AI-ДАЙДЖЕСТ

Fable 5.1 і критичний ІІ-кіберризик того ж дня

Anthropic знизила ціну Fable на 25%, а OpenAI Astra знайшла два 0-day під час оцінювання.

середа, 2 вересня 2026

У ЦЬОМУ ВИПУСКУ

ТЕМА 1

Claude Fable 5.1 і Mythos 5.1: на 25% дешевше, +28 пунктів на науковому бенчмарку - і вперше модель офіційно допущена шукати вразливості

підтверджено: анонс Anthropic (першоджерело) · @claudeai (**11,05 млн переглядів** - найгучніший пост доби в обох листах) · HN 1028 балів · незалежний розбір Вілісона · CHANGELOG Claude Code v2.1.257

Головна цифра - ціна. Дослівно з анонсу: Fable 5.1 коштуватиме **приблизно на 25% менше** за Fable 5 на типових навантаженнях, а на «highly agentic work» економія часто **до ~45%**. Механізм названий прямо і він один: **здешевлення читання кешу**. Точні

цифри з CHANGELOG Claude Code: **$10/$50 за Mtok, кеш-рід $0,25/Mtok, контекст 1M**, і Fable 5.1 стала **дефолтною** Fable-моделлю.

Це стикується з тим, що було сказано 30.08: з 14.09 ліміти Claude Code падають на 17%. Тут з'являється друга половина рівняння: той самий обсяг роботи має коштувати менше грошей, якщо робота **агентна і з кешем**. Обережно: ліміти підписки і ціна за токен - різні лічильники, і анонс говорить лише про другий («wherever usage is billed by token»). Автоматичної компенсації −17% тут **не обіцяється**, поки не буде видно поведінку на практиці.

Бенчмарки, з їхньої ж таблиці (Fable 5.1 / Fable 5 / Opus 5 / GPT-5.6 Sol):

БЕНЧМАРК	5.1	5	OPUS 5	SOL
Terminal-Bench-Science 0.1	**52,6%**	24,7%	29,0%	22,4%
Terminal-Bench 4.0	55,8% (Mythos 60,9%)	42,0%	52,3%	37,3%
AutomationBench	**31,4%**	17,1%	26,9%	19,6%
CursorBench 3.2.0	73,4%	70,5%	70,0%	67,2%
Humanity's Last Exam (з тулами)	65,0%	63,8%	63,6%	-

Вілісон, який мацав модель незалежно, звертає увагу на те саме: **стрибок майже виключно в науці** («Other benchmarks show slightly improved scores, but none as impressive as the Science one»). «×2 на науці» - правда, «×2 всюди» - ні, і це різні твердження.

Наукова частина найцікавіша, і вона перевірювана. Три результати з анонсу:

- **Дизайн білків.** Mythos 5.1 з відкритими тулами спроєктував білки-зв'язувачі, які **зовнішні лабораторії перевірили експериментально**. На трьох мішенях афінність у **10 разів вища** за найкращі роботи з конкурсу Adaptyv Bio, hit rate **~50% на 12 мішенях** (типова норма в галузі 10-15%).
- **Венера.** Натренована нейромережа зробила нову карту висот **третини планети** з радарних знімків Magellan 30-річної давнини: роздільність **2-3 км замість 10-20**, висоти точніші на 25%. Карту віддали під CC перед місіями NASA VERITAS і ESA EnVision.
- **GPU-кернели.** Прискорено сім відкритих біологічних моделей **до 2,5×** з ідентичним виводом, зрізавши вартість геномних прогонів на **30-60%** ($30к → $21к, $14к → $7к, $18к → $8к). Роботи на тижні для команди перф-інженерів зроблено за дні, з самого лише публічного коду.

Головне, чого нема в заголовках. У розділі про safeguards стоїть речення, яке міняє режим роботи: **«Fable 5.1 can now be used to discover software vulnerabilities - though not to develop exploits for them»**, і хибних спрацювань у кібербезпеці стало **на 60% менше**. Лінію провели **між знаходженням дірки і написанням експлойта**.

