AI-ДАЙДЖЕСТ

# NVIDIA прокачала Opus 5 обгорткою, не моделлю

*ARC-AGI-3 з 30% до 100% без зміни ваг; окремо 175 балів на HN про тиху обрізку effort у Claude Code.*

неділя, 23 серпня 2026

---

У ЦЬОМУ ВИПУСКУ

1. NVIDIA підняла Claude Opus 5 з 30% до 100% на ARC-AGI-3 - не міняючи модель. Різниця тільки в обгортці агента
2. Користувачі кажуть, що Anthropic тихо зрізала шкалу effort у Claude Code. 175 балів на HN, у чейнджлозі - тиша. Але доказова база слабша за заголовок
3. Той самий сюжет з третього боку: однакові ваги на одному GPU дають різні токени - і винна не модель, а бекенд
4. Леві: вузьке місце в поширенні AI - не моделі, а evals. І поруч цифра: $100 млн/рік на інференс без жодного офлайн-евала
5. Продовження вчорашнього: Ox Alpha, який учора був «чия модель - невідомо», сьогодні протестований Молыком і виявився не фронтиром
6. Новий роадмап MCP: агентні примітиви, єдиний HTTP-транспорт і ідентичність агента замість вставлених API-ключів
7. Munder Difflin: опенсорсний харнес, щоб тримати «офіс власних клонів». #1 на GitHub Trending, 3 696 зірок, 263 бали на HN
8. Молык: Codex і Claude Code вже нормально заповнюють форми з пошти без нагляду. І одразу сам пише умову, на якій це працює
9. Cursor відкриває Grok Bot для компаній - та сама тема, яку вчора я свідомо відклав як «без анонсу і без цифр»
10. Брокман про темп: «рік тому 10 млрд токенів на рік давали пам'ятну табличку - тепер це тиждень»

---

ТЕМА 1

## NVIDIA підняла Claude Opus 5 з 30% до 100% на ARC-AGI-3 - не міняючи модель. Різниця тільки в обгортці агента

Блог NVIDIA Developer (21.08), анонс @NVIDIAAI - 70 балів на HN, другий тред на розбір.

**Цифри з першоджерела, дослівно з підзаголовка:** «The research project elevates **Claude Opus 5 from a 30% model baseline to 100%** as part of the complete AVO agent system, showing that **system design - not model capability alone** - can unlock frontier-level long-horizon performance».

Система зветься **AVO** (Agentic Variation Operators). Результати:

| ЩО | ЦИФРА |
|---|---|
| ARC-AGI-3 public set | **100.00 RHAE**, усі **183 рівні** у **25 середовищах** |
| Ефективність | на **12% менше дій** у середовищі, ніж VISTA |
| GPU-ядра (та сама архітектура) | 500+ напрямків, 40 версій ядра, **+10,5% до FlashAttention-4** на DGX B200 |

**Тут варто поправити джерело, з якого тема прийшла.** У збір вона зайшла постом @ClementDelangue (CEO Hugging Face, 17,2 тис. переглядів): «**NVIDIA built its own coding harness to optimize CUDA GPU kernels** and achieved a 100% score on ARC-AGI-3». Звучить так, ніби ядерний оптимізатор випадково взяв бенчмарк. У самій статті сказано інше: AVO це **загальнопризначений агент**, «For ARC-AGI-3, we connected **the same general-purpose agent** to a different task interface. **The underlying agent remains the same**; only the environment-specific tools and evaluation change». Оптимізація ядер і ARC - це дві демонстрації однієї архітектури. Не побічний ефект однієї з них. Пост Клема лінка на джерело не має взагалі, тому цифри взяті з блогу NVIDIA.

**Чому це важливо.** Стаття каже дослівно те, на чому тримається будь-який агентський харнес: «evaluating a model **is not the same as** evaluating an agent». Практичний висновок: коли щось не виходить, першою гіпотезою має бути **обгортка** - що в контексті, які інструменти, як ловиться помилка, чи є пам'ять між кроками. Чотири механізми AVO - persistent memory, supervision, tool-use і **recovery from failure**. Останній у більшості саморобних харнесів і відсутній, а без нього збій у фоновому запуску проходить мовчки. [proven - цифри від виробника, незалежної перевірки ARC-команди на публічному лідерборді ще нема, див. «чого не перевірив»]

---

ТЕМА 2

## Користувачі кажуть, що Anthropic тихо зрізала шкалу effort у Claude Code. 175 балів на HN, у чейнджлозі - тиша. Але доказова база слабша за заголовок

Тред @argofowl, HN - 175 балів, 27 коментарів.

