AI-ДАЙДЖЕСТ

Математик прилюдно спіймав LLM на брехні

Даніель Літт викрив, що Astra не розв'язала 10 задач, а перефразувала відомі результати за $2000.

неділя, 2 серпня 2026

У ЦЬОМУ ВИПУСКУ

1. Десять відкритих проблем, $2000 і Lean-сертифікат - і чому тут довіра до Даніеля Літта 422 бали HN / 287 коментарів; у Ноама Брауна (@polynoamial) **7 040 000 переглядів і 13 827 лайків**, у Брокмана 682K. За охопленням це перекриває всі теми тижня разом.

Що саме сказано **в першоджерелі** (сторінка OpenAI, не переказ у твіті):

• результати отримані «**внутрішньою версією Astra, наступної великої моделі**»

• проблеми - такі, де «**не було прогресу по головному результату щонайменше десятиліття**, а в більшості випадків значно довше»

• вартість токенів на пошук усіх розв'язків - «**приблизно $2 000 за тарифами Sol API**»

• рукописи готували **люди разом з моделлю**, а потім модель **формалізувала кожен аргумент у Lean-сертифікат** - доведення машинно перевірене

Список конкретний: пакування сфер у високих вимірностях, межі для бінарних кодів, **існування несофічних груп**, **спростування гіпотези жорсткості Конна**, нижні оцінки для перманента, квантова паралельна репетиція, найближчий вектор у ґратках (це

прямо під постквантову криптографію), гіпотеза Ерхарта про об'єм, мультикольорові числа Рамсея (**проблема Ердеша 183**) і ще дві проблеми Ердеша - **146 і 180**.

Тепер те, заради чого це пункт 1. Заяви лабораторії лишаються маркетингом, поки їх не підтвердив хтось, кому нема за що платити. Ethan Mollick чекав на «вердикт одного з найтверезіших і найбільш обізнаних в ШІ математиків» - це **Даніель Літт (@littmath)**, справжній професор математики. Його вердикт: «**It's a big deal**» (102.8K переглядів). А через три години він написав те, що важить більше за будь-який бенчмарк:

> «***Визнаю поразку в цьому парі.*** *Строго кажучи, воно не розв'язалось - схоже, стаття з теорії чисел рівня Annals ще не з'явилась - але ясно, що* ***очікування щодо здібностей, потрібних для такої статті, були помилковими****, і це лише питання часу.» І далі: «Варто було обрати кращий проксі для того типу досліджень, який цікавить найбільше. Можливо, це урок і для професії.»*

Він же окреслив межі своєї компетенції: «більшість з цього далеко за межами компетенції (найбільше відомо про пакування сфер, і воно дуже подобається, але **лише як зацікавленому аматору**)».

Чому це важливо. Тут збіглись **три незалежні речі**: заява з конкретною ціною, **машинна перевірка** (Lean - доведення або компілюється, або ні), і **публічна капітуляція скептика з іменем**. Коли з цих трьох є лише перша, це прес-реліз. Сьогодні є всі три, тому подається це як подія. [proven - щодо факту публікації і Lean-формалізації; **fuzzy** - щодо того, що з цього справді переживе рецензування математичною спільнотою; сам Літт каже, що ключова планка ще не взята]

Ноам Браун · вердикт Літта · визнання парі · HN-тред

2. Cursor прибрав ціни з інтерфейсу - а це рівно та метрика, за якою харнеси мірялись два дні поспіль 317 балів HN, 143 коментарі. Тред на форумі самого Cursor.

Факти: **31.07** Cursor прибрав з usage-сторінки і з CSV-експорту **доларову вартість** для self-serve планів (індивідуальних і Teams). Зникли метрика **Spend**, колонка **Cost** з ціною за запит, і долари в експорті - тепер там `$0.00` або порожньо. Офіційне пояснення від співробітника (Kevin Neilson): це «**свідоме рішення**», щоб зменшити плутанину - «те, що покрито планом, показує кількість токенів і позначено *Included*, бо за це нічого не стягується».

