AI-ДАЙДЖЕСТ

METR: Anthropic розвиває моделі вдвічі швидше

METR: Anthropic покращує свої моделі вдвічі швидше, ніж рік тому, і переважно завдяки AI.

субота, 7 листопада 2026

У ЦЬОМУ ВИПУСКУ

1. METR: Anthropic розвиває свої моделі приблизно вдвічі швидше, і більшу частину прискорення дає AI
2. Anthropic: AI може взяти 27-52% часу науковця вже з теперішніми моделями
3. DeepMind: AI-агенти за тижні знайшли дві слабкості в RSA-1024
4. IPO Anthropic: покупці говорять уже про $1,1-1,3 трлн
5. Anthropic про пам'ять агентів: «сни» виправляють нотатки
6. Tesla закриває виробництво Cybertruck
7. Chainalysis: нападники отримали дані входу всіх клієнтів
8. Hightouch: дешева модель плюс Opus для складних випадків дає точність Opus
9. Frontier Model Forum: спільна шкала для оцінки джейлбрейків у кібербезпеці
10. Конгрес викликає Дженсена Хуанга на закритий брифінг

ТЕМА 1

METR: Anthropic розвиває свої моделі приблизно вдвічі швидше, і більшу частину прискорення дає AI

ДЖЕРЕЛА METR, 05.11 Metr · Ajeya Cotra, 06.11 Planned-obsolescence · HN, 06.11 Hacker News · Arjun Panickssery, 06.11 Lesswrong · HN, 06.11 Hacker News

З 12.08 по 29.10 співробітники METR працювали всередині Anthropic. Вони мали доступ до внутрішніх метрик претренінгу, RL, бенчмарків внутрішніх моделей і до інтерв'ю з людьми. Питання було одне: наскільки AI вже прискорює розробку самого AI.

Головна цифра METR така. З жовтня 2025 року внутрішній індекс можливостей Anthropic (AECI, версія Epoch Capabilities Index) росте в 1,9 раза швидше, ніж до того. Довірчий інтервал 95% становить від 1,5 до 2,8 раза. Тобто за рік лаба пройшла приблизно два роки колишнього прогресу. METR певний, що причина не в швидшому нарощуванні обчислень для тренування. Більшу частину прискорення сьогодні METR пояснює тим, що Anthropic широко використовує AI у власних дослідженнях. Вузьких місць, які зупинили б прогрес у наступні 12 місяців, METR не побачив.

У звіті є важливе застереження. Anthropic мала право редагувати непублічні дані і вирізала більшість доказів, які METR вважав найпереконливішими. Компанія пояснила це тим, що ці дані могли б прискорити інших розробників. Тому публіка має покладатися на судження METR більше, ніж хотів би сам METR.

Аджея Котра з METR розвиває цифру далі, від свого імені. Якщо загальний прогрес іде вдвічі швидше, то праця дослідників мусить прискорюватись значно сильніше: за її оцінкою, приблизно в 10 разів. Співробітники Anthropic в опитуваннях кажуть про прискорення близько 12 разів. Котра пише, що Anthropic у документах до IPO вже визнала: компанія «may imminently meet» поріг «dramatically accelerated progress» зі своєї політики безпеки.

Скептична позиція теж є. Арджун Панікссері на LessWrong описує кілька своїх проєктів з агентами: рушій для Scrabble, бот для Ultimate Tic-Tac-Toe, формалізацію доказів з філософії. Його висновок: агент добре дотискає задачі зі швидким зворотним зв'язком, але погано обирає напрям. Агент пропускає головну статтю в огляді літератури і ставить експерименти, які нічого не можуть показати.

Чому це важливо: це перший незалежний замір того, про що лаби досі говорили лише самі. Дві позиції не суперечать одна одній. Прискорення реальне там, де результат можна швидко перевірити, а більшість роботи над моделлю якраз така. Для решти галузей висновок практичний: швидкість тепер залежить від того, наскільки задачу можна оцінити автоматично.

