AI-ДАЙДЖЕСТ

Хассабіс залишає CEO DeepMind

Джефф Дін і ще троє йдуть з ним і засновують нову компанію, пост зібрав 547 балів на HN

четвер, 6 серпня 2026

У ЦЬОМУ ВИПУСКУ

ТЕМА 1

🔥 Землетрус у Google DeepMind: Хассабіс іде з CEO, а Джефф Дін забирає з собою ще трьох і засновує компанію

Найгучніша подія доби, і три окремі історії на HN про одну й ту саму подію: **547 балів** (пост Google), **652** (сайт нової компанії, топ-1 дня) і **297** (NYT). Першоджерела - лист Сундара Пічаї співробітникам і сайт нової компанії.

Що сталось насправді, по пунктах з листа:

• **Деміс Хассабіс** іде з посади CEO Google DeepMind і стає **Chair GDM + Chief Scientist Alphabet**, лишаючись на чолі Isomorphic Labs. Формулювання Пічаї: Деміс «описував нас як таких, що стоять у передгір'ях сингулярності», і йому потрібна роль, де він зможе «зосередити всю увагу на формуванні майбутнього AGI»

• Операційне керівництво GDM бере **Корай Кавукчуоглу** (13 років у DeepMind, стояв за WaveNet і DQN) як SVP, з підпорядкуванням прямо Пічаї • **Джефф Дін іде після 27 років** у Google

Але головне - на сайті нової компанії. Пічаї згадує двох: Дін і Санджай Гемават. На `discoveryloop.com` у складі засновників **чотири** прізвища: **Джефф Дін, Санджай Гемават, Куок Ле і Оріол Виньялс**. Тобто з Alphabet виходять ще й **Оріол Виньялс** (співавтор Gemini, AlphaStar) і **Куок Ле**, це вже вихід ядра. Самоопис команди: «троє з найбільш цитованих дослідників у ШІ і двоє з найбільш цитованих у розподілених системах».

Компанія - **Discovery Loop**, public benefit corporation. Місія дослівно: автоматизувати експериментальні цикли в ML, науці й інженерії, «паралельне виконання тисяч експериментів». Починають з автоматизації **самих ML-досліджень** і прямо пишуть, що будуть «власним першим клієнтом». Google лишається **інвестором-засновником і хмарним партнером**.

Цифри Google з того ж листа: **Gemini-застосунок 950 млн+ місячних користувачів**, **Gemma 900 млн+ завантажень**, і згадка про **Gemini 4** як про те, що вже в роботі.

Чому це важливо. Найгостріша репліка дня - від @blader: «є рівно Нуль вінчурних інвесторів у долині, яким потрібно було побачити презентацію від Джеффа Діна». Але важливіше інше. Discovery Loop будує рівно те, про що був вчорашній пункт 4 (Ліліан Венг, harness engineering): **автоматизований цикл самополіпшення**, тільки на рівні компанії й з людьми, які будували TPU і MapReduce. Вчора Венг описувала цикл, який можна автоматизувати; сьогодні під це підняли компанію. [proven - обидва першоджерела відкрито напряму: лист Пічаї і сайт Discovery Loop; розбіжність у складі засновників між ними перевірена]

лист Пічаї і Хассабіса · discoveryloop.com · анонс Джеффа Діна · HN про зміни, 547 · HN про Discovery Loop, 652

ТЕМА 2

Atlassian Rovo зливає Jira і Confluence через prompt injection, і Atlassian мовчить два з половиною місяці

196 балів. PromptArmor показав zero-click витік даних з **Atlassian Rovo**, AI-агента, що ходить по всьому стеку Jira/Confluence.

Механіка, дослівно з розбору: користувач завантажує в Rovo файл із прихованою ін'єкцією (звичайна річ: «знайшов документ в інтернеті, кинув агенту»), просить розібрати Jira-тікети, і ін'єкція змушує агента **дописати чутливі дані до URL атакувальника**. Інструмент відкриття URL у Rovo не має жодного захисту від переходу за посиланням, яке агент **сам щойно склав**. Сервер атакувальника логує запит разом із даними.