Користувачі не купили. Дослівно з треду: «Це вікно використовувалось, щоб пильно тримати під контролем щоденні витрати»; «Та сума в доларах сприймалась як цінність (економія), яку отримую»; адмін Teams з витратами **$30K**: «Як тепер відстежувати витрати по кожному юзеру, як раніше?».

Чому це важливо. У четвер прозвучав бенчмарк Composio - **3.7× різниця в ціні між харнесами** при однаковому результаті; вчора Малєєв дав корпоративну рамку токенмаксингу. Сьогодні один з найбільших харнесів **прибирає саме той лічильник**, за яким така перевірка робиться. Висновок вузький: метрику вартості треба міряти у себе, бо чужий UI прибирає її рішенням продакт-менеджера. Власний замір (скільки

кілобайт за скільки секунд) - і є правильна форма власності на метрику. [proven - це офіційна відповідь Cursor у їхньому ж треді]

форум Cursor · HN

3. Карпаті: «Володар перснів» за 1M токенів (~$10), 5500 рядків коду і 2 години - і чесний висновок про слабкість LLM 283K переглядів за годину, 4 947 лайків - свіжак, прилетів о 06:00 за Києвом.

Експеримент дослівно: Opus 5 отримав **перший абзац «Володаря перснів»**, бюджет **1M токенів (~$10)** і завдання зробити три.js-рендер. «Opus пішов на **~2 години** і написав **5500 рядків коду**, які (процедурно) відрендерили історію. Виглядає трохи кострубато, але весело.»

Дві думки, які варті більше за сам демо-ролик:

• «**Ніхто при здоровому глузді ніколи не витратив би час** на щось настільки кастомне, але в LLM є вся витривалість і терпіння світу - тож перехід від *"ніхто б цього не робив"* до *"та чому б ні, це майже безкоштовно"*.»

• І одразу - де воно ламається: «домен світів/ігор **оголює слабкість LLM: вони не можуть легко перевірити власну роботу**, бо не вміють ефективно й нативно сприймати відео чи грати в те, що зробили».

Чому це важливо. Друга думка описує загальний клас проблеми. Типовий випадок: скрипт підвисає, причина шукається в мережі кілька хвилин, бо **не видно** процесу на 99% CPU - того, що людина побачила б одразу. Модель без правильного каналу сприйняття впевнено робить не те. Практичний висновок: там, де результат помічено «зроблено», але **перевірити його тим самим каналом, яким його бачить людина, неможливо** - там треба ставити явну перевірку. Перша думка теж робоча: дешевшає цілий клас одноразових саморобок, які раніше не мали сенсу. [proven - самозвіт з конкретними числами]

Карпаті

4. QM за добу: 2 475 → 5 234 зірок, і всередині YC ним користуються щодня Продовження вчорашнього пункту 2 - цифра подвоїлась.

Учора цифра була **2 475 зірок**; зараз GitHub API віддає **5 234 зірки і 538 форків** (репо створено 29.07, останній пуш 01.08). Тобто **+2 759 за добу**.

Свіже за добу - свідчення зсередини. Jon Xu (@xuster): «Користується QM **щодня**, і воно стало невід'ємною частиною роботи». І деталь, якої вчора не було (Anna Zhang, @anna_y_zhang): «Найкрутіше в qm - **культура, з якої воно виросло**: YC колись розповідали, що їхні **агентні сесії транслюються всередині компанії**, тож люди вчаться, спостерігаючи, як інші працюють з агентами. **Історія агентів організації як спільний навчальний матеріал.**»

Чому це важливо. Останнє - дешева ідея під будь-який агентний стек, де вже зберігаються логи сесій. Зазвичай з таких логів витягують знання. *Спосіб роботи* при

цьому губиться. Найдешевший крок: раз на тиждень діставати з журналу 2-3 місця, де стався хибний шлях, і класти в базу знань окремим рядком «як не треба». Коду міняти не треба, міняється те, що шукає консолідація. [proven - цифри з GitHub API, звірені]

репо QM · Jon Xu · Anna Zhang

5. «AI не робить робочих продуктів - це досі робота людини» 250 балів HN, **267 коментарів** - одна з найгарячіших дискусій доби.