ТЕМА 2

Anthropic: AI може взяти 27-52% часу науковця вже з теперішніми моделями

ДЖЕРЕЛА Anthropic, 05.11 Anthropic · @emollick, 07.11 @emollick · HN, 06.11 Hacker News

Економісти Anthropic проаналізували 207 000 дослідницьких сесій у Claude і Claude Code. Аналіз даних займає приблизно третину сесій у Claude і майже половину в Claude Code. Для ідей і збору даних науковці використовують Claude менше, хоча в опитуванні саме ці задачі вони назвали найважливішими. Найбільше Claude використовують інформатика, математика і статистика. Найменше використовують медицина і лабораторні науки.

Далі автори будують модель. Якщо теперішні можливості впровадити широко, AI може взяти 27-52% часу дослідника на одну роботу, медіана 38%. Це дає прискорення від 1,35 раза в мистецтві до 1,96 раза в інформатиці і 2,02 раза в статистиці. У гіпотетичному сценарії, де AI робить усю нефізичну роботу, вихід в інформатиці і статистиці зростає до 15 разів. У хімії і біології ріст обмежений лабораторією: у 3-4 рази. Близько 40% опитаних науковців, які вже працюють з AI, чекають, що до 2030 року AI сам вестиме повні дослідницькі проєкти.

Чому це важливо: цифри цього звіту і пункт 1 описують одне явище з двох боків. METR бачить прискорення в одній лабі, де майже вся робота когнітивна. Anthropic показує, що в науці загалом межу ставлять фізичні експерименти. Прогноз «AI прискорить науку в N разів» без розбивки на поля мало що каже.

ТЕМА 3

DeepMind: AI-агенти за тижні знайшли дві слабкості в RSA-1024

ДЖЕРЕЛА @pushmeet, 06.11 @pushmeet · @GoogleDeepMind, 06.11 @GoogleDeepMind · стаття про підписи Googleapis · стаття про факторизацію Googleapis · The Wall Street Journal, 06.11 WSJ

Команда DeepMind Science разом з криптографами Надею Хенінгер (UC San Diego і Inria) та Еммануелем Томе (Inria)

випустила дві роботи. Агенти працювали під наглядом людей. Пушміт Колі пише, що такі результати раніше займали роки, а тут зайняли тижні.

Перший результат стосується цифрових підписів. Перед підписом стандарти доповнюють повідомлення фіксованими даними, padding. Агенти показали: коли нападник контролює приблизно половину блоку або більше, padding не захищає взагалі. На ключі з уже розв'язаного академічного бенчмарку атака пройшла з витратою CPU у 3,5 раза меншою. Для випадку з коротким 32-байтним відбитком новий алгоритм, за теоретичною оцінкою, кращий за попередній приблизно в 1 600 разів.

Другий результат стосується факторизації багатьох ключів однакової довжини. Нападник платить один раз за велике спільне обчислення, а потім кожен наступний ключ RSA-1024 обходиться приблизно в 30 разів дешевше, ніж з попередніми пакетними методами. Читачі в X одразу уточнили: 30 разів стосуються ціни додаткового ключа в пакеті.

Ціна зламу одного окремого ключа не падає в 30 разів.

RSA-2048 і довші ключі ці результати не зачіпають. WSJ виносить у заголовок ширшу думку: AI може зламати старе шифрування раніше, ніж квантові комп'ютери.

Чому це важливо: криптографія спирається на припущення, що певна задача важка, і запас міцності досі змінювався роками. Ці роботи показують, що запас тепер може змінитись за тижні. Системам, де ще живуть ключі RSA-1024 або старі схеми підпису, пора переходити на 2048 біт і більше, навіть без квантового комп'ютера.

ТЕМА 4

IPO Anthropic: покупці говорять уже про $1,1-1,3 трлн

ДЖЕРЕЛА Bloomberg, 06.11 Bloomberg · HN, 06.11 Hacker News · Financial Times, 06.11 FT · BBC, 06.11 BBC

Вчора тут писали, що великі інвестори сумніваються в оцінці $1,5 трлн. Сьогодні Bloomberg дає нові цифри від джерел. Потенційні покупці обговорюють оцінку $1,2-1,3

трлн, а дехто говорить і про $1,1 трлн. До обвалу очікування були $1,4-1,5 трлн, а на початку жовтня доходили до $2 трлн. Формальний маркетинг розміщення стартує наступного тижня. Anthropic коментар не дала.

Bloomberg рахує масштаб четверга: Nasdaq 100 упав на 6,4% і втратив $2,3 трлн вартості, це рекорд за один день.

