AI-ДАЙДЖЕСТ

Амодеї сам відповів на звинувачення

Гендиректор Anthropic вийшов у X з двома постами і визнав провину.

неділя, 16 серпня 2026

У ЦЬОМУ ВИПУСКУ

ТЕМА 1

Даріо Амодеї особисто відповів на публічне звинувачення двома довгими постами. Найважливіше в них - визнання провини

Найбільша історія доби, і вона розгорталась ланцюжком, який варто пройти повністю.

Як усе почалось. У подкасті All-In інвестор **Ґевін Бейкер** (@GavinSBaker) заявив: «Мені казали **кілька людей, яким я довіряю**, що Даріо казав...» - далі про те, що Anthropic вважає, що може лишитись **єдиною компанією у світі**. І додав: «Наполегливо відраджував би Даріо колись комусь таке казати».

Спростування. @_sholtodouglas (Шолто Дуглас, дослідник Anthropic, 558 тис. переглядів) відповів різко: «**Абсолютно неправда.** Тейки Ґевіна подобаються, але той,

від кого він це почув, **бреше**, щоб воно лягало в наратив, у який дехто так відчайдушно хоче, щоб ви вірили».

А далі те, чого зазвичай не буває. Бейкер виклав розгорнуту відповідь по суті (1.5 млн переглядів): проблема в тому, що чутка **правдоподібна**, бо збігається з публічною позицією Даріо. Його теза: якщо ШІ **може** бути небезпечним, є два шляхи - «сконцентрувати його в руках кількох обраних компаній і політиків через регуляцію, або **широко розподілити**». І цитує Цукерберга: «Уявлення, що ШІ настільки небезпечний, що єдиний безпечний шлях - **крайня концентрація влади**, саме по собі виглядає проблемним».

І тоді написав сам Даріо - двома частинами, 2.7 млн переглядів на другій. Обидві прочитані повністю в оригіналі.

Частина 1, про регуляцію: вибір «концентрація або розподіл» він називає **хибною дилемою** і наводить конкретику, яку можна перевірити: SB 53 (підтримали) **звільняє компанії з виручкою нижче $500 млн**, тобто б'є по фронтир-лабах і **щадить малих**. «Дуже стараємось робити пропозиції, які **сповільнюють фронтир-компанії, водночас даючи перевагу меншим конкурентам**. Це шкодить бізнес-інтересам фронтир-лаб і допомагає челенджерам, зокрема **відкритим вагам**».

Частина 2, і заради цього абзацу варто читати весь ланцюг. Замість захисту Даріо визнає головну претензію: «Думаю, **найточніша критика ШІ-компаній, включно з Anthropic, - що вони ще не виконали своїх гучних обіцянок** принести користь світу. Це **цілком на нас**, і саме цю критику варто робити - замість усього цього про меседжинг і маркетинг».

І там же особисте, чого в корпоративних заявах не буває: «**Втратив батька через гепатит С** за кілька років до появи препаратів прямої дії (софосбувір), які **виліковують 95% пацієнтів** і, ймовірно, вилікували б його». Плюс пряма відмова від піару: глянцева кампанія з позитивним ухилом не спрацює; спрацює те, що **справді вилікує рак**.

Чому це важливо. Три речі, і всі прикладні.

① **Той самий фільтр, що був виведений 15.08 на історії Ашенбреннера, тільки з другого краю.** Тоді висновок був: гучність прогнозу і його точність - різні осі. Тут свіжий приклад, як **чутка від «кількох людей, яким я довіряю»** проходить у подкаст з мільйонною аудиторією і живе там добу, поки її не спростує людина з самої компанії. Джерело «мені казали» джерелом не є. Рівно тому в дайджесті стоїть лінк на пост.

② **Практичне про регуляцію і гроші.** У суперечці лежить розвилка, яку варто розуміти: Бейкер прямо каже, що **меседжинг Даріо допомагає рухам проти дата-центрів у США** і що «майже всі великі компанії, крім Anthropic, підписали лист Дженсена». Це рамка, щоб читати наступні новини про регуляцію ШІ без паніки.

③ **Найкорисніше - як він тримає удар.** Відповідь по суті, з конкретними номерами законів, з визнанням найсильнішої претензії проти себе і **без переходу на особистості**.

