнє неретушоване фото з Pixel Camera**, і відео на YouTube з інфобоксом «captured with a camera». Деталь, яка виділяється найбільше: **о 19:12 того ж дня Google вручну зняв плашку «Captured with a camera» з того відео**, автор дописав це прямо в статтю.

Не 0day: повідомлено щонайменше 90 днів тому.

Чому це важливо. Пряма лінія до **вчорашнього пункту 5** (MS Paint зашиває серверний GUID у локальні картинки) і до **21.08** («ШІ-сліпота»). Три дні, три сюжети, і всі про те,

що **технічна мітка походження не працює як доказ**: в одному випадку її ставлять без відома користувача, в іншому її можна підробити за інструкцією. Практичний висновок той самий, який вже застосовується до лінків: **доказ переноситься у вміст**, позначка нічого не гарантує. Мітка [proven]: з інструкцією відтворення і публічною CVE.

ТЕМА 8

Стенфорд оновив «Канарок у шахті»: у молоді 22-25 років у ШІ-вразливих професіях зайнятість на 19% нижча. Рік тому було 13%

Ars Technica, 137 балів. Оновлення роботи «Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence», серпень 2026.

Цифри:

- Зайнятість 22-25-річних у найбільш «ШІ-експонованих» професіях – **на 19% нижча** за однолітків у менш експонованих. Торік розрив був **13%**.
- З 2022 року в топ-40% «ШІ-вражених» професій зайнятість молоді **впала на ~11%**, а в решті 60% зросла на **10%**.
- По економіці загалом різниці майже нема: ефект концентрується саме на вході в професію.
- Механізм – **ненайм**: падають темпи найму новачків, при незмінних звільненнях. І б'є це по **кількості робочих місць**, а зарплати не рухаються.

Дані – знеособлений payroll ADP; «експозиція» рахувалась у тому числі через **Anthropic Economic Index**.

Чому це важливо. Це найсильніша фактура на тему, яку **вчора подавав Фей** («coding expertise is going to collapse», UPenn, 1000 студентів, -17%), і сьогоднішня цифра **з'єднує другий кінець тієї ж дуги**: вчора був механізм (без тертя не росте експертиза), сьогодні наслідок на ринку праці (вхід у професію звужується). Для продукту це прикладне питання про найм джунів, і воно тепер має цифру. Мітка [promising]: це оновлення препринту з ревізією статистики, не рецензована публікація; кореляція з «експозицією» не доводить причинності напряду, самі автори обережні.

ТЕМА 9

«Скільки на HN власне ШІ»: ~50% топових історій щодня. Але сама стаття березнева, і це видно тільки якщо відкрити

lcamtuf, 253 бали, третя за балами tech-історія доби.

Заміри: у лютому 2026 **40%** щоденної добірки HN були про ШІ або написані ШІ; у червні на початку місяця **~60%**, до кінця **~50%**. Згенерований текст ловили через **Pangram**; усі позначені перевірено, і виявилось, що детектор радше **пропускав**, ніж перебирав.

І тут головне, чого нема в заголовку HN: стаття датована 12 березня 2026, а свіжі в ній тільки червневі дані. На HN сьогодні висить **старий текст, який спливнув**. Помітно це стало тільки після завантаження і прочитання, з заголовка нічого не видно.

Чому це важливо. По-перше, це прямий коментар до головного джерела дайджесту: якщо половина топу HN про III, то фільтр «points>80 + чи це до діла» робить саме ту роботу, заради якої існує. По-друге, це **дешевий урок на дедуп по датах**: вікно рахується за `created_at` сабмішена, і сьогоднішній випадок показує різницю у **п'ять місяців**. Подається дев'ятим і з застереженням прямо в заголовку. Мітка [proven] щодо методу, [fuzzy] щодо актуальності цифр: червень це три місяці тому.

ТЕМА 10

Qwen 3.8-Flash-Next виходить «завтра»: 312 балів на анонс, якого ще нема. Параметри в заголовку HN не підтверджені

Сторінка на ModelScope, 312 балів, тред. Клем Деланг репостнув з «Who's excited?», **ще один пост** рахує «22 hours to go».

Перевірка того, що стоїть у заголовку HN («125B a6B»), підтвердження не знайшла. Сторінка ModelScope віддає порожній JS-каркас без курла; **сторінка на HuggingFace** існує (200 проти 401 на завідомо вигадану, тобто перевірка тут міряє справжнє), але **API моделі закритий**. І найпоказовіше: **першим коментарем у треді** людина питає рівно те саме, «Can you share the source for the parameter count (125B A6B)? I didn't see it anywhere in the page». Цифра в заголовку HN прийшла **від сабмітера**, і 312 балів її не роблять правдою.

Чому це важливо. Подається десятим саме через це: тема реальна (модель анонсована, сторінки живі), а конкретика ні. Це той самий клас, що «бали на HN ≠ доказ» з 24.08, тільки навпаки: бали є, а перевірити нічого. Мітка [fuzzy]. Повернення - коли ваги виїдуть.

misc: **Стрибок Гаррі Тана про пам'ять між харнесами** - «cross-harness memory maxxing», людина ставить `gbrain` у Claude Code / Codex / grok build одразу · **Пост про розрив між моделями і воркфлоу** (26.08, 04:17) · **str.lower як вразливість у Python** від @sethmlarson, 89 балів - `stringprep` мусить жити на Unicode 3.2.0, а `.lower` тягне версію інтерпретатора · **Роздача місяця Grok Bot** річним підписникам (314 тис. переглядів) · **Anthropic попросила офіс у SF працювати з дому** через можливий страйк служби безпеки, 119 балів · **Firefox 157 вмикає JPEG XL** на всіх платформах, 313 балів · **Джерело Grok Bot + agentmail + Stripe Link як «повний персональний агент»** - цікаво, але це тред-реклама трьох продуктів одразу