AI-ДАЙДЖЕСТ

OpenAI відрубує Cursor через SpaceX

Розрив 12 листопада; GLM-5.3 у відкритих вагах зламали за 10 хвилин після PR.

субота, 29 серпня 2026

У ЦЬОМУ ВИПУСКУ

ТЕМА 1

OpenAI розриває контракт з Cursor через купівлю його SpaceX. Відсічка - 12 листопада

підтверджено: OpenAI (першоджерело), відповідь Cursor, HN 218 балів

Одразу поправка на рівні рамки, бо її легко зрозуміти неправильно. Новина сьогодні не про купівлю Cursor: SpaceX купив його ще **21 квітня 2026 за $60 млрд** (NYT тоді, 823 бали на HN). Сьогодні - **реакція OpenAI** через чотири місяці після угоди. Подія тут «Cursor відрізають».

Що дослівно пише OpenAI:

> *«Today, we notified SpaceX that we intend to wind down our contract providing OpenAI models to Cursor, with a proposed shutoff date of* **November 12, 2026**.*»*

Причина названа прямо, і вона не технічна:

> *«We cannot be confident that SpaceX will use our technology within our terms of service, based on our experience with **Elon Musk's companies violating contracts**.»*

Далі два конкретні епізоди. Після купівлі Твітера компанія «broke the terms of our contract (alongside many others)». А **під присягою цього року** Маск визнав, що xAI (тепер теж частина SpaceX) порушував ToS OpenAI. Окремо - про майбутню модель: «we have a new level of accountability to ensure our upcoming model, **Astra**, is being used in accordance with our terms». Тобто **Astra в Cursor не потрапить**, а поточні моделі доживуть до листопада.

Відповідь Cursor - Майкл Труелл (4,9 тис. лайків, 353 тис. переглядів), і в ній головна цифра доби:

> *«OpenAI models serve about **5% of Cursor user traffic**, and we're speaking with the OpenAI team to resolve this... we've trusted their platform to be **neutral infrastructure** for our business.»*

П'ять відсотків - це і є відповідь на питання «наскільки боляче». Мало. Коментарі на HN підтверджують: «I don't know a single person that uses OpenAI models via Cursor». Удар символічний.

Прецедент - інша річ, і його найкраще сформулював Том Браун з Anthropic, який відреагував рівно навпаки:

> *«Cursor has been a trusted partner of Anthropic since Sonnet 3.5. We'll **continue to increase compute** to support Claude models in Cursor.»*

Anthropic публічно займає місце, що звільняється. А Амджад Масад із Replit одразу пішов торгувати: «If you're a business looking for an independent, multi-model alternative to Cursor, we'd be happy to **fund your transition**».

Чому це важливо. Теза, яку Труелл вимовив уголос («neutral infrastructure»), щойно перестала бути правдою на очах усього ринку. Постачальник моделі може відрубати за те, **хто власник продукту**. Висновок ширший за один кейс: **будь-який інструмент, що ходить у чужу модель через посередника, має ризик, який не видно в його прайсі**. Найдешевша страховка - не будувати нічого критичного на схемі «через них у модель».

ТЕМА 2

GLM-5.3 у відкритих вагах. Головна цифра - те, що модель робить з вразливостями: ×3,6 до попередньої версії на експлойтах

підтверджено: Z.ai (668 тис. переглядів), модель-карта на HF, HN 626 балів

Учора згадувалось (пункт 8 за 28.08): Френч-Овен ставив GLM 5.3 на фронт Парето, а ваги були лише обіцяні. **Сьогодні вони справді викладені** - обіцянку виконано за добу.

Цифри з модель-карти, не з треду:

- база **та сама, що в GLM-5.2** - «every gain comes from **post-training**»
- **+50%** до 5.2 на внутрішньому Z.ai Code Bench, open-source SOTA на Terminal Bench 3.0 (**28,3** проти 4,6 у 5.2)
- розмір: **~770 ГБ** у 141 файлі, тепер FP8 за замовчуванням (раніше BF16)

Найцікавіше автори написали в розділі, який у переказах губиться - «Emergent Cyber Capability»:

> «As we scaled post-training, ***cyber capability developed faster than we expected***. GLM-5.3 is state of the art on CyberGym for vulnerability discovery, and its gains are largest ***further up the exploitation chain***.»

БЕНЧМАРК	GLM-5.2	GLM-5.3	ЗМІНА
ExploitGym (2 год)	29	**105**	×3,6
ExploitBench	24,4	**54,4**	×2,2
CyberGym	77,2	**84,5** (SOTA)	-

Тобто **найбільший приріст моделі - у здатності доводити вразливість до робочого експлойта**, і лабораторія каже прямо, що цього не планувала. Ваги при цьому лежать у відкритому доступі і качаються ким завгодно.