Як зразок комунікації в конфлікті це вартісніше за будь-який тред про «як писати в LinkedIn».

Чого не перевірено: сам подкаст All-In не прослухано, цитата Бейкера взята з відео-кліпу в треді; хто саме сказав Бейкеру ту фразу, невідомо, і це **лишається неспростовним і непідтвердженим** одночасно; чи справді SB 53 працює так, як описує Даріо, у тексті закону **не звірено**. [proven щодо того, ХТО і ЩО написав - усі чотири пости прочитані повністю в оригіналі; fuzzy щодо початкової чутки]

Даріо, частина 1/2 · частина 2/2 · Шолто: «абсолютно неправда» · розгорнута відповідь Бейкера

ТЕМА 2

Есей «Робота з ШІ більше схожа на лідерство, ніж на код» - 285 балів, і Саймон Вілісон у коментарях додає спостереження просто по темі

Аллен Барджі написав короткий есей з тезою: код давав визначеність, та сама програма на той самий вхід дає той самий результат, інакше це баг. **Люди ніколи такими не були.** Робота з ШІ ближча до другого: «Дратує, коли до ШІ ставляться як до **компілятора**. Це стає кориснішим, коли до взаємодії ставляться як до **співпраці**».

Головна думка не про промпти: «Хороший промпт допомагає, але **спільний робочий контекст допомагає більше**».

Найцінніше - у коментарі @simonw (Саймон Вілісон, автор `llm`): «Багато людей, які отримують **справді хороші результати** від LLM і агентів, - це люди зі **значним досвідом управління людьми**. Компанії на кшталт Anthropic, схоже, це теж розуміють. Вражає, скільки CTO і CEO Anthropic найняла на **індивідуальні інженерні позиції**».

І одразу чесне обмеження від нього ж: «Керувати агентами **значно легше**, ніж людьми. Не треба зважати на бажання, цілі, думки чи емоційний стан агента. У людей є суб'єктність; в агентів, попри назву, - нема».

Чому це важливо. Це опис того, як влаштована щоденна робота з асистентом на LLM, і воно пояснює, чому працює те, що працює.

«Спільний робочий контекст важливіший за промпт» - це і є проєктні конвенції, записані інструкції та довгострокова пам'ять. Кожен запит «запиши це собі» будував рівно те, про що есей. Різниця між сесією з накопиченими нотатками про граблі й без них у тому, що контекст спільний.

А тепер незручна половина. Вілісон каже, що керувати агентами легше, бо в них нема суб'єктності, але **ціна помилки менеджера тут та сама**. Корекції рівня «не міняй тон», «доводь роботу до кінця сам», «новини з X даються криво» - це менеджерські корекції, і кожна з них дає більше, ніж доналаштування промпту. [proven - есей і тред прочитані повністю]

Есей Аллена Барджі · HN-тред, 285 балів

ТЕМА 3

232× прискорення ядра через «авто-дослідження» - і чому це подається з поміткою «місячної давнини»

410 балів. Автор (sankalp) взяв конкурс GPU Mode на реалізацію батчевої QR-декомпозиції і зайняв **12 місце зі 183 учасників з прискоренням 232× над базовим рішенням**, ганяючи Codex у циклі.

Спершу чесні дужки, бо без них цифра бреше. Пост датований **08 липня 2026** - це місячний текст, який учора виплив на HN. І 232× - це **результат змагання проти навмисно простого baseline**. У проді нічого на 232× не прискорилось.

Тепер про те, що там реально цінне. Метод, який автор описує так: «Агенти **прагнуть щільних циклів зворотного зв'язку**». Задача підходила для авто-дослідження, бо мала готову інфраструктуру перевірки: чекер, що звіряє результат, і бенчмарк. **За 14 днів було зроблено понад 1500 сабмішенів.** Плюс зізнання, яке варте більше за цифру: «**Що краще щось знаєш, то краще можеш промптити LLM**». Автор не був експертом з CUDA, але знав основи, і саме це дозволило ставити правильні питання. Людина просто над ним у таблиці - **principal engineer з NVIDIA**.