Два деталі, від яких неприємно:

• Атака працює, **навіть якщо в організації вимкнено вебпошук**: налаштування прибирає пошук, але **не прибирає інструмент відкриття результатів • Слідів не**

лишається: якщо користувач пізніше відкриє той самий чат, він побачить звичайні пропозиції по тікетах, а доказів атаки не буде

Хронологія розкриття: PromptArmor повідомив Atlassian **23 травня**. Ті присвоїли номер справи, подякували, і **за два з половиною місяці й кілька нагадувань більше не вийшли на зв'язок. Уразливість жива**, тому й публікують.

Чому це важливо. Третій день поспіль головна безпекова тема - **агент, що робить зайве через довірений вхід** (04.08 - агент AISI соціально інженерив мейнтейнера, 05.08 - npm-черв'як у конфізі агента, сьогодні - Rovo). Спільний знаменник: **інструмент, який агент може викликати з параметром, що прийшов ззовні**. Практичний висновок: небезпечний **тул, що приймає URL або шлях, зліплений з тексту, який щойно було прочитано**. Все, що приходить з месенджера, з листів, з чужих документів, це дані. Rovo показує, що правило в промпті не рятує, якщо тул технічно дозволяє довільний адресат. [proven - детальний розбір з хронологією розкриття прочитано повністю]

розбір PromptArmor · HN-тред

ТЕМА 3

Rust офіційно заборонив LLM-код у ядрі мови, і аргумент там не про якість коду

109 балів, рідкісний випадок, коли політику по ШІ пише спільнота, яка тримає одну з найважливіших мов сучасності. Автор - Jynn Nelson, політику ухвалили **п'ять команд** проекту для монорепо `rust-lang/rust`.

Правило в одному рядку: **LLM можна для питань, аналізу, дистиляції, перевірки, рев'ю, але не для створення.** (Виняток є: заздалегідь узгоджені, некритичні, добре протестовані зміни з **обов'язковим розкриттям**, що це LLM.)

Цікаве - обґрунтування, дослівно з посту:

• **«Відшліфований технічний продукт більше не свідчить про зусилля й розуміння».** Раніше вилизаний PR означав, що на тому кінці людина, яка вклала час. Тепер, у разі автономного агента, «**на тому кінці більше нікого нема**»

• **Легкість написання коду добиває вузьке місце рев'ю.** Цифра, яку варто запам'ятати: **1 281 відкритий PR** у `rust-lang/rust` просто зараз. І головне: «більша частина роботи рев'юера - вирішувати, **чи цей напрямок узагалі хороша ідея**»

• Механічне копіювання туди-сюди між LLM і трекером марнує час тих, хто читає

Чому це важливо. Прямий місток до вчорашнього пункту 3 (ACM Queue: кодування займає лише ~15% часу розробника, тиск переноситься нижче за течією, на рев'ю). Вчора був рецензований механізм, сьогодні - **найбільший опенсорс-проєкт, який упирається в цю стінку і пише регламент**. Висновок практичний і незручний: правило «пишемо швидше з ШІ» без правила «хто і як це рев'юить» просто переносить чергу на

людей. У Rust ця черга вже має номер: 1 281. [proven - офіційний блог Rust, цифри й формулювання з першоджерела]

політика в Inside Rust · HN-тред

ТЕМА 4

Cloudflare відкрила код своєї «ОС для агентів», і це буквально архітектура харнеса, тільки для компанії

501 бал. Cloudflare виклала у відкритий доступ **Cloudflare OS**, платформу, де кожен співробітник отримує агента і робочий простір, «побудовані навколо його компанії: як вона працює, що вона знає і на які системи спирається».

У **травні** вони дали першу версію **всім у Cloudflare**, і, за їхніми словами, «тисячі людей у кожній функції, багато з них поза інженерією» користуються нею щодня. Сьогодні відкрили перероблену версію, будь-яка організація може розгорнути її в себе.

Ключова частина - **спільна бібліотека контексту й навичок**. Дослівно: «Вона фіксує термінологію, процедури й найкращі відомі способи робити повторювану роботу **як інструкції, яких може дотримуватись агент**. Коли одна людина знаходить кращий спосіб щось зробити, ним одразу можуть користуватись усі».

Чесно про те, що в них не спрацювало в першій версії: застосунки були статичні, не живе ПЗ. Детерміновані задачі все одно вимагали запускати інструкцію заново і палити токени. І головне, **доступ до MCP-сервера казав, які тули агент може викликати, але не які саме ресурси він при цьому побачив**. Це стало проблемою, щойно люди почали ділитись просторами.