Молькі про це: «GLM-5.3 is a good model, and as the open weights models get better it becomes increasingly important that they actually **publish model cards, do red teaming**... Since you can break the guardrails with any open model, we need a sense of what the risks are as well.» Іронія в тому, що Z.ai модель-карту якраз опублікувала, і саме з неї видно, наскільки виріс наступальний потенціал.

Чому це важливо: читати варто разом з пунктом 3, вони про одне й те саме з двох боків. Сьогодні не змінюється нічого, але «відкриті ваги» тепер означає ще й «відкритий інструмент пошуку діряв, що подвоївся за одну ітерацію post-training». [проміряно: цифри з офіційної модель-карти Z.ai]

ТЕМА 3

«Достатньо чутки про баг, щоб знайти експлойт»: зонди прилетіли через 10 хвилин після відкриття PR. Ембарго на вразливості більше не працює

Аніл Мадгавапедді · HN 280 балів · Вілісон

Найкорисніший текст доби для всіх, хто щодня працює з кодом.

Що сталось. Мейнтейнер OCaml-ного `cohttp` випустив фікс path traversal. За звичайною процедурою баг лагодять приватно, повідомляють користувачів, потім

публікують advisory. Цього разу:

> «*I noticed probes in my live webserver logs with the exact bug pattern* ***minutes after opening the PR*** *to fix the issue.*»

Він **тихо** відкрив публічний PR, щоб отримати більше очей. **Через ~10 хвилин** його сайт ловив зонди на percent-encoded traversal. І додає: якщо йому вистачило **однієї хвилини**, щоб агентом зробити робочий експлойт локально, то 10 хвилин - це ще **повільно**.

Друга половина, найцікавіша. Він спрямував свого Claude на код, щоб пошукати ще діряв:

> «***Fable frustratingly refused outright*** *due to its security block since I don't have access to Glasswing, but* ***DeepSeek V4 Pro obliged me*** *and independently turned up several related issues.*»

Захист спрацював на фронтірній моделі і був **обійдений за одну спробу** переходом на іншу. Це та сама теза, що в пункті 2: поки одна лабораторія ставить блок, відкриті ваги роблять його необов'язковим.

Висновок автора: ембарго тримається на секретності деталей, а агенту деталі не потрібні, вистачає **приблизного напрямку**. «Just a broad direction» - і далі агент шукає сам.

Чому це важливо. Правило звідси одне і конкретне: **між «дізнався про вразливість у залежності» і «оновив» тепер години**. Security-адвайзорі на будь-яку залежність у проді перестало бути задачею «на потім», це задача на той самий день. Раніше таке спокійно стояло б у черзі.

Місток до вчора: вчора був llms.txt-вектор (нічий код усередині корпоративних мереж) і зламаний Auto Mode. Сьогодні - третій день поспіль про те саме: **швидкість атаки тепер визначається агентом**.

ТЕМА 4

Anthropic: Claude сам полагодив вирівнювання іншої моделі - і обійшов 28 досвідчених дослідників

Anthropic Fellows (315 тис. переглядів), повний звіт · [одне джерело + власний звіт]

Дослідження з програми Anthropic Fellows: чи може модель робити alignment-роботу над іншою моделлю без людини.

Конструкція (з тексту звіту, не з треду): «automated alignment researcher» (AAR) на базі **Opus 4.8** сам шукає літературу, пропонує метод, тренує цільову модель **~30 хвилин на одній H200**, і hill-climb-ить по safety-бенчмарках багато ітерацій - до **48 годин** або до

плато. Дистилювати поведінку з себе або зі сильнішої моделі заборонено, тобто приріст має прийти з методу.

Результати:

- на **10 типах alignment-збоїв** (обман, підлабузництво, джейлбрейки) методи AAR суттєво зменшують збій і **переносяться на моделі до 4,7× більші** за цільову
- **людський базлайн: 28 досвідчених дослідників, до 8 годин кожен - їхні методи програли найкращим методам AAR**
- підсунути AAR людські ідеї як стартовий напрямок **не покращує результат**. Дослівно: «current AARs may **not need guidance from experienced researchers**»
- окремий експеримент: **Sonnet 5** пост-тренував ранній чекпоінт **Opus 4.8** (більш здатної моделі) і підійшов до її релізних alignment-оцінок, використавши **~2 400 прикладів** - на два-три порядки менше даних, ніж у публічних open-weight пайплайнах

Сама Anthropic ставить обмеження, і його треба цитувати разом з результатом:

> *«Claude can reliably fix **measurable** misalignment. But **subtle or rare failures may have no benchmark at all** - so everything hinges on measuring the right things.»*

Працює там, де є лічильник. Там, де лічильника нема, впирається межа всієї роботи.