Стратег HSBC Макс Кеттнер вважає, що оцінка близько $1 трлн зменшила б ризик різких коливань після старту торгів.

FT у заголовку пише інше: Anthropic не відступає від цілі $1,5 трлн і готує окрему пропозицію для роздрібних інвесторів у Британії.

У п'ятницю ринок частково відіграв падіння. За BBC, Nasdaq Composite закрився на 1,1% вище, S&P 500 на 0,3%. Азія продовжила падати: Kospi у Кореї втратив 6%, SK Hynix понад 10%.

Чому це важливо: від $2 трлн до $1,1 трлн за місяць зникло майже стільки, скільки коштувала вся Anthropic у травні ($965 млрд).

IPO першої лаби frontier-моделей стане ціною для всієї галузі, тому ця ціна важлива і для OpenAI, яка залучає кошти за оцінкою $1,25 трлн. Пункт 1 додає контекст: інвестори оцінюють компанію, яка, за METR, сама прискорює свій розвиток.

ТЕМА 5

Anthropic про пам'ять агентів: «сни» виправляють нотатки

ДЖЕРЕЛА Anthropic, 06.11 Anthropic · @AnthropicAI, 06.11 @AnthropicAI

Інженери Anthropic описали, як Claude веде пам'ять між сесіями. Пам'ять тут означає файли нотаток, які агент пише сам і читає в наступній сесії. «Dreaming» означає окремий прохід без користувача: агент перечитує логи, виправляє помилкові нотатки і будує нові структури.

Найнаочніший приклад дає Pokémon Red. Sonnet 3.7 записував у пам'ять усе підряд: 413 файлів у 225 папках, один значок за 10 000 кроків. Fable 5 мав дев'ять файлів у двох папках і один загальний `global.md`, і заробив чотири значки. Fable 5 не записував те, що видно на екрані, і зберігав лише стратегії та перевірені правила.

Жорсткі правила шкодять сильним моделям. Коли Opus 5 і Opus 4.7 отримали правило «записуй лише помилки», вони втратили більшу частину виграшу від пам'яті. Старшим моделям, навпаки, підказки допомагали: Sonnet 4.6 з інструкцією «читай пам'ять перед відповіддю» майже подвоїв результат.

Окремий тест стосується вікі. Claude будував вікі за 350 000 файлів (близько 420 млн токенів) компанії Stainless.

За 128 сесій «снів» точність на 25 питаннях зросла з 13% до 48%. Пошук разом з вікі дав 61% проти 53% лише з пошуком і зекономив близько 20 000 токенів на питання.

Чому це важливо: головний урок простий. Агенту треба пояснити, для чого пам'ять, і не диктувати, що записувати. Друге: пам'ять старіє, і без перевірки проти логів хибна нотатка повторює ту саму помилку щоразу.

Для будь-якої системи з довгою пам'яттю регулярна звірка нотаток з першоджерелом коштує дешевше за наслідки.

ТЕМА 6

Tesla закриває виробництво Cybertruck

ДЖЕРЕЛА Reuters через NBC News, 06.11 Nbcnews · The Guardian, 06.11 Guardian · Ars Technica, 06.11 Ars Technica · HN, 06.11 Hacker News

Ілон Маск написав у X, що Tesla згорне виробництво Cybertruck на заводі в Техасі приблизно до середини 2027 року.

Звільнені потужності підуть під Cybercab, робота-таксі компанії. Маск назвав Cybertruck «найкращою машиною Tesla» і «кінцем епохи».

Цифри продажів пояснюють рішення. За Kelley Blue Book, торік Tesla продала в США 20 237 Cybertruck, на 48% менше, ніж 38 965 у 2024 році. Перед запуском Маск обіцяв 250 000 на рік. Тред на HN зібрав 903 бали і 1 753 коментарі.

Чому це важливо: Tesla прибирає продукт заради ставки на автономність і роботів. Того ж тижня ринок продавав усе, що пов'язане з AI, тому Tesla подвоює ставку саме тоді, коли інвестори сумніваються в ній найбільше.

