

OpenAI знизив ціну GPT-5.6 Sol

\$4/\$20 за млн токенів, промо тримається до 21 листопада

субота, 22 серпня 2026

У ЦЬОМУ ВИПУСКУ

1. OpenAI зрізав ціну GPT-5.6 Sol на понад 20%. Пішов у прайс і звирив: \$4/\$20 за млн - і промо тримається не «3 місяці», а до 21 листопада
2. Kar1 додав перемикач «прибрати платні лінки з видачі» - і це найпопулярніша сторінка доби на HN, 1045 балів. Найцікавіше: сама фіча - один рядок у чейнджлозі, решту дописала аудиторія
3. Дан Лу: «Немає причин, щоб софт був повільним». Його агент за місяць зробив JIT-компіляцію регексів у ripgrep - прискорення 2-4× на довгих запитах і 7% на контрольних
4. DHH зібрав \$8 млн під Linux на десктопі: вісім патронів по мільйону, серед них CEO Stripe, Shopify, Cloudflare, Dell і Джек Дорсі
5. Три різні автори за добу написали одне й те саме: люди перестали читати машинний текст. Один з них - про фізичну неможливість сфокусуватись
6. Codex на AWS Bedrock: кеш-записи з'їдали 85% рахунку. \$1 386 за чотири дні там, де мало бути близько \$200
7. Thinking Machines роздає Inkling безкоштовно на OpenRouter - але тільки агентним харнесам, і за телеметрію
8. Анна з Архіву: ШІ-компанії скуповують і знищують паперові книжки. Тема гримить на HN двома тредами, але сам текст - від 5 серпня
9. Naval: «Фронтір-лабораторії мотивовані боятись - це моральне виправдання закритості»
10. DeepSeek тихо відкрив зір: deepseek-v4-flash-vision-exp, 472 бали і жодного анонсу

ТЕМА 1

OpenAI зрізав ціну GPT-5.6 Sol на понад 20%. У живому прайсі \$4/\$20 за млн, промо тримається до 21 листопада

Офіційний анонс @OpenAI - 1,6 млн переглядів, найгучніший пост доби в обох листах. Другий пост тієї ж гілки уточнює канали. Реакція @gdb: «towards giving our customers the lowest price on the market for any task, as well as the highest ceiling on capability».

Цифри з живого прайсу, не з твіту. Курл на openai.com сліпий (403 на все, включно з вигаданою адресою), тому сторінку тарифів відкрито залогіненим браузером:

МОДЕЛЬ	ВХІД	КЕШОВАНИЙ ВХІД	ВИХІД
gpt-5.6-sol	\$4.00	\$0.40	\$20.00
gpt-5.6-terra	\$2.00	\$0.20	\$12.00
gpt-5.6-luna	\$0.20	\$0.02	\$1.20

Довгий контекст рівно вдвічі дорожче (\$8/\$30 у Sol).

Дві речі, яких у твіті нема, а в прайсі є: • Дедлайн промо названий точно: «GPT-5.6 Sol's promotional pricing is available at least through **November 21, 2026**». Твіт каже розмите «for the next 3 months», документація дає дату і слово «at least». Це мінімальна гарантія з конкретним числом.

- **Data residency** коштує +10% для моделей, випущених після 05.03.2026. У свіжому Claude Code (v2.1.239, вчора ввечері) з'явився дзеркальний рядок: оцінки `/cost` тепер враховують 1,1× премію за **US-only inference**. Обидва вендори одночасно почали показувати надбавку за резидентність окремим рядком.

Дедуп проти вчора і поправка. 19.08 пунктом 7 йшло «GPT-5.6 Sol подешевшав удвічі» з цінами \$2.50/\$15 і поміткою «але тільки на OpenRouter». Сьогодні здешевлення прийшло в офіційний прайс OpenAI, і воно **менше** за те, що місяць тому давав OpenRouter: \$4/\$20 проти \$2.50/\$15. Там був маршрутизатор зі своєю маржею, тут вендорський прайс-лист. Але якщо тримати в голові тільки цифру з 19.08, сьогоднішня новина виглядатиме як подорожчання.