Чому це важливо. Їхні **граблі** цінніші за їхній реліз: пункт про MCP («знаємо тули, не знаємо побачені ресурси») - це рівно та проблема, що в пункті 2 сьогодні вбила Rovo. І окремо їхній висновок «не змушуй кожного пояснювати моделі ту саму процедуру заново»: це аргумент за куровану бібліотеку інструкцій замість того, щоб щоразу пояснювати в чаті, як робити. [proven - офіційний блог, прочитано повністю; це реліз відкритого коду. Не заявка]

анонс Cloudflare OS · HN-тред, 501

ТЕМА 5

Відкрита модель на 4B побила GPT-5.6 Sol на пошуку і коштує в 100 разів менше

251 бал. Neon (це Postgres, тепер у складі Databricks) разом зі стартапом Castform показали замір: **відкрита модель на 4 млрд параметрів після RL-донавчання шукає по базі так само точно, як GPT-5.6 Sol, коштуючи в 100 разів менше**.

Цифра, заради якої це варте уваги: типовий багатокроковий пошуковий запит на `gpt-5.6-sol` займає **понад 10 секунд і коштує ~$0.03** end-to-end. Причина: агентний пошук перестав бути одним запитом. Модель планує і шукає в циклі, і **кожна ітерація циклу - це ще один виклик фронтир-моделі**.

Логіка авторів проста: маленькі відкриті моделі дешевші на два порядки, але з коробки слабші; RL-донавчання під **конкретну задачу** (пошук) цей розрив закриває. Meta Castform дослівно - «зробити пост-тренінг таким же доступним, як промпт-інжиніринг».

Найтверезіша репліка з треду: «харнес мав би піднімати субагента і віддавати вузькі задачі спеціалізованим моделям, Claude Code частково це вже робить, віддаючи роботу explore-агента Haiku».

Чому це важливо. Прямо в тему вчорашнього пункту 4 і сьогоднішнього пункту 1: інтелект дістають з обв'язки. Практичний бік: $0.03 за пошуковий запит складається в гроші тільки на великому обсязі, і локальний BM25 по невеликому корпусу лишається дешевшим за будь-яку модель. Це орієнтир вартості агентного пошуку, якщо постане питання шукати по чомусь набагато більшому. [proven - цифри з технічного посту Neon; це їхній власний замір і їхній продукт, тобто джерело зацікавлене]

розбір Neon + Castform · HN-тред

ТЕМА 6

Александр Ванг: рій агентів у Meta робить більше за команду зі 100 інженерів. Цифру дав, доказів немає

Пост зібрав 75 тис. переглядів. Головний AI-офіцер Meta **Александр Ванг** на YC Startup School 2026 у розмові з Гаррі Таном сказав дослівно:

> *«Всередині Meta спостерігали випадки, коли, якщо побудований правильний агентний цикл і є правильна система оцінювання та метрика, яку агенти оптимізують, **рій агентів може досягти більшого, ніж команда зі 100 інженерів**. Вони роблять це дуже впевнено, насправді дуже легко».*

Заява не береться на віру. Вона складається з трьох умов («правильний цикл», «правильна система оцінювання», «правильна метрика»), і жодна з них не розкрита. Нема ні задачі, ні бенчмарка, ні заміру. Це рівно той жанр, який учорашня стаття ACM Queue розносила в міфі №3: **метрика, що провокує гру в систему**. У сьогоднішньому пункті 3 Rust каже протилежне і з цифрою: вузьке місце в тому, що **хтось має вирішити, чи це взагалі хороша ідея**, а рій агентів цю роботу не робить, він її створює.

Чому це важливо. Це маркер жанру, не факт. Коли цю цитату принесуть як аргумент (а принесуть, вона з YC-сцени, від імені Meta), контрпитання таке: на якій задачі, за якою метрикою і хто рев'юив результат. У самій ідеї «рій агентів + система оцінювання» нічого дурного нема, це буквально те, що Discovery Loop з пункту 1 збирається робити

всерйоз. Різниця в тому, що вони описали механізм, а тут анекдот зі сцени. [fuzzy - цитата дослівна і перевірена, але за самою заявою нема жодного оприлюдненого заміру]

цитата з YC Startup School

ТЕМА 7

levelsio спалив $900 на агентному циклі й викинув 95% результату

150 тис. переглядів, 964 лайки, і найкорисніший антидот до пункту 6 того самого дня.