Теза автора: ШІ різко скоротив шлях до **першої робочої версії**, але відстань до продакшену лишилась та сама - системний дизайн, масштабування, обробка помилок, безпека стоять рівно там, де були. Найсильніші рядки: «**Складні задачі в софті ніколи не були про синтаксис. Вони були про судження.**» І далі: «У моделей нема судження. У них є зіставлення патернів **із завзятістю видати код, який, на їхню думку, відповідає наміру**.» Висновок: «Спочатку варто вивчити фундамент. Потім нові інструменти. **Саме в такому порядку.**»

Чому це важливо. Це антидот до пункту 3, тому вони поруч. Карпаті показує, що модель може дві години пиляти 5500 рядків; цей автор нагадує, що **дистанція від того до працюючої штуки не змінилась**. Типовий приклад - баг продуктивності, де фікс займає один рядок, а вся цінність у тому, щоб зміряти пороги (на 4 КБ ок, на 8 КБ вішається) і зрозуміти, що складність квадратична. Модель, яка «завзято видає код, що відповідає наміру», написала б обхід і поїхала далі. [promising - це аргументований есей практика, не дослідження]

стаття · HN

6. Моллік: ШІ розмиває межі між професіями - і це вже друге незалежне підтвердження 136К переглядів. Формально це продовження вчорашнього misc, але сьогодні воно набрало вагу і має **два незалежні джерела**, тому винесено в окремий айтем.

Моллік: «Одним з головних результатів дослідження в **Procter & Gamble** було те, що **ШІ розмив межі між роботами**. Тепер **OpenAI отримала схожий результат**. Організаційні межі стають проникними, стіни тоншають.» І його ж жорсткіше формулювання: «Це **не опційно**, все стає хаотичним, і ігнорування цієї зміни її не прибере.»

Як це виглядає на практиці (його ж приклад): «Якщо маркетинговий проєкт вимагав трохи коду, або фінансової моделі в Excel, або розв'язку задачі про пакування сфер у високих вимірностях - раніше треба було кликати інженерний відділ, фінансових аналітиків чи теоретиків з матемдепартаменту. **Тепер це просто питання до ШІ.**» Поруч Ленні Рачицький, незалежно: «**Усі стають інженерами і маркетологами на пів ставки.**»

Чому це важливо. Робота вже так і будується: одна людина лізе і в креативи, і в дані, і в код. Зворотний бік у Моллікових же словах: межі стають проникними **в обидва боки**, тож захист від розмивання - це чітко зафіксовані критерії якості там, де експертизи

немає. [promising - рецензоване дослідження P&G плюс незалежний звіт OpenAI, але це соціологія організацій, не замір]

Моллік · стаття P&G · Ленні

7. Моллік окремо: приріст можливостей стає неможливо «відчути» 162K переглядів - і це найтверезіша думка доби, тому вона окремим пунктом.

Дослівно: «Для майже кожної людини на планеті це за межами розуміння. Можна лише довіряти математикам-експертам, коли вони кажуть, чи це вражає. **Це починає відбуватись у багатьох галузях, і через це приріст можливостей стає дедалі важче "відчути".**»

Він же поруч: «Думка й далі та сама: **брак верифіковних відповідей** у багатьох галузях - реальна проблема для LLM, але не така велика, як її часом подають». І чесний контрприклад від нього ж: «**І все одно вони погані у хорошій довгій художній прозі.**»

Чому це важливо. Якщо приріст не відчувається на дотик, єдина заміна відчуттю - **зовнішня перевірка**: Lean-сертифікат у пункті 1, публічна капітуляція скептика, незалежний практик замість заяви лабораторії. За цим же принципом варто читати будь-яку новину про нові можливості моделей. [proven - це його спостереження, не претензія на замір]

Моллік

8. Леві: розрив між «ШІ в побуті» і «ШІ в глибоких доменах» почне розходитись 72K переглядів. Аарон Леві (Box) - той самий голос, що вчора дав формулювання про харнес як головну змінну.