У треді практики підтверджують шаблон незалежно: один прогнав FlashAttention-оптимізацію на DeepSeek-V4-Flash за 1-2 години **за $0.2**; інший описує той самий цикл `бенчмарк → профайл → верифікація → покращення` як щоденний робочий режим. Ключова умова від нього: «Поки можна вказати агенту "**ось правильний baseline, переконайся, що оптимізації його проходять**" - їх можна лишати працювати».

Чому це важливо. Найпрактичніший пункт випуску.

Рецепт розкладається на три умови, і вони або є, або нема: **машинна перевірка правильності** · **числова метрика** · **дешевий швидкий прогон**. Де це вже є, там такий цикл можна вмикати хоч завтра. Де нема (наприклад, «зроби дайджест краще») - **ніякий цикл не допоможе**, бо нема на що хіл-клаймити, і результат доводиться правити людиною.

Друге, важливіше: **1500 сабмішенів за 14 днів** - це людина, яка 14 днів тримала цикл і додавала ідеї, коли він застрягав у локальному максимумі. Рівно та рамка, що в пункті 2: керування. [proven - стаття прочитана повністю; **сам код і сабмішени не перевірялись**, таблицю лідерів не відкривали, цифра 232× - заява автора]

Пост про авто-дослідження · HN-тред, 410 балів

ТЕМА 4

Топова історія HN доби - «ШІ не переграє математиків, він їх перепам'ятовує». І перший же коментар питає, чому вона взагалі на першому місці

447 балів, найбільше за добу. Теза Давіде Піффера: коли ШІ розв'язує складну математичну задачу, це пояснюють зростанням інтелекту, хоча є простіше пояснення. **Робоча пам'ять.**

«Людський математик може тримати в голові **лише невелику кількість незнайомих елементів одночасно**. Модель може тримати весь текст задачі, сотні проміжних рівнянь, кілька відкинутих підходів, визначення й обмеження **всередині свого контекстного вікна**». І образ, який пояснює все: «**Папір не робить розумнішим. Він розширює робочу пам'ять**».

Але подається з двома застереженнями, і вони суттєві.

Перше: **найвищий коментар у треді - про накрутку**: «Метакоментар: **як цей пост потрапив на 5 місце головної сторінки з одним апвоутом через дві хвилини після сабміту?**» Відповіді розділились, переконливого пояснення в треді нема. 447 балів тут **менш надійний сигнал якості**, ніж зазвичай.

Друге: пост датований **04 серпня**, знову не вчорашній.

Найзмістовніша репліка по суті - від `d--b`: «Щойно ШІ робитиме аргументи, які потребують робочої пам'яті на **сотні елементів**, не буде способу ці аргументи зрозуміти… **доведеться довіряти системі**».

Чому це важливо. Прикладний висновок тут вужчий, ніж здається, і він про робочий процес.

Якщо перевага - це **обсяг утримуваного контексту**, то корисність падає рівно там, де контекст рветься: нова сесія без пам’яті, застарілі інструкції, факт, що не доїхав у довгострокові нотатки. Саме тому в таких конвеєрах будують нічну консолідацію пам’яті. І це аргумент **проти** рефлексу «зараз перечитати все заново»: дешевше тримати контекст цілим, ніж щоразу відновлювати.

А теза `d--b` - це вимога перевірності, сказана з другого краю. Поки цифра дається **з лінком на скрипт або першоджерело**, довіряти системі не треба. [promising щодо самої тези - це есей з посиланнями на дослідження робочої пам'яті, не нове дослідження; **статті Alloway і Passolunghi не перевірялись**, беруться як цитовані автором]

«ШІ не переграє математиків» · HN-тред, 447 балів

ТЕМА 5

a16z: розрив у витратах на ШІ між компаніями - 600×. Ґреґ Брокман: «звучить точно». Але в самій статті цифра інша

@a16z виклали Charts of the Week з тезою: «Дикий розрив в адопції: **топ-1% витратників на ШІ витрачають у понад 600 разів більше, ніж медіанна компанія**». @gdb (Ґреґ Брокман, OpenAI) процитував це коротким «sounds accurate», 76 тис. переглядів.