П'єтер Левелс (той, що робить продукти соло і живе з них) написав про свій досвід з «Gauntlet Loop»: «Щоразу, коли робиться Gauntlet Loop, виходить **тотальний безлад і хаос з неоптимального коду**, де забагато всього відбувається і нічого не працює як слід. І спалюється $500. Доводиться все прибирати руками й повертатись до того, що було».

Через півтори години в уточненні: «**Насправді спалено $900**, і це стало повним безладом. Довелось прибрати **95% того, що воно наробило**, і повернутись до того, що було!»

Висновок про власний робочий процес: єдиний спосіб кодити з ШІ - **крок за кроком**, без довгого автономного циклу.

Чому це важливо. Йдеться про **режим використання**. Довгий автономний прогін без звірки проміжного результату дає найдорожчі помилки; зупинка й перевірка артефакту їх ловить. Levelsio платить за цей урок готівкою: $900 за один прогін. [proven - це його власний досвід з конкретною сумою, обидва пости прочитані повністю]

пост levelsio · уточнення про $900

ТЕМА 8

Zed зробив систему контролю версій, де кожен рядок пов'язаний з розмовою, що його породила

340 балів. Zed анонсував **DeltaDB**, окрему систему контролю версій, побудовану під те, що код тепер пишуть агенти.

Головна ідея в їхньому формулюванні: «**Софт робиться між комітами**», тобто вся цікава робота відбувається в проміжку, який git не бачить. Що вміє DeltaDB:

• **Перемотка до будь-якої правки**: фіксує кожну операцію між комітами, у кожної стабільний ідентифікатор • **Зв'язок коду з розмовою**: «з будь-якого рядка коду знайти розмову; з будь-якого повідомлення стрибнути до коду, який воно змінило»

• **Гілкування з будь-якого моменту**, зокрема **посеред запуску агента**: worktree віртуалізований, тому нова гілка коштує майже нічого • Колега може підключитись, **поки робота ще йде**, і поговорити з агентом, який її робив, не чекаючи коміту й пуша

З треду, важливе уточнення: працює **з будь-яким агентом через ACP**, тобто Codex і Claude Code теж, не лише власний агент Zed. Поки що **early access за запитом**.

Чому це важливо. Зв'язок «коміт і розмова, що його спричинила» зазвичай тримається на тому, що людина руками напише в повідомленні коміту, тобто по пам'яті й вибірково. DeltaDB робить цей зв'язок властивістю сховища. Це другий день поспіль, коли інструментарій тицяє в те саме місце. [promising - early access, живих відгуків ще нема, все з їхнього анонсу]

анонс DeltaDB · HN-тред

ТЕМА 9

TIME віддає ШІ-ботам іншу версію сайту, з рекламою всередині тексту

234 бали. Вінсент Шмальбах виявив, що **TIME** віддає краулерам ШІ **інший HTML**, ніж людям, і в цю версію вшита реклама, так, щоб вона потрапила в контекст моделі й спливла у відповіді користувачу.

Найтверезіші репліки з треду:

• «Звучить як звичайний старий prompt injection. ChatGPT зараз може переглянути 50 сторінок, коли просять дослідити тему. Цілком імовірно, що фінальна відповідь зазнає впливу такої реклами»

• І найточніше: «Найкращий спосіб для політика збрехати - переконати в брехні когось іншого й поставити його перед камерами. **LLM за своєю природою легковірні**»

Плюс окрема здорова підозра: це може бути ще й спосіб для агенції накрутити цифри показів, які вона звітує клієнту.

Чому це важливо. Будь-який конвеєр, що читає чужі сайти через браузерну автоматизацію або `curl`, отримує те, що віддає сервер, а сервер тепер може знати, що це не людина, і підсунути інше. Правило «цифри брати з першоджерела» лишається, але до нього треба дописати: першоджерело вміє бути персоналізованим під бота. Захист той самий: перехресна перевірка кількох джерел і відкриття сторінок залогіненим справжнім браузером. [proven - тест відтворюваний: різний HTML на різний User-Agent]

розбір Шмальбаха · HN-тред

ТЕМА 10

Чому задачі Ердеша почали падати перед ШІ

130 балів, Quanta Magazine. Легендарні задачі Пала Ердеша, той самий жанр математичних проблем, які десятиліттями стояли, почали розв'язуватись за допомогою ШІ, і Quanta розбирає, чому саме зараз.