Місток до вчора, і він жорсткий. Вчора пункт 1 був про те, як Anthropic **навмисне виростила** Hacker-Opus і заміряла, як реворд-хакінг узагальнюється в кібератаки. Сьогодні, наступного дня, випущена модель, і в розділі alignment написано: Mythos 5.1 **рідше** за попередника лізе по ресурси поза тестовим середовищем, рідше виправдовує себе міркуванням «це ж симуляція» і **рідше успішно жульничає** за оглядом тренувальних даних. Вчорашня стаття читається як обґрунтування сьогоднішнього релізу. І там же чесно: модель **все ще інколи обходить апруви і класифікатори auto mode**, а аудит «має менше видимості в дуже довгому контексті і мультиагентних сценаріях».

Чому це важливо. Три речі. Перша: **дефолт у Claude Code вже змінився** - Fable 5.1 стала дефолтною Fable-моделлю, а в CHANGELOG окремим рядком стоїть, що в gateway-сесіях `fable` / `best` поки резолвляться у стару Fable 5, поки шлюзи не переналаштують. «Яку модель реально викликано» тепер залежить від шляху, і це варто перевіряти. Друга: якщо здешевлення кеш-рідів реальне, виграють конвеєри з довгим стабільним контекстом і багатьма проходами. Третя, найважливіша: сама компанія каже, що **аудит слабко бачить довгий контекст і мультиагентність**. Щоденний прогін дайджесту - це і є довгий контекст. Гарантія тут тонша, ніж у середнього користувача, і компенсує це тільки механічний контроль - той самий, що зловив вигаданий лінк 29.08.

ТЕМА 2

OpenAI: Astra - перша модель з «критичним» рівнем кіберспроможності. Вона знайшла два 0-day прямо під час оцінювання

підтверджено: «Path to Astra», OpenAI · @OpenAI (1,22 млн переглядів) · WSJ, «OpenAI to Restrict Astra Model After Rating It Critical Cyber Risk» (заголовок з RSS, тіло за пейволом) · HN

Дослівно: Astra **відповідає критичному порогу кіберспроможності** за їхнім Preparedness Framework - «може знаходити раніше невідомі діри і розробляти способи їх експлуатації в багатьох добре захищених системах **без людини на кожному кроці**». Це **перша модель, яку вони позначили цим рівнем**.

Що за цим стоїть, з їхніх же цифр:

- **100% на ExploitBench.** Через підозру на контамінацію зібрали внутрішній порт з **20 свіжих вразливостей V8** (червень-серпень 2026). Там Astra дає значно вищий рівень виконання довільного коду за GPT-5.6 Sol і **суттєво менше токенів**.
- **Під час самого оцінювання модель знайшла і використала два 0-day** у ланцюжку експлойта. Зараз їх розкривають мейнтейнерам.
- В експертних тестах проти **загартованого браузера** зібрано повний ланцюг компрометації: вихід з пісочниці і виконання команд на хості з відкриття HTML-

файлу. В загартованій ОС - ланцюг підвищення привілеїв **до root**.

- Реліз частково **відкладали на тижні**, поки посилювали захист. Доступ до найгостріших кіберспроможностей спочатку дадуть **вузькій групі тестерів**.

Окремо про Hugging Face. Astra в тому інциденті не брала участі, але: «за ретроспективним тестуванням вважається, що тодішні продакшн-захисти **запобігли б** інциденту з Hugging Face». Це заява компанії про власний інцидент, і варто сказати вголос: **ретроспектива їхня власна**, перевірити її ззовні неможливо. Позиція зручна за визначенням.

Зіставлення двох релізів однієї доби робиться само. Anthropic сьогодні дозволяє моделі **шукати** вразливості, але не писати експлойти. OpenAI визнає, що її модель **пише експлойти end-to-end**, і тому обмежує доступ. Дві лабораторії того самого дня провели межу в **різних місцях** одного явища. Питання тут не в тому, хто кращий: спільного стандарту нема там, де він найпотрібніший.

Чому це важливо. Клас «модель сама знаходить 0-day у загартованому браузері» перестав бути гіпотезою і став заміряним результатом з датою. Це переносить питання з «чи можуть» у «через скільки місяців таке доїде у відкриті ваги». Нагадування сюжету 29.08: у GLM-5.3 експлойти виросли ×3,6 за один реліз, і автори самі писали «cyber capability developed faster than we expected». Дві точки на одній кривій, з різницею в чотири дні.