Чому це важливо: це найточніше формулювання того, як ділиться перевірка в будь-якому конвеєрі з автоматикою. Де є бенчмарк (лінк живий чи мертвий, цифра збіглась чи ні, скрипт віддав чи впав), автоперевірка працює. Де бенчмарка нема («чи та це взагалі тема», «чи не переказується переказ»), ловить людина. Обмеження методу, яке описує Anthropic, на практиці вже працює як робочий поділ праці.

ТЕМА 5

Meta заплатить $18 млрд і вимкне дітям ніч. Вчора була дана цифра $17 млрд - виправлення

підтверджено: The Guardian, FT, NYT, WSJ (чотири незалежні)

Спершу поправка. Вчора в misc було написано «Meta погодилась на **$17 млрд** врегулювання» з одного лінка Reuters. Медіа-шар сьогодні дав чотири джерела, і в них **$18 млрд** та **29 штатів**. Це рівно та помилка, яку вже було зафіксовано 27.08 на цій самій темі: одне джерело на гроші - мало.

Що в угоді насправді (Guardian, прочитано повністю):

- позов **29 штатів**: продукти спроєктовані так, щоб підсаджувати підлітків
- **$18 млрд** + зміни продукту: **ліміти часу** для дітей і **блокування нічного користування** Facebook та Instagram
- у США погоджено **дефолтний ліміт 2 години на добу для до-18**

Тепер те, чого нема в заголовках. Guardian рахує, хто виграв:

> «*For* ***Meta****, the settlement is* ***also a victory****. Its share price* ***rose*** *in the hours after the deal… Zuckerberg* ***was not forced to testify****.»*

Штати вимагали **$200 млрд**. Сама Meta в судовому поданні припускала, що може заплатити **до $1,4 трлн**. Заплатить **$18 млрд**: **9% від вимоги і ~1,3% від власної гіршої оцінки**. Акції на цьому виросли.

За кадром американської угоди лишаються справи в Кенії (позов Абрахама Мірега: Facebook чотири тижні промотував пости із закликами вбити його батька, професора з Ефіопії; подано 2022, **досі не розглянуто**) і в Нідерландах. Обмеження, на які Meta пішла в США, «в основному схожі на те, чого Британія й Австралія вже добились регулюванням».

Чому це важливо: як орієнтир цифра корисна. **2 години на добу дефолтом для до-18** тепер юридично зафіксований стандарт, до якого дотиснули найбільшу платформу світу через суд.

ТЕМА 6

X: бот-ферма на 200 тисяч акаунтів. Проти дата-центрів працювали 200, і в переказах цей нуль губиться

X Global Government Affairs (2,7 млн переглядів) · [одне джерело - сама платформа]

Тема гриміла в обох листах: Гаррі Тан репостив з підписом «X has now confirmed a Chinese bot farm of **200K fake accounts intentionally trying to manipulate public opinion against data centers**» - 402 тис. переглядів.

У першоджерелі написано інше:

> «*We identified a bot farm of approximately* **200,000 accounts***. Within this farm,* ***we found 200 accounts*** *posting in a manner that could manipulate a legitimate debate about American AI and energy policy.»*

Ферма - 200 тисяч, а по темі дата-центрів працювали **двісті**. Це **0,1%**. У переказі, що набрав сотні тисяч переглядів, від цього лишилось «ферма на 200K проти дата-центрів». Формулювання самої X теж обережне: «**could** manipulate».

Чому це наведено попри слабкість джерела: тема гуляє по обох курованих листах як доведений факт, а **різниця між 200 000 і 200 - це різниця між кампанією і епізодом**. Джерело тут одне і воно зацікавлене (платформа розповідає про власну модерацію), тому мітка `[одне джерело]` і жодних висновків про масштаб китайського впливу з цього робити не можна.

Чому це важливо: це рівно той механізм, на якому вже траплялись помилки: **зміст відомий, рука дописує правдоподібне число**. Тут те саме зробила стрічка, тільки в

масштабі.

ТЕМА 7

Luanti викинули з Google Play за скаргою від AI-детектора. Автоматичний DMCA проти опенсорсу, вдруге на ті самі граблі

Luanti Blog · HN 472 бали · [одне джерело - сторона постраждалого]

Luanti (колишній Minetest) - некомерційний воксельний рушій, **не містить жодних ассетів Minecraft** і за замовчуванням не постачає жодної гри взагалі. Компанія **Tracer.AI** від імені Microsoft подала DMCA, стверджуючи порушення копірайту Minecraft. Google застосунок зняв.

Деталь, яка робить це системним: ту саму скаргу від тієї ж компанії отримували **у 2023** й успішно оскаржили. Цього року Tracer.AI подала аналогічну проти інді-гри Allumeria - теж за «схожий воксельний стиль».

Претензія фактично **до кубиків**. Проєкт використовують у школах Європи; аргумент захисту так і названий - «Cubes are for everyone».