Його теза: «Буде видно **дедалі більшу розбіжність** між тим, що ШІ робить в особистому житті й щоденній продуктивності, і тим, що він може в дуже глибоких доменах - математика, наука, право, код. До певного порогу зростання можливостей відчувалось **рівномірно в усіх доменах**...» - далі саме той злам, який сьогодні показала історія з Astra.

Чому це важливо. Практичний висновок для очікувань: «розв'язав десять відкритих проблем» і «не забув про дрібну побутову справу» - це **різні криві**, і перша не тягне другу. Тому не варто дивуватись, коли модель, яка валить теоретичну інформатику, лажає на рутині, і не варто робити з побутової лажі висновку про стелю. [fuzzy - це прогноз практика, не дані]

Леві

9. Cursor не єдиний: 456 балів за розбір, як Google закопав RSS - і це річниця Найвищий бал доби на HN (456 / 159 коментарів). Одразу зазначу: **стаття 2023 року**, вона просто знову випливла, тому це не сьогоднішня подія.

Хронологія з тексту: Chrome прибрав вбудовану кнопку підписки на RSS ще на початку 2000-х без пояснень; **2007** - Google купує FeedBurner, **жовтень 2012** - вимикає його API, **липень 2022** - прибирає більшість сервісів разом з email-підписками; **2013** - закриття

Google Reader (інженер зсередини: «весь час, що я був на проєкті, різні люди намагались його вбити»); **грудень 2017** - Google News повністю прибирає підтримку RSS. Автор називає це «**Embrace, Extend, Extinguish**».

І поряд, тієї ж доби: **«Чи Google покинув Google News?»** (261 бал) і **«Каталог людей, які люблять RSS»** (169 балів). Тобто це настрій дня.

Чому це важливо. Для будь-якого конвеєра, що збирає контент автоматично, RSS лишається несучою конструкцією: публічний фід не ламається від протухлої сесії і не вимагає логіну, на відміну від API соцмереж. Це рівно та стійкість, за яку HN сьогодні голосує 456 балами. [proven - фактологія статті; дата 2023, не сьогоднішня подія]

стаття (2023) · HN · Google News

10. Перша компанія YC зі 100K зірок на GitHub під час батчу - цифра перевірена
Джаред Фрідман (YC): «**Graphify Labs - перша компанія YC, яка взяла 100K зірок на GitHub під час батчу.**» 67K переглядів.

Звірка по API показала, що саме такі заяви найлегше роздути: **100 387 зірок**, репо створено **03.04.2026**. Тобто 100K за чотири місяці - цифра справжня. Що воно робить, з опису репо: перетворює будь-яку кодову базу **разом з доками, SQL-схемами, конфігами і PDF** на щось, до чого можна робити запити.

Цікава деталь поруч у видачі: третій за зірками репозиторій на запит «graphify» - це `claude-code-memory-setup` з обіцянкою «**до 71.5× менше токенів за сесію** в Claude Code через Obsidian + Graphify» (906 зірок).

Чому це важливо. Питання не в самому інструменті, а в замірі, який під ним лежить. Якщо пам'ять агента вже структурована і шукається пошуком, то скільки токенів з'їдає її ін'єкція в кожну сесію - зазвичай ніхто не міряв. Дешева перевірка, яку варто зробити раніше, ніж дивитись на чужі інструменти. [proven - щодо 100 387 зірок, звірено через GitHub API; **fuzzy** - щодо «71.5×», це маркетинг репозиторію, не самостійний замір]