Твердження дослівно: «anthropic enrols **fable 5** sessions on claude code **2.1.236+** into an experiment that shrinks the effort scale, older versions and **opus 5 are left alone**... if "high" feels like "low" for you, you're in the test group». І окремо: «since 2.1.237 the model reads "high" effort as **10 out of 100**, the exact number "low" used to be - and the changelog doesn't say a word».

**Що перевірено.** Чейнджлог Claude Code витягнуто через GitHub API (raw, не рендер), прочитані записи **2.1.235 - 2.1.241**. Про зміну шкали effort - **жодного слова**: 2.1.237 це «Fixed prompt caching for sessions using an LLM gateway» плюс новий стиль виводу

Concise; 2.1.240 і **2.1.241 (вийшов сьогодні о 03:52 за Києвом)** - обидва «Bug fixes and reliability improvements». Частина заяви «в чейнджлозі нічого нема» **підтверджена**.

**А от сама метрика тримається слабко.** Найвищий коментар на HN б'є точно в метод: «the evidence, as best I can tell from the tweet, is that they **asked Claude what effort level it was set to**. But **how would the model even know that?**» Тред далі розходиться: один каже, що effort передається системним промптом і тому питання валідне, інший - що так у Qwen, але не в більшості моделей. Вимір зроблено **запитуванням моделі про саму себе**, а це рівно той клас доказу, який уже названий «індикатор замість даних».

Найтверезіша пропозиція в треді - від @toolshed_labs: «If enrolment is gated on the client version, then **pinning the previous client and rerunning the same prompts is the test**. Everyone comparing this week to last week **moved the client and the server at once**». Це єдиний спосіб відрізнити реальну зміну від відчуття.

**Чому це важливо.** Міняти щось наосліп рано. Заява стосується **Fable 5**. **Opus 5** її не зачіпає взагалі. Практичне: якщо здається, що сесія «отупіла», у Claude Code є `/effort`, і перевіряти треба **поведінку на однакових промптах**. Питати модель, як вона себе почуває, марно. [fuzzy - сам факт експерименту непідтверджений; підтверджено лише те, що чейнджлог мовчить]

---

ТЕМА 3

## Той самий сюжет з третього боку: однакові ваги на одному GPU дають різні токени - і винна бекенд

Level1Techs, лонгрід з експериментами - 231 бал на HN, тред.

Автор бере **офіційний BF16-чекпоінт Qwen3.6-27B** на одній RTX PRO 6000 Blackwell, вимикає все, що могло б заплутати (eager execution, без CUDA graphs, без prefix caching, без квантизації ваг і KV-кешу), і міряє, наскільки різні бекенди розходяться в тому, який токен наступний. Результат: «For the first several thousand tokens, **every run of the model agreed**... Then **in later portions of the prompt, backends began disagreeing**». Розходження з'являється **на довгому контексті** - саме там, де живе агентська робота.

Порада автора про заміри: «Do not crank temperature to zero and paste in 3 test prompts then call it good/bad. **Zero-shot tests are not a good analog of most agentic tasks.** You need **long-context tool-calling** and domain specific knowledge evaluations».

**Чому цей пункт стоїть поруч з другим.** Три різні історії за добу говорять одне: **«та сама модель» - це ілюзія**. У NVIDIA різницю дала обгортка (30% до 100%), у Claude Code люди підозрюють серверний експеримент, тут - інший бекенд на тому самому залізі. Практичний висновок: **«здається, гірше працює» - не діагноз**. Або замір робиться на однакових промптах з довгим контекстом і тул-колами, або висновок не робиться взагалі. [proven - авторські заміри з описаною методикою і графіками]

ТЕМА 4

## Леві: вузьке місце в поширенні AI - evals. І поруч цифра: $100 млн/рік на інференс без жодного офлайн-евала

@levie (CEO Box) - 14 тис. переглядів: «AI diffusion is **far more rate limited by having good evals** than most realize. The kind of evals that you see for every model release are incredibly helpful, but only tell you the shape of general AI progress and the relative capability level of models».

Він цитує @BrendanFoody: «It is insane how many enterprises I meet that are spending **$100M / year on inference** and **don't have offline evals** to determine which model to use».