Чому це важливо. Для тих, хто сидить на підписці, прайс-лист чужого вендора нічого не міняє сьогодні. Цікавіше інше, і воно про **вимірювання**: обидва вендори за добу навчилися показувати надбавку за резидентність окремим рядком в оцінці вартості. Більшість пайплайнів вартість власних сесій не міряє взагалі. Поки все на підписці, це не болить; питання стає живим у той день, коли щось поїде через API з потоковою оплатою.

ТЕМА 2

Kagi додав перемикач «прибрати платні лінки з видачі» – найпопулярніша сторінка доби на HN, 1045 балів. Сама фіча це один рядок у чейнджлозі, решту дописала аудиторія

Чейнджлог Kagi, тред на HN – перше місце за добу з великим відривом (наступний матеріал має 703).

Дослівний обсяг фічі – **одне речення**: «we've added a setting for removing paywalled links from search results automatically». Решта випуску про віджет акцій, рендер Markdown у повідомленнях, пошук по тредах і два десятки багфіксів з номерами іш'ю.

Чому це цікаво як показник. 1045 балів набрала **сама можливість не бачити пейволли**. Тред це референдум про стан вебу, Kagi тут лише привід. Другий момент з того самого

чейнджлогу: у списку виправлень стоїть іш'ю «"CHATGPT" Wikipedia article is flagged as slop». Тобто механізм позначання «сміттєвого» контенту існує, і він **помилково спрацював на статті Вікіпедії**. У треді це майже цікавіше за фільтр пейволлів: платний пошук уже фільтрує й «якість», і в цього механізму бувають хибні спрацювання.

Чому це важливо. У будь-якому конвеєрі, що перевіряє джерела, пейволли ріжуть перевірку: тіло статті недоступне, лишається помітка «за пейволлом». Kagі цю проблему **ховає**: лінк зникає з видачі, стаття лишається недоступною. Корисніше зворотне – знати **заздалегідь**, що джерело платне, і не тягнути його як перевірене. Для цього достатньо явного списку платних доменів поруч зі списком тих, що віддають 403 роботам. Це сусідня категорія, і Kagі для неї не потрібен.

ТЕМА 3

Дан Лу: «Немає причин, щоб софт був повільним». Його агент за місяць зробив JIT-компіляцію регексів у `gprgr`: прискорення 2-4× на довгих запитах і 7% на контрольних

168 балів, **оригінал**, **тред**. Прочитано повністю.

Це пряме продовження пункту 4 від 19.08, тодішнього «бенчмарк-апокаліпсису»: агент за місяць обігнав Rust regex у 1,4× і виявився вдесятеро повільнішим на контрольному наборі. Сьогодні наступний крок того самого експерименту, і Лу тримає ту саму планку чесності.

Що зроблено: • той самий рушій FRE, перевчений так, щоб не бути переобученим на `rebar` • далі ідея, яку раніше не робили, бо дорого: компілювати нативний код регексу в окремому потоці, поки `gprgr` уже шукає звичайним матчером, і перемкнутись, коли компіляція готова • результат: 2-4× на довгих простих запитах, ~7% на репрезентативних холдаут-запитах з його реальної історії • ключова фраза: «ми можемо просто спробувати цей експеримент за кілька хвилин людського часу» – код-хірургія, яка для людини була б помітною роботою

Найцінніша теза про економіку. Він цитує Майкла Маліса: був мем «ШІ не допомагає, бо код ніколи не був складною частиною». У багатьох областях це правда, але JIT-компілятори саме той випадок, де писати код і був складною частиною. Рідкість JIT-ів пояснюється тим, що історично вони були задорогі, щоб окупитись. Тепер бар'єр упав. Марк Брукер додає: ймовірний результат це динамічний софт, підігнаний під конкретне навантаження. Не під клас навантажень.

