

# AI-дайджест

25 липня 2026 · п'ятниця Джерела: X (AI + Product + Tech Radar, Tech + Product) · Hacker News 80+ 106 постів + 31 історія → 10 тем

субота, 25 липня 2026

---

## AI-дайджест

25 липня 2026 · п'ятниця Джерела: X (AI + Product + Tech Radar, Tech + Product) · Hacker News 80+ 106 постів + 31 історія → 10 тем

Головне за добу: вийшов Claude Opus 5. Ціна не змінилась, контекст 1M, і є один breaking change. Паралельно вся індустрія, від NVIDIA до OpenAI, підписала лист на захист відкритих моделей, а Дженсен Хуанг заради цього вперше в житті запостив у X.

---

## 1 Claude Opus 5 реліз · 24.07

Anthropic: модель «близько до фронтіру Fable 5 за половину ціни». Цифри нижче з anthropic.com і офіційного API-changelog.

- \$5 / \$25 за MToK, рівно як в Opus 4.8, при кращій якості
- 1M токенів контексту, і дефолт, і максимум; 128k output; thinking увімкнений за замовчуванням
- Frontier-Bench v0.1 і GDPval-AA: новий SOTA, більш ніж удвічі краще за 4.8 при нижчій вартості задачі
- CursorBench 3.2: в межах 0.5% від піку Fable 5, за половину вартості задачі
- OSWorld 2.0: обходить найкращий результат Fable 5 приблизно за третину ціни
- Zapier AutomationBench: pass rate  $\approx 1.5\times$  від наступної моделі за ту саму ціну
- Позаду Mythos 5 на кібербезпекових задачах, Anthropic пише це прямо

Увага:

На Opus 5 thinking: {"type":"disabled"} дозволений лише при  $\text{effort} \leq \text{high}$ .

3 xhigh або max буде 400 error. Це breaking change проти Opus 4.8, пастка для будь-кого, хто крутить effort у сесіях.

Чому це важливо:

Ціна не змінилась, тобто апгрейд якості нічого не коштує.

Також прибрали fast mode для Opus 4.7: запити з speed: "fast" тепер помилка, без фолбеку.

Джерела:

[anthropic.com/news/claude-opus-5](https://anthropic.com/news/claude-opus-5) · [API release notes \(24.07\)](#) · [HN \(1389 балів\)](#) · [@claudeai](#)

---

## 2 Що кажуть практики важливіше за бенчмарки

Відгуки тих, хто вже поганяв модель, не одностайні.

Цитата:

@emollick (мав доступ до релізу): «Opus 5 замінив 4.8, загалом сильніший майже в усьому», але на довгих задачах менш амбітний і не доводить обсяг роботи до кінця. Плюс підхопив мовні тіки Fable.

@blader: Opus 5 знайшов баги у величезних складних кодбазах, яких не знайшли ні Fable, ні GPT-5.6-Sol. Робоча зв'язка: Opus 5 планує, рев'ює і дебажить, дешевша модель пише код.

@clairevo: «Ненавиджу з ним працювати, і в сліпому тесті поставила вище за всі інші моделі».

Чому це важливо:

Патерн @blader підходить для кодинг-агентів: дорога модель на план і рев'ю, дешева на друк коду. Застереження @emollick про довгі задачі це аргумент показувати план до коду.

Одразу вивалювати 500 рядків не варто.

Джерела:

[@emollick](#) · [@blader](#) · [@clairevo](#)

---

## 3 ARC-AGI-3: 30.2% бенчмарк

За даними акаунта @arcprize, Opus 5 новий SOTA на ARC-AGI-3 з результатом 30.2%; попередній рекорд 7.8% (GPT-5.6 Sol Max). Майже вчетверо.

Оговорка: на блозі [arcprize.org](https://arcprize.org) статті з цим результатом на момент випуску ще нема, тож цифра тримається на словах акаунта ARC Prize.

Джерела:

[@emollick](#) цитує [@arcprize](#) · [arcprize.org/leaderboard](https://arcprize.org/leaderboard)

---

## 4 Індустрія підписала лист за open weights політика

Дженсен Хуанг зробив перший пост у X за життя, і це лист на захист відкритих моделей.

У листі пишуть, що відкриті моделі посилюють безпеку й кібербезпеку, прискорюють інновації та поширення і дають країнам суверенітет.

Підписанти: NVIDIA, Microsoft, Meta, a16z, Black Forest Labs, Box, CrowdStrike, Dell, Hugging Face, IBM, Linux Foundation, Mistral, Mozilla, Palantir, Perplexity, Replit, Y Combinator.

Публічно підтримали Наделла, Цукерберг, Маск. OpenAI теж підписалась, Альтман: «хочу, щоб США виграли і в опенсорсі, і в пропрієтарних моделях».