Перевірка першоджерела показує інше формулювання. У тексті статті (автор Moses Sternstein, 14.08) написано: «топ-10% фірм витрачають **~50× більше на працівника, ніж медіана**». 600× - це інший зріз (медіана проти топ-1%), він у графіку, не в тексті. Обидві цифри реальні, але **у твіт пішла гучніша**, і саме її процитував Брокман.

Плюс застереження, яке автор дає сам і яке в твіт не влізло: «**Дані Ramp схиляються до техкомпаній, це варто тримати на увазі**».

Найзмістовніше в статті - висновок з нього: «"Використовуй найкращу модель завжди" - можливо, хороший підхід для деяких компаній, але **точно не для всіх**». І дані BCG по 107 публічних компаніях: два верхні квінтилі за споживанням токенів мали помітно **швидше зростання виручки**.

Чому це важливо. Дві речі, друга практична.

① **Це рівно та механіка, через яку вчора сталася помилка на $15 млрд FT.** Цифра живе в заголовку і твіті в одному вигляді, а в тексті в іншому. Правило підтвердилось: **іти в першоджерело навіть тоді, коли цитує CEO OpenAI**. Авторитет того, хто ретвітнув, не робить цифру перевіреною.

② **Про вибір моделі.** «Найкраща модель завжди» - не оптимум. Розумна розкладка виглядає так: аналіз і генерація тексту йдуть на важкій моделі, транскрипція на **локальній моделі**, а рутинні перевірки за розкладом взагалі без LLM, звичайним скриптом. Це те, що стаття називає ефективним фронтиром. [proven - стаття прочитана, обидві цифри звірені з текстом; **графік з 600× не відкривався** (він картинкою), береться з підпису у твіті; **звіт BCG не читався**]

Charts of the Week (першоджерело) · твіт a16z · «sounds accurate» від @gdb

---

ТЕМА 6

Хайтен Шах: як робити продукти, коли софт став дешевим. 300 тис. переглядів і механіка, яку можна забрати сьогодні

@hnshah виклав есей прямо в X, 300 тис. переглядів. @dharmesh (засновник HubSpot): «Абсолютно сподобався цей пост. Наче Хайтен **читає думки**».

Ядро тези: «Раніше **дефіцит був у реалізації**. Можна було мати десять хороших ідей і потужність серйозно взятись за дві. Вартість побудови нав'язувала пріоритезацію, хотілось того чи ні. **Це швидко змінюється**... Зрештою "чи можемо ми це побудувати?"

стає **нудним питанням**. І лишається те, що завжди було важчим: **чи варто це будувати?**»

А далі найпрактичніше - **дивитись на механіку під ними**, з розбором на прикладі пул-реквесту: хтось пропонує шматок роботи → робота стає **видимою іншим** → фідбек чіпляється **прямо до неї** → її можна інспектувати **асинхронно** → апрув **міняє її стан** → історія **лишається**. І висновок: ця механіка «може заїхати досить далеко за межі розробки», у маркетинг, фінанси, дизайн, юридичку, **роботу, згенеровану ШІ**.

Одразу з застереженням від автора: це **не** «зроби таблицю фіч конкурентів і клонуй» - «це, мабуть, один з найшвидших способів перетворити хороший продукт на купу лайна».

Чому це важливо. PR-механіка Шаха описує, чого зазвичай бракує в роботі з асистентом.

Коли робиться щось нетривіальне - править розклад задач, збирається замовлення, готується публічний драфт - воно **не проходить жодну зі стадій**: робота не стає видимою до того, як зроблена, фідбек чіпляється постфактум у чат, стан не міняється апрувом, історія розсипана між чатом і нотатками. Частково це закривається правилом «незворотне тільки з явного дозволу», але тільки для незворотного.

Найдешевший крок, якщо колись захочеться зробити такий процес передбачуванішим: **видимість роботи до її завершення**. [proven - есей прочитаний повністю в оригіналі]

Есей Хайтена Шаха · реакція @dharmesh

ТЕМА 7

Стенфордський професор виклав власну оцінку біобезпеки після статті про згенеровані ШІ геноми. Рідкісний жанр: вчений сам called out ризики своєї роботи

@BrianHie (Браян Хай, професор Стенфорду, Arc Institute) після публікації статті про **згенеровані ШІ фаги** написав особистий текст про біобезпеку. Пост ретвітнув @patrickc (Патрік Коллісон, Stripe).