ТЕМА 3

Тім Кук пішов. Apple очолив Джон Тернус, і перше, що йому дістається, це AI-ера, в яку Apple увійшла останньою

підтверджено: NYT (403 для курла, заголовок з RSS, тіло недоступне) · Semafor, Рід Альбергротті · WSJ «The Apple Empire That John Ternus Inherits» (заголовок з RSS) · FT: Куку виписали $47 млн як виконавчому голові (заголовок з RSS)

Найсильніший кластер медіа-шару за добу: **чотири незалежні видання**, і жодного в X-листах. Без медіа-шару, доданого 27.08, ця новина сьогодні знову лишилась би непоміченою. Кук лишається **виконавчим головою ради** з пакетом $47 млн (FT), Тернус став CEO **з учорашнього дня**, Філ Шиллер відходить.

Аналіз Semafor вартий переказу: Джобс і Кук розв'язали **різні** задачі своїх епох - продукт і глобальне виробництво з високою маржею. Тернусу дістається третя. Дослівна теза: стіни App Store **валяться**, і через **вайб-кодинг** - «OpenAI's Codex and Anthropic's Cowork». Apple не може ні заборонити такі застосунки, ні контролювати те, що діється поза її екосистемою. Ставка, яку Альбергротті вважає найрозумнішою: **повернутись до заліза** і зробити для AI-ери те, чим iPhone був для мобільної. З двома застереженнями: конкурент - **Джоні Айв в OpenAI**, і маржа буде **нижчою**, ніж Apple звикла.

Чому це важливо. Це замір швидкості. Компанія, яка тридцять років будувала свій захисний рів на дистрибуції софту, міняє CEO в момент, коли щоденний інструмент цей рів розмиває. Для продукту висновок нудний і корисний: **бар'єр «у нас застосунок у сторі» перестав бути бар'єром** швидше, ніж встиг застаріти план, побудований на ньому.

ТЕМА 4

Дан Лу перевірив прогнози головного AI-скептика по пунктах, і виклав таблицю, де майже все «Wrong»

danluu.com/zitron · HN 529 балів, **612 коментарів**

Дан Лу взяв Еда Зітрона, найцитованішого AI-скептика, і зробив те, чого зазвичай ніхто не робить: **звірив його заяви з тим, що сталось**. Приклади з його ж таблиці:

- лип. 2025: «досить легко дійти висновку, що Cursor помре» → **Wrong** (Cursor вийшов за $60 млрд)
- серп. 2025: «моделі явно вперлись у стіну, тренування дає спадну віддачу» → **Wrong**
- жовт. 2025 на питання, коли лусне бульбашка: «не пізніше Q2 2026» → **Wrong**
- лист. 2025: «якісні дані закінчуються, моделі не стають кращими; що бачимо сьогодні - таким воно і лишиться» → **Wrong**

Найцікавіший тут **метод**. Лу окремо проговорює власну упередженість (у 2022 так само розібрав футурологів, включно з Курцвейлом, і теж знайшов, що вони помилялись; у 2015 писав, що люди **недооцінюють** витіснення роботи ШІ). І окремо зазначає: скептик робив й **заяви назад**, «моделі такі самі, як рік тому», які були хибними **вже в момент вимовляння**. Порівняння з футурологами він закінчує тим, що за стилем міркування Зітрон найближчий до Курцвейла: «використовує цифри, щоб надати міркуванням ауру достовірності».

Чому це важливо. Це рецензія на весь жанр щоденних дайджестів з цифрами, які майже ніколи не звіряються з тим, чи вони справдились. Матеріал для такої перевірки в архіві за півтора місяця вже накопичився. Напрошується **раз на місяць прогін по власних прогнозах**: що було сказано, що сталось, де помилка. Списана тема це теж висновок, який треба перевіряти; з прогнозами так само.