**Чому це важливо.** Та сама думка, що в пунктах 2-3, тільки з боку бізнесу. Типова ситуація в маленькій автоматизації: евалів нема взагалі, якість міряється тим, чи була скарга. На одному користувачі це працює, далі ні. Найдешевший перший крок: зафіксувати десяток типових запитів з очікуваним результатом і ганяти їх, коли міняється модель або промпт. Без цього будь-яке «стало гірше/краще» лишається відчуттям. [proven як теза про практику; цифра $100 млн - заява практика без розкритої вибірки]

ТЕМА 5

## Продовження вчорашнього: Ox Alpha, який учора був «чия модель - невідомо», сьогодні протестований Молыком і виявився не фронтиром

Учора (пункт про Inkling і misc) Ox Alpha проходив з поміткою «чия це модель, не оголошено, не вгадано». Сьогодні є замір.

@emollick, 39 тис. переглядів: «I am **not as blown away** by the mystery model Ox Alpha as people appear to be. In my very early experiments, it seems fine, but **not at the frontier even among open weights**». Тест - його стандартний tbigl-шейдер «неоготичне місто», той самий, на якому він ганяв Kimi K3. Через сім годин підтверджує: «This is a **consistent view across every test I am running**. Nice model, but not frontier and **I am not sure why there has been so much buzz** as if it is».

Контекст, чому взагалі шум: OpenCode роздає її безкоштовно тиждень - 1M контексту, мультимодальність, zero data retention, «capacity for 100T tokens per day».

**Чому це важливо.** Гучний безкоштовний доступ ще не якість. Витрачати час на міграцію під стелс-модель варто після того, як її проміряв хтось із відомою методикою. Молык - саме такий випадок: у нього один і той самий тест місяцями, тому його «не фронтир» має вагу. [promising - один рецензент з одним тестом, зате з послідовною методикою]

ТЕМА 6

## Новий роадмап MCP: агентні примітиви, єдиний HTTP-транспорт і ідентичність агента замість вставлених API-ключів

Офіційний блог MCP (22.08, автори - Lead Maintainers David Soria Parra і Den Delimarsky), HN - 183 бали.

П'ять напрямків, з них три реально стосуються теми:

• **Agentic messaging primitives** - «Modern agentic workloads **no longer fit the standard request-and-response pattern**». Тобто server-initiated events (вебхуки і канали), «so clients **aren't left polling** for results», плюс дозрівання розширення Tasks (SEP-2663).

• **HTTP-native transport unification** - один транспорт на все, включно з локальними серверами, що говорять Streamable HTTP поверх stdio.

• **Agent identity** - найважливіше: «MCP authorization today is built around **a person approving access in a browser**... but more and more of the callers are **agents running as cloud workloads with their own identity**, acting on behalf of a user **who isn't present**». Рішення будують на DPoP, Workload Identity Federation і token exchange - «rather than **pasted API keys and long-lived tokens**».

**Чому це важливо.** Останній пункт описує стан справ у будь-якій автоматизації за розкладом: агент ходить **без нікого біля браузера** і тримається на довгоживучих токенах і куках у профілі браузера. Кожна історія «сесія протухла» - це воно. Стандарт ще не готовий, тож поки що тільки спостереження; але коли DPoP і delegation доїдуть, це рівно той шлях, яким запуски за розкладом мають ходити замість вставлених ключів.

У HN-треді найвищий коментар - про **code mode** (Cloudflare) як альтернативу: замість вантажити всі MCP-інструменти в контекст, дати моделі писати код, що їх викликає. Аргумент «progressive discovery - kind of late to the party».

ТЕМА 7

## Munder Difflin: опенсорсний харнес, щоб тримати «офіс власних клонів». #1 на GitHub Trending, 3 696 зірок, 263 бали на HN

Сайт, репо (MIT, 410 форків), HN-тред.

Ідея: обгортка над **CLI-агентами, які вже встановлені** - «works with your **existing subscriptions** (uses hourly limits)», підтримує 12 провайдерів (Claude Code, Codex, Gemini CLI, OpenCode, Cursor, Copilot...). Клони «захоплюють воркфлоу» і мають спільну пам'ять: «Every clone you run **shares that memory**, so the next one you spin up **starts already knowing how you work**». Все локально: «Your code, your keys, your existing subscription - **nothing leaves your machine**». Платне - тільки Teams (приватна хмара 24/7 і E2E-мережа між клонами колег).