Що він каже, дослівно і збалансовано в обидва боки. **Про поточний ризик, заспокійливо і конкретно:** експерименти були «безпечні й контрольовані», згенеровані фаги лишаються **близько в просторі послідовностей** до шаблону дикого типу ΦX174, повністю de novo фагів вони **не намагались** робити. Плюс бар'єр входу: потрібні «час, ресурси і **команда мультидисциплінарних експертів**» з ML, біоінформатики й мікробіології. Висновок: «Я б оцінив поточну загрозу перенесення цих технік з фагів на синтез патогенних, високо-новітніх, згенерованих ШІ вірусів, що інфікують людей, як **дуже низьку**».

Але далі чесна половина: інтерес публіки викликаний «тим, **які можливості можуть бути попереду**». Їхня стаття - «**перші згенеровані ШІ геноми**», і генома-масштабний

дизайн відкриває функції, недосяжні на рівні окремих генів. «Згенеровані ШІ біомолекули або ШІ-керовані мутації **можуть допомогти завдати значної шкоди людям**».

Чому це важливо. Прямої дії нуль, галузь непрофільна, і оцінити біологію по суті немає можливості.

Але тут ідеальний зразок редакційного правила, яке тримається з 07.08: **коли автор розповідає про власну роботу, шукай, де він применшує**. Особливість цього тексту в тому, що шукати не довелось: він сам відокремив «зараз ризик низький» від «далі буде інакше», дав перевірювані деталі під перше і не сховав друге. Це протилежність жанру «felony humble-bragging», який розбирався на прикладі Black Hat.

Практичний фільтр звідси на майбутнє, він переноситься куди завгодно: **заява про безпеку варта чогось тоді, коли поруч лежить те, що робить її фальсифікованою**: які саме експерименти, які межі, що конкретно не робили. Без цього «ми все зважили» дорівнює «повір мені, бро» з пункту про FHE. [proven - есей прочитаний повністю за лінком з t.co; **саму наукову статтю про фаги перевірити неможливо** і оцінити біологічні твердження неможливо; це виклад автора про власну роботу, тобто зацікавлена сторона]

«Present and future of safe biological AI» · твіт @BrianHie

ТЕМА 8

«Інженери зроблять що завгодно, аби не вчитися з історії» - 147 балів, і найкраще пояснення того, чому це вигідно

Есей horn.gg: індустрія систематично перевідкриває вже вирішене, перейменовуючи його.

Топовий коментар пояснює механіку грошима: «**Можна заробити багато грошей, роблячи щось на вигляд новим і невідкритим**, і відповідно виставляючи себе розумним і на вістрі».

Чому це важливо. Це прямий сусід «нудних технологій», якими вимірюється прогрес з 14.08, тільки з боку стимулів: у людини, яка **продає** новинку, є фінансова причина не помічати, що це вже було.

Практично, фільтр для стрічки: коли наступного разу зайде «новий патерн для агентів», перше питання - «**як це називалось раніше**». Половина «агентних оркестрацій» - це черги задач і воркфлоу-двигуни, які старші за них обох. [proven - есей і топові коментарі прочитані]

Есей · HN-тред, 147 балів

ТЕМА 9

«Досить слати мені величезні пул-реквести» - рант на 142 бали, який став актуальним саме зараз

Рант розробника про PR, які неможливо рев'ювити. Свіжості темі додає очевидний контекст: **агенти генерують великі дифи легко**.

Найцікавіше - у треді, де обговорюють механічне рішення і одразу знаходять його межу. Пропозиція: зробити CI-джобу, що **відхиляє PR понад N рядків**, бо «агенти розуміють технічні обмеження, як-от падіння CI». Заперечення, миттєве і влучне: «Ага, **кумедний спосіб отримати величезні стоси PR, кожен з яких окремо незрозумілий**».

Чому це важливо. Пряма рима з пунктом 6: ліміт розміру - це метрика-проксі, і оптимізація під неї дає **гірший** результат при формально кращій цифрі. Те саме, що «зелений CI» ≠ «код правильний».