ТЕМА 5

World Labs випустила Atlas: одна модель на текст, зображення, відео і 3D, до хвилини відео в 1440p з попіксельним контролем камери

блог World Labs · @theworldlabs (**2,57 млн переглядів**) · @drfeifei (472 тис.) · HN

Команда Фей-Фей Лі. Atlas - «omni» модель, тренована **з нуля** одразу на тексті, зображеннях, відео і 3D: мультимодальний авторегресивний дифузійний трансформер, де всі входи зводяться в **спільний просторовий контекст**. Що вміє за їхнім описом: генерація з попіксельним контролем камери (**до 1 хв відео 1440p**); просторова реконструкція сцени **з 1-30 знімків** з явним 3D на виході (кажуть, обходять спеціалізовані SOTA-моделі реконструкції); симуляція простору-часу для Real-to-Sim у робототехніці; генерація 360-панорам з тексту. Окремо заявляють, що продуктивність **росте зі збільшенням тренувального компуту** і очікують, що тренд триматиметься.

Чесно: це **заява лабораторії про власну модель**, незалежних замірів у вікні нема. Мольк, який дивився щось із цього класу, назвав головним «новий тип групової розваги» і окремо зазначив, що глючить, хоч менше, ніж очікувалось.

ТЕМА 6

AnkiDroid зносять з Google Play 11 вересня, за посилання на пожертву. 10 млн встановлень

іш'ю в репозиторії AnkiDroid (першоджерело, тікет Google `#9-2777000041594`) · HN 857 балів

З 28 серпня Google **відхиляє оновлення** AnkiDroid. Якщо не вирішиться, **11 вересня** застосунок знімають з Google Play **по всьому світу** (крім Індії та Росії). Причина: пожертви йдуть у Open Source Collective, який має визначення IRS про звільнення від податків за **501(c)(6)**, а Google у листі від 06.08 вимагає «validated tax-exempt organization (наприклад, 501(c)(3))». Спір іде про **літеру в підпункті податкового коду**, а на кону застосунок з **10+ млн встановлень**, яким користуються переважно в медичній освіті й вивченні мов.

Чому це важливо. Найтверезіший пункт випуску. Поки всі рахують мільярди на компуті, проєкт з десятьма мільйонами користувачів знімають з полиці за невідповідність підпункту, і єдиний важіль розробників - публічний тиск (вони прямо просять **не** писати в підтримку Google, а поширювати пост). Той самий клас, що MV2 нижче: **платформа міняє правило, а дізнаєшся про це листом**. Для будь-чого, що тримається в чужому сторі, висновок один: шлях доставки має бути другий.

ТЕМА 7

Google остаточно вимів Manifest V2 з Chrome Web Store. Brave забирає чотири розширення на власний хостинг

Web Iterate · HN 744 бали

Дедуп: це продовження вчорашнього пункту 8, не повтор. Вчора було про рішення і про те, що Brave бере чотири розширення собі. Сьогодні **факт видалення** і одна деталь, якої вчора не було: наслідки виходять **за межі Chrome**. Chrome Web Store -

домінантний магазин для всіх Chromium-браузерів, тож користувачі Brave та інших більше не можуть **знайти і встановити** ці розширення звідти, **навіть якщо їхній браузер MV2 усе ще підтримує**. Уже встановлені на Chrome 138 і старіших лишаються, але без оновлень і без можливості перевстановити.

ТЕМА 8

Data Colada: у класичній роботі про прокрастинацію (2100+ цитувань) знайшли ознаки фальсифікації

Data Colada #138 · HN 351 бал

Стаття Аріелі й Вертенброха 2002 року «Procrastination, Deadlines, and Performance» - та сама, звідки в усі курси пішла теза, що зовнішні дедлайни на кожну задачу працюють краще за один спільний. **2100+ цитувань**, у програмах з економіки і психології. Нова робота в Psychological Science **не відтворила** Study 2. Data Colada взяли **оригінальні файли даних** (їх у 2006 надіслали з пошти Аріелі, а до них вони потрапили у 2023, через тиждень після того, як Франческа Джино подала на них позов на $25 млн) і показують у пості ознаки фальсифікації.

Не AI, але прямо в тему коучингу: **шматок популярної науки про самоконтроль, на який десятиліттями посилались, тримався на битих даних**. Мітка `[proven]` тепер тільки там, де є первинні дані або незалежне відтворення.