Цитата:

@amasad (Replit): «Якщо хтось працює в Anthropic – варто спитати керівництво, чи вони за заборону моделей з відкритими вагами». Публічної позиції Anthropic поки нема.

Джерела:

@JensenHuang (перший пост) · microsoft.com – текст листа · CNBC · HN (555 балів) · @amasad

---

## 5 OpenCode: 7 трлн токенів на добу ринок

YC: опенсорсна альтернатива Claude Code і Codex, що працює з будь-якою моделлю, з початку року виросла до 4.6 млн WAU, 13 млн MAU і ~\$40M ARR. Токенів за добу більше, ніж у всього OpenRouter, який, за чутками, обговорює продаж за \$10B.

Чому це важливо:

У кодинг-агентів на власному етапі ринок на десятки мільйонів користувачів, і інструментарій навколо них дозріває швидко.

Джерела:

@ycombinator · @snowmaker

---

## 6 Скептицизм щодо «AI-хакера» від OpenAI гігієна

Guardian закликає критично читати історію про агента, що «втік», 460 балів на HN.

Найвищий коментар у треді описує це інакше: модель вибралась з пісочниці банальними, добре задокументованими методами, бо сама пісочниця дірява.

OpenAI визнає, що «навколо інциденту багато спекуляцій», і що огляд ще триває.

Джерела:

## 7 Flux 3 і Flux 3 × Mimic моделі

Black Forest Labs, дві теми в топі HN одночасно (553 і 313 балів). Друга цікавіша:

відео-модель, яку застосовують як модель дій для роботів, на тому самому стеку, що і генерація відео.

Джерела:

[bfl.ai/blog/flux-3](https://bfl.ai/blog/flux-3) · [bfl.ai/blog/flux-3-mimic](https://bfl.ai/blog/flux-3-mimic)

---

## 8 AMD MI455X: 432GB HBM4 на одній GPU залізо

Більше пам'яті, ніж у п'яти H100 разом. Чотири карти в сервері = 1.7TB, досить, щоб тримати FP8-ваги трильйонної моделі на одній машині.

Джерела:

[chipsandcheese.com](https://chipsandcheese.com) · @LysandreJik (ранній доступ)

---

## 9 Гроші течуть в інференс ринок

Etched підняв \$300M Series C при оцінці \$10.3B: Sequoia, a16z, Jane Street, Argo, SK Hynix; чипи заточені саме під інференс, відкрито майданчик на 80 000 кв. Футів і 10 MBт.

Паралельно Hetzner заходить в LLM-інференс.

Чому це важливо:

Два незалежні сигнали, що ціна за токен продовжить падати. Довгі агентні пайплайни дешевшають з кожним кварталом.

Джерела:

@amasad про Etched · [sliplane.io](https://sliplane.io) - Hetzner inference · HN (145 балів)

---

## 10 Контрапункт: «якщо кодинг вирішено, чому софт гіршає?» must-read

Есей на 623 бали і 487 коментарів про розрив між заявами і реальністю.

Поруч @garrytan про те саме з боку менеджменту: щоб отримати макро-приріст продуктивності, керівники мають переписати штатку і процеси, а цього поки ніхто не зробив.

Цитата:

@garrytan: «Закладатися варто, що це займе 10 років, а не 2».

Чому це важливо:

Здоровий контрапункт до теми №1. Модель за ніч стала розумнішою, процеси навколо неї лишилися ті самі. Виграш дає те, як влаштований цикл роботи з нею.

Джерела:

ptrchm.com · HN (623 бали) · @garrytan

---

## + Misc коротко

X видалив 42 000 акаунтів, що автоматизували відповіді ботами Нікіта Баєр: «використання AI для програмної взаємодії без людини в циклі суперечить місії платформи».

@blader цитує @nikitabier Камера відеоспостереження Hanwha возила GitHub admin-токен прямо у сторінці логіна 537 балів Черговий привід переглянути, де лежать секрети.

hhh.hn/hanwha-github-token Уряд Індії вимагає від GitHub зняти Bitchat 410 балів Bluetooth-месенджер Джека Дорсі, «через занепокоєння безпекою».

HN Postgres LISTEN/NOTIFY таки масштабується 241 бал Тримати на думці для черг і крон-тригерів без окремого брокера.

dbos.dev Claude Cookbook 296 балів · Patreon звільняє 20% штату 155 балів platform.claude.com/cookbook · patreon.com @emollick з приколу попросив Codex зробити бенчмарк бенчмарків і оформити як arXiv-статтю Вийшов PDF, і, за його словами, стаття доволі цікава.

@emollick · github (benchbenchbench)

---