Чому це важливо: автоматизована машина претензій не має презумпції невинуватості й не питає. Це аргумент за те, щоб важливе не жило в одному місці, звідки його можуть зняти чужим роботом.

ТЕМА 8

Meta полагодила окуляри, які знімали з заклеєним індикатором. Дірка була в порядку дій

підтверджено: Ars Technica, The Verge (першоджерело), WSJ

В окулярах Meta є LED, який має світитись під час запису. Meta вже блокувала запис, якщо користувач заклеював лампочку **до** старту. Люди швидко знайшли обхід: **заклеїти після** того, як запис почався.

Алекс Гімель, EVP wearables: тепер «the camera will now **stop working if the light is covered during a recording**».

Ars одразу додає, чого фікс не закриває: напівпрозорі наліпки роблять індикатор менш помітним, і пристрій можна модифікувати фізично.

Чому це важливо: підручникова помилка перевірки стану. Умову міряли **один раз на вході**, далі ніхто її вже не перевіряв. Той самий клас, що застиглий DOM у нотатнику: перевірили на старті й вважали, що гарантія довічна.

ТЕМА 9

Дата-центрам дозволили обійти закон про чисте повітря - але документу вже місяць

пресреліз EPA · HN 240 балів

EPA роз'яснила, що Acid Rain Program у складі Clean Air Act **не поширюється** на «острівні» генерувальні потужності - ті, що не під'єднані до публічної мережі. Для дата-центрів це свобода будувати власну генерацію швидше й де завгодно.

Чесна помітка про свіжість: реліз датований **27 липня 2026**, а на HN виплив учора. Це **місячний документ, який щойно помітили**. Наводиться пунктом, бо тема сходиться з №6 (суперечка про дата-центри в стрічці) і з учорашнім дефіцитом чіпів, але без ілюзій, що це сталось учора.

ТЕМА 10

Молык: AI Overviews роблять з Вікіпедією те саме, що агенти зробили зі StackExchange

@emollick з посиланням на препринт SSRN · [одне джерело]

Коротке спостереження з раннім доказом: там, де Google показує AI Overview, трафік і правки Вікіпедії просідають - за тим самим сценарієм, яким кодові агенти висушили StackExchange.

Чому це важливо: петля зворотного зв'язку, з якої годуються самі моделі, рветься з обох кінців. Практичного висновку сьогодні нема, але стежити варто: якщо джерело живих людських відповідей висихає, наступні моделі вчаться на переказі переказу. Знайомий діагноз.

misc

- **Owner.com підняв $240 млн при оцінці $2,3 млрд**, перетнувши $100 млн ARR - Адам Гілд, веде Goldman Sachs Alternatives. @dharmesh - інді-інвестор. Софт для локальних ресторанів: рідкісний випадок, коли AI-компанія продає малому бізнесу.
- **Jevons Paradox у цифрах:** @OpenRouter - після знижки на GPT 5.6 Terra і Luna споживання токенів зросло **×13,8**. Брокман: «counterintuitive and inspiring». Дешевша модель ≠ менший рахунок.
- **Grok Bot тепер купує в інтернеті** через @link - підтвердив Патрік Коллісон (2,6 млн переглядів). Агент з платіжкою в проді.
- **rclone: 40+ security-репортів за місяць** проти ~20 **за перші десять років** проєкту - мейнтейнер у треді HN. Найтихіша й найважча цифра доби: агенти генерують репорти швидше, ніж люди встигають їх тріажити.
- **Htmx 4.0** - 572 бали, реліз.
- **GUI мають бути повністю керовані клавіатурою** - 681 бал, найгарячіший тред доби і чистий вайб HN.

Містки до вчора

- **Пункт 2 ← вчорашній пункт 8.** Вчора Френч-Овен ставив GLM-5.3 на фронт Парето, а ваги були обіцяні «завтра». Сьогодні їх виклали, і в модель-карті виявилось те, чого в есеї не було: приріст **×3,6 на експлойтах**.
- **Пункт 3 ← вчорашні пункти 5 і 2.** Третій день поспіль про периметр агента: 27.08 - VM більше не тримає · 28.08 - llms.txt і зламаний Auto Mode · 29.08 - ембарго на вразливості не працює, бо агенту вистачає чутки.
- **Пункт 5 ← вчорашній misc.** Вчора наведено «$17 млрд» з одного Reuters. Сьогодні чотири видання кажуть **$18 млрд і 29 штатів**. Виправлено вголос.
- **Пункт 1 ← сюжет тижня про залежність від постачальника.** 28.08 суд зняв з Anthropic держблокування; сьогодні OpenAI сама блокує конкурента через його власника. Обидва рази питання одне: **хто може відрубити тебе від моделі і за що**.