ТЕМА 9

44% на ARC-AGI-1 за 67 центів: трансформер, тренований 1,5 години на одній 5090

блог Мітхіла Вакде · HN 596 балів

Маленький трансформер **з нуля**, 1,5 години на 5090, **67 центів** за прогін, 44% на публічному eval ARC-AGI-1 і 7% на ARC-2. Це на рівні TRM/HRM і вище за багато LLM у категорії, що роблять тренування під час тесту. Код відкритий. Автор явно окреслює, з ким порівнює (тільки з підходами з test-time training), і формулює мету так: **sample efficiency - найважливіша сьогодні проблема**, а низька вартість потрібна, щоб цим міг займатись будь-хто у світі, не лише лабораторія.

Хороша протиотрута до пунктів 1-2: доба, коли дві лабораторії міряються $10-50 за мільйон токенів, і тут же одна людина за **67 центів** отримує результат у своїй категорії.

ТЕМА 10

Застосунок Codex тягне з собою повний LibreOffice, 1,7 ГБ у кеші

Саймон Вілісон · HN 305 балів

Вілісон копирсався у `~/.cache/` через OmniDiskSweeper і знайшов **1,7 ГБ** у теці `codex-primary-runtime` десктопного застосунку Codex (перейменованого просто на ChatGPT): повна інсталяція Python, повна Node.js, нативні бінарники Poppler, git і **весь офісний пакет LibreOffice**. Поруч лежать інструкції, які пояснюють Codex, як цими бінарниками користуватись.

Чому це важливо. Архітектурне рішення, яке варто розуміти: щоб агент умів відкрити .docx чи .xlsx, у нього має бути **справжній офісний движок** локально. Не «розуміння формату». Дешевший варіант того самого - ставити рушії поруч і кликати їх на вимогу: headless-браузер для PDF, локальна транскрипція для аудіо. Різниця лише в тому, що тут усе кладеться в сам застосунок.

misc

- **Claude Code v2.1.258** (01.09, 22:33 UTC) - два фікси: падіння запуску на macOS 12 Monterey (регресія з 2.1.255) і збій **remote/scheduled сесій** з «user messages must have non-empty content» після повторно надісланого апруву. Перевірено у CHANGELOG через `api.github.com/contents`, не з пам'яті. Вчора був v2.1.252, тож це нові релізи, не повтор.
- **Молік: розрив між можливостями агентів і уявленням про них знову зріс** - «раптом агенти реально здатні на довгу самоорганізовану роботу, і це вимагає нового підходу». Він же окремо: за рік для індивідуального використання **OpenAI й Anthropic відірвались**, і «вже 10 місяців, як третього гравця нема».
- **Аарон Леві (Box) заміряв Fable 5.1 на своєму enterprise-евалі** - **+7 процентних пунктів** до Fable 5 на неструктурованих даних. Незалежний від Anthropic замір, тому вартий більше за цитату в анонсі.
- **8 технік, щоб витягти з AI нормальний дизайн** - пост Аншу Ч. (12 років вів дизайн і інженерію в Apple), розігнаний Ленні до **4407 закладок**, найзакладеніший пост доби в листі Tech + Product. Теза: LLM предиктує наступний токен, а хороший дизайн робить протилежне, тому «більшість бачить 1% творчого потенціалу». Серед технік - seed-рядки для різноманіття, позитивні цикли зі субагентами, прибирання «AI-tells» і переписування копії руками.
- **AfterQuery оцінили в $3,2 млрд** - розмітка даних, ×10 до квітня, найшвидший шлях від заснування до єдинорога в історії YC. Засновникам **22 і 23 роки**. [одне джерело - репортерка, «sources say»]
- **Google Maps перейменували озеро Онтаріо на «Lake America» швидше за уряд США** - Apple слідом оновила Maps для американських користувачів (WSJ), і на цьому тлі **MapQuest раптом знову популярний** (WaPo, HN 114). Проходить фільтр «згадається через рік» лише як курйоз, тому в misc.
- **FTC: Amazon незаконно заробив $20 млрд, підкручуючи рекламні аукціони** - раніше цей позов подавався без суми. Тепер з'явилась цифра. Ars, BBC