Чому це важливо. Це та сама рамка, що вчорашній Карпаті про зношення абстракцій, тільки з іншого кінця і з цифрами. Скрізь, де «занадто дорого оптимізувати» колись було правильним рішенням, воно могло протухнути. Аргумент «це задорого для разової задачі» тепер треба перевіряти щоразу.

ТЕМА 4

DHH зібрав \$8 млн під Linux на десктопі: вісім патронів по мільйону, серед них CEO Stripe, Shopify, Cloudflare, Dell і Джек Дорсі

163 бали, **оголошення** – прочитано повністю. У X це найгучніший пост доби після OpenAI: **@dhh**, 1,2 млн переглядів, 9,9 тис. лайків, репостнув **@patrickc** – один з патронів.

Список патронів дослівно, по \$1 млн кожен: Тобі Лютке (Shopify), Патрік Коллісон (Stripe), Майкл Делл (Dell), Джек Дорсі (Block), Метью Прінс (Cloudflare), Брендан Ірайб (Sesame, співзасновник Oculus), Джейсон Фрайд (37signals) і сам DHH.

Фонд неприбутковий: тримає торгові марки, фінансує інфраструктуру і підтримує open-source проекти, від яких залежить Omarchy. Формулювання DHH: «це смішна сума грошей, тож є намір зробити так, щоб вона протрималась довго».

Чому це важливо. Технічно нічого не міняється. Але \$8 млн приватних грошей у десктопний Linux у 2026-му варті уваги вже тим, хто їх платить: люди, чиї компанії тримають половину інфраструктури, якою всі користуються (Stripe, Cloudflare, Shopify). Поруч у стрічці лежить **@levelsio про те, що Mac Search не працює**, і **Тео про перехід на Linux заради агентів** («файлова система швидша»). Це настрій. Заміру продуктивності агентів на macOS проти Linux немає, і вгадувати його не варто.

ТЕМА 5

Три різні автори за добу написали одне й те саме: люди перестали читати машинний текст. Один з них про фізичну неможливість сфокусуватись

Найсильніший матеріал – **«Я стаю III-сліпим»**, 306 балів, **тред**. Прочитаний повністю.

Рафал Цимерис описує конкретний симптом: він **фізично не може** проаналізувати документ, який йому прислали. «Здається, ніби це читається, але неможливо сфокусуватись на змісті». Закінчується нескінченним пінг-понгом з відправником про речі, які в документі вже написані.

Три його приклади впізнавані до болю: дизайн-док, схожий на копіпаст з Клода з характерним жаргоном («This cuts just through it», «The first gate is real»); 20-сторінкова маркетингова колода, де розумна стратегія перемішана з технічною маячнею; вимоги, які читаються як **внутрішні роздуми моделі**, не впевненої у власних рішеннях.

Пояснення механізму найцінніше: він вважає себе **«предтренованим»** на III-постах у LinkedIn, розсилках і сайтах. Мозок навчився ловити ознаки машинного тексту і **пропускати його не думаючи**, як банерну сліпоту.

Дедуп і чому це окремий пункт. Вчора був пункт 5 – **dontpastetheai.com**, 997 балів: там **прохання** не надсилати машинний текст. Позавчора – лінтер проти «load-bearing» у

білді, боротьба з **машинною мовою**. Сьогодні те саме з боку **читача, який уже не може**. Три дні, три різні місця, зростаюча гострота: прохання, інструмент, симптом. Трендом це названо тільки тому, що це третій незалежний матеріал поспіль.

Чому це важливо. У Цимериса є маркер, що влучає точно: «пітчити кожну дрібницю як прорив». Це головна хвороба щоденного формату, який підштовхує подавати рядок у чейнджлозі як подію. Найкраща конкретна порада з тексту: якщо документ описує чекбокси в RBAC, не продавати це так, ніби щойно винайшли вогонь. Запобіжник тут один - проміряний власний голос, зафіксований десь поза головою автора.