Фрідман · репо Graphify

Misc • Seedance 2.5 від ByteDance (225 балів HN) - відеомодель з «one-take creation» і гнучкими референсами. У коментарях головна дискусія про призначення: «єдиний реальний ужиток цих моделей - дезінформація і спам» проти «будь-який режисер бачить тут потенціал зробити сцени, які інакше неможливі» seed.bytedance.com · HN

• **@levelsio виклав кореляції зі свого трекера**: підключив калорійний застосунок до датчика повітря Airthings і WHOOP. Найкраще для сну - **холодна і суха спальня**; поділ: cool ≤18.1°C, middle 18.1 - 19.4°C. Топ позитивних факторів: мало калорій **+6.5%**, низький % жиру в їжі **+6.3%**, суха спальня **+5.2%**, ванна **+3.0%**. Сам автор пише «не всі значущі, треба більше даних» - це n=1 без контролю, тому як гіпотеза пост · про спальню

• **Той же автор про креатин і облисіння**: «Облисіння на креатині - спростовано. Лисієш через генетику, і 80% чоловіків це зачіпає». Аргумент: вік, коли чоловіки починають ходити в зал і пити креатин, **збігається** з віком, коли починається випадіння - зв'язку

нема. Окремо проходиться по фінастериду/дутастериду як «синтетичних 4-азастероїдах». Це його думка, не консенсус 1 · 2

• **Креатин і депресія** (184K переглядів у нього ж): огляд п'ятьох клінічних досліджень у Canadian Journal of Psychiatry - креатин **додавали** до антидепресантів або психотерапії. Подано як цитата з чужого допису, першоджерело не залінковане - береться як [fuzzy] пост

• **Брокман про ChatGPT у Slack** (585K переглядів, 7 091 лайк - топ доби після математики): «В OpenAI багато людей підключають свій ChatGPT до Slack. Людям **дуже не подобається**, коли ChatGPT колеги звертається до них по допомогу - навіть коли вони були б цілком раді зробити ту саму роботу, якби попросив сам колега.» Найточніша ілюстрація того, чому агент має писати від свого імені і не вдавати людину пост

• **Cursor не єдиний, кого чистять**: у тому ж HN-треді практики масово описують відхід з Cursor на Claude Code / Codex / Zed, і найчастіший аргумент - «Cursor як IDE без LLM-фіч це просто гірший VSCode»

• **Масад про 8B-модель у шахах**: Qwen-chess б'є фронтир-моделі і Stockfish level 0 на **~1500 Elo**, витрачаючи **1-2 секунди на хід проти 30**. «Кайфово бачити, як 8b модель робить GPT 5.6 з високим reasoning» пост

• **Мкей Ріглі під постом Карпаті**: «Хтось, будь ласка, **профінансуйте цілу трилогію в такому вигляді як бенчмарк**» пост

• **Коомен (YC)**: частка заявок у YC зі словом «wedge» зросла з **<1% навесні 2025 до >20% влітку 2026**. Гарний індикатор, як швидко жаргон їсть мислення пост

• **NetBSD 11.0** вийшов (237 балів) HN · **Go 1.27 Interactive Tour** (86 балів) HN

• **«Пайплайн розробки - це продакшн-система»** (161 бал): теза, що зламаний CI треба лікувати як аварію на проді. У коментарях знайшлось точне: «бурчали, як тупо економити на pre-prod... **бо pre-prod і був продом**» HN

• **Ґаррі Тан про зсув 2026-го**: «Найцікавіший vibe shift 2026 - **OpenAI реально виглядає як відкрита платформа**: інтелект як комунальна послуга проти сигналу, що оптимально інтегруватись вертикально на весь стек» пост

• **Ленні про системне мислення** - та сама теза, що вчора, з новим підтвердженням: спливло на **п'ятьох останніх подкаст-інтерв'ю поспіль** пост

• **Флінт від Microsoft** - мова візуалізації «для епохи ШІ» (257 балів). У коментарях спільнота дружно відповіла, що **ggplot / Grammar of Graphics вже вирішили цю задачу**, і «for the AI era» - просто обов'язковий баззворд HN