**Чому це важливо.** Форм-фактор тут той самий, що у більшості саморобних агентських конструкцій: CLI-агент на власній машині, на вже оплаченій підписці, зі спільною пам'яттю між сесіями. Munder Difflin додає до нього **паралельний офіс клонів**. Цінне в цій новині одне: #1 дня на GitHub Trending підтверджує, що локальний харнес на своїй підписці став мейнстрімом. У треді є і різке «this project is cringe», тож консенсусу нема.

---

ТЕМА 8

## Молік: Codex і Claude Code вже нормально заповнюють форми з пошти без нагляду. І одразу сам пише умову, на якій це працює

@emollick, 21 тис. переглядів: «In terms of everyday usefulness and saving time, **Codex & Claude Code are very capable** of doing the thing where you ask them to "**fill out the forms that I got an email about**" and they do it well & **without further intervention** from you. Really nice for **low-risk** time-consuming stuff».

І окремим постом - застереження, яке в переказах зазвичай губиться: «To make this work you need to run the ChatGPT or Claude apps on your computer & **turn on browser control, which means you need to trust the models to do that**. Definitely **check the work carefully first** until you understand them».

**Чому це важливо.** Для будь-кого, хто вже пустив агента в залогінений браузер, ключове слово тут **low-risk**. Межа проходить по зворотності дії: читати, збирати, складати чернетку - так; тиснути «оплатити» і слати листи від чужого імені - ні. Молік ставить цю умову сам, у тому ж треді.

---

ТЕМА 9

## Cursor відкриває Grok Bot для компаній - та сама тема, яку вчора свідомо відклали як «без анонсу і без цифр»

@mntruell (CEO Cursor) - **522 тис. переглядів**, найгучніший пост доби в обох листах: «We're **opening up Grok Bot access to a small set of enterprises this weekend**. Reply if you'd like for us to swing by your office in San Francisco and onboard your team tomorrow or Monday».

**Явний місток.** Учора Grok Bot стояв у секції «свідомо не пішло» з формулюванням: «форм-фактор "канали для агентів" цікавий і лягає у вчорашній пункт про Slack Code, але це **обговорення в трьох постах без анонсу і без цифр**. Тримається на оці». Сьогодні є анонс від самого CEO, тож тема піднімається. **Конкретики так і нема**: ні що вміє бот, ні скільки компаній, ні ціни. Гучність дає посада автора і формат «приїдемо до вас в офіс завтра». [fuzzy]

ТЕМА 10

## Брокман про темп: «рік тому 10 млрд токенів на рік давали пам'ятну табличку - тепер це тиждень»

@gdb, 72 тис. переглядів, цитує **@xeophon** (Florian Brand): «Not even a year ago, **10B tokens in a year** was a big achievement and got you a plaque. Now I do **>10B tokens a week**».

Масштаб, який пояснює решту випуску: коли обсяг росте в 50 разів за рік, у лабораторій з'являється рівно той стимул, який обговорюється у пункті 2 (тихо економити комп'ют), а в компаній - та проблема, про яку каже Леві в пункті 4 ($100 млн на інференс наосліп). Це заява ентузіаста про власне споживання. Звітної цифри за нею нема. [fuzzy]

**misc:** @awilkinson зібрав «Bookshelf» на сайті через Claude Code - найпідкреслюваніші книжки з Kindle плюс власні коментарі з розсилки, 23 тис. переглядів · Саймон Вілісон: рецензувати кожен рядок - не єдиний спосіб перевірити агента «Eyeballing every line of code has never been the most effective way to validate a change» · @emollick проти ELI5 за замовчуванням - 52 тис. переглядів: «you aren't five! You don't need things **dumbed down**, you need things **explained differently**» - краще формулювання «поясни, спираючись на те, що вже відомо» · @amasad: «"Pretty soon" turned out to be 3 months» 119 тис. переглядів · Replit за тиждень: 7 релізів включно з імпортом скілів з GitHub · @patrickc показує запуск Inkling безкоштовно 118 тис. переглядів - вчорашній пункт 7, тепер однорядковою командою · @avlok: на SpaceX 75% акціонерів беруть акціями замість кешу при звичайних 2% · @levelsio і німецькі податки - 236 тис. переглядів, «Germany does NOT take 50%. They take 49.3%», джерело OECD · @bentossell: «i never update chrome, i always update my agent apps» · hdiutil оголошений застарілим у macOS 27 173 бали - тред, стосується будь-яких скриптів з .dmg