ТЕМА 6

Codex на AWS Bedrock: кеш-записи з'їдали 85% рахунку. \$1 386 за чотири дні там, де мало бути близько \$200

146 балів, [ish'ю openai/codex#37674](#) - прочитане повністю, закрити, 12 коментарів.

Суть: нативний провайдер `amazon-bedrock` у Codex CLI **не може ввімкнути явне кешування промπτу** для GPT-5.6 Sol. Запит не містить ні `prompt_cache_options`, ні `prompt_cache_breakpoint`, а конфіг провайдера дає лише транспорт і авторизацію, без трансформації тіла запиту.

Цифри автора з Cost Explorer (4 доби, 05-08.08):

ЗАПИТІВ	КЕШ-ЗАПИСІВ	ВАРТІСТЬ КЕШ-ЗАПИСІВ	РАЗОМ
3 656	171,94 млн токенів	\$1 182	\$1 386

85% витрат на модель - це кеш-записи, з яких нічого не читалось: у локальній сесії 76 запитів дали 6,7 млн `cache_write_input_tokens` і **нуль** `cached_input_tokens`. Середньо ~88 тис. кеш-записів на запит. Автор чесно позначає: це оцінка з usage-даних, не фінальний інвойс AWS.

Чому це важливо. Найпрактичніший пункт випуску, попри скромні бали. Схема відмови узагальнюється за межі Codex і Bedrock: **проміжний шар (гейтвей, провайдер, роутер) мовчки втрачає параметр, і про це стає відомо з рахунку**. Нічого не падає, все працює, метрики зелені. Найближчий аналог - **вчорашній фікс у Claude Code v2.1.237**: «Fixed prompt caching for sessions using an LLM gateway or custom base URL», той самий клас бага у тому ж інструменті, полагоджений позавчора. Правило варті запам'ятовування: **якщо між користувачем і моделлю є проміжний шар, треба перевіряти, що кеш реально читається. Не тільки пишеться.**

ТЕМА 7

Thinking Machines роздає Inkling безкоштовно на OpenRouter, але тільки агентним харнесам і за телеметрію

Анонс @thinkymachines через репост Коллісона (54 тис. переглядів), [документація](#) - перевірена, 200.

Дослівна умова: модель безкоштовна на OpenRouter **лише з агентними харнесами**, на найближчі кілька тижнів, і дані **використовуватимуться** (відв'язані від акаунтів), щоб покращити поведінку моделі. Це збір реальних агентних трейсів, чесно названий у самому анонсі.

Поруч у тій самій стрічці - [Ox Alpha](#), стелс-модель на OpenRouter: 1 млн контексту, вхід текст+зображення+відео, 230 балів на HN. Коллісон про неї: «It's very impressive». Чия вона, не оголошено, і вгадувати не варто.

Чому це важливо. Механіка нова і варта фіксації: **безкоштовний доступ в обмін на агентні трейси**. Це третій за тиждень сюжет про те, що агентна телеметрія стала валютою (18.08 - автономний агент у Jira, 20.08 - Linear з цифрами по задачах агентів). Хто збирається щось ганяти через OpenRouter, читає умови про дані. Не тільки ціну.

ТЕМА 8

Анна з Архіву: ШІ-компанії скуповують і знищують паперові книжки. Тема гримить на HN двома тредами, але сам текст від 5 серпня

703 бали на [дзеркалі.pk](#) і ще 542 на [.gl](#) - два окремі треди про один матеріал, разом більше за будь-що інше за добу.

Теза: компанії через посередників купують мільйони вживаних книжок, сканують і **знищують** їх, щоб отримати дані «не торкані машинами» до 2022 року. Мотиви, які називає автор: не дати відсканувати конкурентам, зняти юридичні ризики, і банально - знищити дешевше, ніж сканувати без ушкоджень. Заклик до волонтерів: сканувати рідкісні книжки, поки вони є.