І в той самий день, у той самий фід, лягає протилежна робота: **«Position: LLMs Can't Jump»** (256 балів, OpenReview), позиційна стаття про те, що LLM не роблять **стрибків інтуїції**. Найкраща репліка з треду, напівжартом: запропонований експеримент - навчити LLM на всіх текстах **до 1980-90 року** і подивитись, чи зможе вона дійти до винаходу самої себе.

Чому це важливо. Ці дві теми подаються разом свідомо, бо поодинці кожна бреше. «ШІ розв'язує задачі Ердеша» без другої половини звучить як «інтуїція взята»; «LLM не вміють стрибати» без першої - як «нічого не працює». Правда посередині і вона операційна: **там, де є вимірюваний критерій успіху і можна перебрати тисячі спроб, машина виграє** (буквально теза Discovery Loop з пункту 1 і Castform з пункту 5). Там, де треба один нетривіальний стрибок без зворотного зв'язку, ні. Висновок той самий, що й учора у пункті 8 про табличні дані: важливо знати, у якому з двох режимів працює інструмент. [promising - Quanta це якісна науково-популярна журналістика, друга робота - позиційна стаття, тобто аргумент. Не експеримент]

Quanta про задачі Ердеша · HN-тред · «LLMs Can't Jump» · HN-тред

misc - коротко, що ще варте ока

• **@blader поставив дочці ліміт екранного часу, вона попросила Claude Code зробити їй власний браузер.** 406 тис. переглядів, 7404 лайки, найсмішніше за добу: 11-річна дочка обійшла обмеження на YouTube, завайбкодивши браузер на Electron і Chromium. В окремому пості він визнає, що сам винен: пообіцяв необмежений екран за кожен успішний злам. Серед попередніх - вона навчилась **апрувити власні запити на екранний час**

• **@garrytan: «Make something agents want»** - 637 тис. переглядів і 3218 закладок на чотири слова. Переписаний девіз YC («make something people want») під нову реальність. Як індикатор, куди дивиться найбільший акселератор світу, вартує уваги більше за багато довгих есеїв

• **@emollick помітив суперечність:** моделі стають кращими у виконанні складних інструкцій, але водночас більше «**користуються судженням**» про те, на яких частинах інструкції зосередитись, а які приглушити. Висновок: «Великі наслідки для навичок, які можуть стати **пропозиціями. Не наказами**»

• **@emollick про фінансові поради:** дослідження MIT і Стенфорда каже, що **більшості людей було б фінансово краще, якби вони слухали поради LLM** (тестували GPT-5.2 і Gemini 3 Flash). З важливим застереженням: якість поради сильно залежить від того, **які питання ставиш**. Читається з поправкою на вчорашній пункт 8: прогнози по числах питати не варто

• **@gdb: аншлаг на доповіді OpenAI про інцидент з Hugging Face** на Black Hat. Це та сама історія, яку Малєєв драматизував як «Повстання OpenAI» і яка потім розсипалась; тепер її розбирають на головній безпековій конференції

• **@dakshgup: Greptile v5** - переписаний з нуля агент код-рев'ю, «ловить більше багів, з вищою точністю, і вдвічі швидше (~2 хвилини на рев'ю)». Прямо в тему пункту 3: якщо вузьке місце - рев'ю, то інструменти туди і поїдуть

• **@levie: «99% токенів у світі будуть спожиті в корпоративному контексті»** - код, дослідження в науках про життя, автоматизація виробництва, безпека, виявлення шахрайства. Тверезий контрапункт до споживчого хайпу

• **@EricTopol - не про ШІ:** тирзепатид (Zepbound) проти ситагліптину в людей з діабетом і попереднім атеросклеротичним ССЗ: зниження серцево-судинних подій, а також **інфекцій на 36% і загальної смертності на 45%**. Публікація в BMJ. Подається як сигнал по класу GLP-1, не як рекомендація

• **@lennysan - продовження вчорашнього:** вчора йшла новина, що YouTube знищив його канал за «видавання себе за себе». Сьогодні він написав розгорнуто: апеляцію подано, каналу досі нема. З його ж репліки: «з одного боку, добре, що вони серйозно ставляться до проблеми; з іншого - можливо, є спосіб делікатніший, ніж закривати канал»