Висновок конкретний: цінність зміни в репо в тому, щоб вона була **розповідною**, одна зміна = одна причина. [proven - рант і тред прочитані]

«Stop sending me huge PRs» · HN-тред, 142 бали

ТЕМА 10

«ШІ-психоз Cloudflare» - 106 балів, і корисне тут суперечка в треді

Пост opensauce.it звинувачує Cloudflare, що захоплення ШІ коштує їм надійності, з відсиланням на серію збоїв.

Це подається заради того, як тред його розібрав. Захист: «Баги трапляються постійно. Вони приблизно **зростають з масштабом. Не спадають**... цей конкретний баг ідеально проходить крізь сито: кілька сервісів, важко помітити на рев'ю, **невидимий при малому трафіку**». Атака: «Баги - так, збої - ні. Якщо збої ростуть з масштабом, отримується **один-два безкоштовні проходи**. Після цього це або невміла інженерія, або невміле керівництво». І контр-удар: «В AWS буває більше одного збою на рік - вони що, несерйозні?»

Чому це важливо. Тема непрофільна, але **фраза «невидимий при малому трафіку» - це діагноз половини граблів** у будь-якій браузерній автоматизації.

Застиглий DOM, `aria-selected` замість вмісту, обрізаний URL - жоден з них не падає **гучно**. Вони віддають правдоподібний результат, і саме тому виявляються через дні. Сьогодні такий випадок було зловлено ще раз: лінк на статтю Хайтена побудовано «за логікою», отримано **HTTP 200** і сторінку «Session not found». Двохсотка тут доказом не є. [proven - пост і тред прочитані; **самі інциденти Cloudflare перевірити неможливо** по їхніх постмортемах, беруться як заяви сторін у треді]

«Cloudflare's AI Psychosis» · HN-тред, 106 балів

misc - коротко, що ще варте ока

• **Привиди в Unicode** (195 балів) - найкращий технічний лонгрід доби. У японському стандарті JIS X 0208 з 1978 року живуть **ієрогліфи-привиди** (幽霊文字), у яких **нема джерела** і ніхто не знає, що вони означають. Розслідування 1997-го з'ясувало походження: 峯 виник як **помилка копіювання** - «山 над 女» надрукували окремо, вирізали, наклеїли на аркуш і скопіювали, а лінія стику паперу прочиталась як риска. Ці символи досі в Unicode

• **@naval**: «**Не можна створити Бога і посадити його на повідець**» - 3.2 млн переглядів, найвірусніше в обох листах за добу. У misc свідомо: це афоризм, аргументу тут нема, і в добу, коли Даріо пише дві сторінки конкретики про SB 53, різниця між жанрами особливо помітна

• **@levelsio: рекорд доходу з X - $24 744 за місяць** з розкладкою (ad rev share $9 307 + $7 032, підписки $1 389, книга $4 658, мерч $773). Цікаве що **дохід від самого X - $19 312/міс**

• **@emollick оновив гру 1987 року через Codex** - Infocom відкрили вихідники, він попросив агента перенести «Nord and Bert» у сучасний UX. І деталь про робочий процес: «Усе це зроблено **з ChatGPT на телефоні**, який віддалено підключався до Codex на комп'ютері під час подорожі»

• **@emollick про інді-розробників ігор** (77 тис. переглядів): їх «карають за ШІ **суворіше, ніж великих розробників**», хоча вони найбільш обмежені в ресурсах. Висновок: «більшість інді-розробників мають стимул **використовувати ШІ і приховувати це**»

• **@garrytan**: «У когось ще Codex Desktop **вилітає на чатах**?» - 32 тис. переглядів, 80 відповідей. Не новина, але маркер: збої в інструментах агентів тепер обговорюють як погоду

• **Jane Street і $15 млрд** (134 бали) - історія з **пункту 7 попереднього випуску** дійшла до HN. Нічого нового по суті: попередній випуск давав її з есеєм «When Genius Fails» і поміткою про пейволл FT

• **RISC-V: They should have known better** (174 бали) - той самий лонгрід, що вже був у misc попереднього разу. Досі в топі, змісту не додалось