Три застереження, без яких цей пункт був би дезінформацією:

1. Матеріал **не сьогоdnішній**. У шапці стоїть 2026-08-05, на HN він вистрілив учора, через два тижні після публікації.
2. Це **гостьовий пост волонтера**, перекладений з китайської. Редакційного статусу він не має.
3. Ключова конкретика, «Project Panama» Anthropic, подається як **розкрита в межах врегулювання на \$1,5 млрд**. Врегулювання про книжки справді існує і публічно відоме, але зв'язок конкретної кодової назви з ним **не перевірений у першоджерелах**, і за цю деталь автор не ручається. Решта посту - заклик і аргументація, без документів.

Чому це важливо. Пункт даний, бо це два перші місця на HN за добу. Але мітка чесна: [fuzzy] на конкретиці, [proven] лише на тому, що текст існує і зібрав ці бали.

ТЕМА 9

Naval: «Фронтір-лабораторії мотивовані боятись – це моральне виправдання закритості»

@naval, 210 тис. переглядів, 3,6 тис. лайків – найгучніша думка доби в листі AI + Product.

«Фронтір-лабораторії мають стимул боятись, бо це дає моральне виправдання лишатись закритими. Китайські лабораторії послідовно займають позицію робити те, що добре для всіх, і бути відкритими. Тож маленькі опиняються в незручному становищі, вболіваючи за китайців...»

Поруч – @emollick з протилежного боку: CEO лабораторій говорили про екзистенційний ризик задовго до того, як їхні компанії стали чогось варті, тобто вони в це вірили; піар це радше те, що вони перестали про ризик говорити.

Чому дано обидва варіанти без вибору. Це два несумісні пояснення однієї поведінки, і в жодного нема даних, лише інтерпретація мотивів. Чесніше показати шов, ніж підсунути одну версію як висновок. Єдиний фактичний коментар: за останній тиждень OpenAI спинила навчання через кібер-можливості (19.08, пункт 1). Це дія. Не заява, і вона лягає радше в бік Молліка.

ТЕМА 10

DeepSeek тихо відкрив зір: `deepseek-v4-flash-vision-exp`, 472 бали і жодного анонсу

Документація, тред. Класичний для них шлях: сторінка в доках.

Що вміє: зображення поруч з текстом (JPEG, PNG, GIF, WebP), формат визначається за вмістом файлу, не за розширенням і не за MIME. Три способи передачі, всі в OpenAI-сумісному форматі Chat Completions, працює і через Responses API. Ліміт тіла запиту – 48 MiB на base64.

Чому це важливо. Для задач із зображеннями (фото, скріни, PDF-аналізи) міняти зараз нічого не треба: та модель, що вже в роботі, з ними справляється. Це опція на випадок масової пакетної обробки картинок, де ціна почне мати значення. OpenAI-сумісний ендпоінт означає, що переключення обійдеться рядком конфігу.

misc: Kobo тепер запускає застосунки – 466 балів, Cobalt відкриває читалку для стороннього коду · Клодетт: «змусь Клода не говорити як стаття з BuzzFeed» – 225 балів, четвертий за добу матеріал про машинний стиль · випадково залогнував сотні тисяч дзвінків на військові бази США – 478 балів, перехоплення e164.agra · ЄС: авторське право не захищає ШІ-контент 186 балів · Codex тиждень замість Клода – враження практика

83 бали · \$250 млн при оцінці \$2,3 млрд на дата-центри **на орбіті** - серед інвесторів Nvidia і Cisco, 288 тис. переглядів · @lennysan: звичка питати «а може ШІ це зробити?» **перед** тим, як робити · @paulg: у капіталізмі виграють ті, хто робить життя клієнтів кращим. Не ті, хто вичавлює копійки - «2× проти 10×» · Еха найняла **CTO Robinhood** і Lyft · 100 тис. годин відео людських рук відкрито для робототехніки, 15 тис. задач