ТЕМА 7

Chainalysis: нападники отримали дані входу всіх клієнтів

ДЖЕРЕЛА Ars Technica, 06.11 Ars Technica `[одне джерело]`

Компанія з аналізу блокчейну Chainalysis повідомила, що нападники отримали імена, логіни, хеші паролів і живі токени всіх її користувачів. Серед клієнтів компанії є ФБР, ICE, ВПС США, Microsoft, Coinbase і Barclays.

Chainalysis описує атаку як «multifront». Нападник використав техніки з допомогою AI-агентів і on-path атаку, тобто перехоплення трафіку на проміжній інфраструктурі. Де саме сталося перехоплення, компанія не каже. Розслідування знайшло витік у понеділок, а повідомила компанія в середу. Скидання облікових даних обов'язкове. Доступ до сервісів Chainalysis повертає поступово, через список дозволених IP-адрес.

Чому це важливо: живі токени небезпечніші за хеші паролів, бо дають доступ без зламу. Клієнти такого сервісу самі є цілями, тому витік у постачальника дає нападнику список і ключі одночасно. Це ще один випадок, де компанія прямо згадує AI-агентів в атаці.

ТЕМА 8

Hightouch: дешева модель плюс Opus для складних випадків дає точність Opus

ДЖЕРЕЛА Hightouch, 04.11 Hightouch · HN, 06.11 Hacker News

Hightouch перевіряє, чи правильно система злила записи клієнтів в одну особу. Людина перевіряє близько 50 таких профілів на годину, а для точної оцінки потрібно понад 1 000. Команда порівняла Claude Opus 5.5 з двома «моделями рішень», Jev і GPT-6 Luna. Такі моделі повертають лише ймовірність від 0 до 1.

Самостійно моделі рішень коштують $0,05-0,12 за 1 000 профілів і відповідають за 0,15-0,20 с. Opus коштує $3,86-5,39 і відповідає за 2,2-2,6 с. На складному тесті Jev має 89,1%, Opus 95,1%. Потім команда зробила гібрид:

модель рішень віддає Opus лише випадки, де сама не впевнена.

Тут вирішила калібровка. Jev помиляється здебільшого з низькою впевненістю, тому гібрид з Jev дав 95,3% на складному тесті за $2,94. Luna помиляється впевнено, тому її гібрид дав лише 91,6%. На легкому тесті гібрид Jev дав 98,8% проти 99,0% в Opus за $0,88 замість $3,86.

Чому це важливо: точність дешевої моделі важить менше, ніж те, чи знає вона, коли помиляється. Схема «дешевий суддя, а складне до великої моделі» переноситься на будь-яку масову перевірку. Перед запуском такої схеми перевірте, чи помилки дешевої моделі справді мають низьку впевненість.

ТЕМА 9

Frontier Model Forum: спільна шкала для оцінки джейлбрейків у кібербезпеці

ДЖЕРЕЛА Frontier Model Forum, 05.11 Frontiermodelforum · @OpenAI, 06.11 @OpenAI

Frontier Model Forum випустив версію 1.0 рамки для оцінки джейлбрейків у кібербезпеці. Рамка оцінює кожен обхід захисту за чотирма категоріями: наскільки він збільшує можливості нападника, наскільки широкі ці можливості, як легко його застосувати і як легко його знайти. Мета практична: лаби мають однаково оцінювати однаковий обхід і першими закривати найнебезпечніші. Рамку підтримали Amazon, Anthropic, Google DeepMind, Microsoft і OpenAI.

Чому це важливо: вчора OpenAI показала модель, яка пише експлойти для перевірених клієнтів. Межа між такою моделлю і звичайною тримається на захисті від обходу. Спільна шкала дає змогу порівнювати, у кого цей захист слабший.

ТЕМА 10

Конгрес викликає Дженсена Хуанга на закритий брифінг

ДЖЕРЕЛА Ars Technica (текст FT), 06.11 Ars Technica · The New York Times, 05.11 NYT

Голови трьох комітетів Палати представників (з Китаю, закордонних справ і розвідки) запросили гендиректора Nvidia на закритий брифінг 18.11. Причина: NYT минулого тижня писала, що розвідка США вважає Хуанга ціллю китайської шпигунської операції. Йдеться про Еліс Хуанг, керівницю сінгапурської хмарної компанії Megaspeed, яка купила чипів Nvidia на мільярди доларів. Nvidia відкидає звинувачення і каже, що клієнти отримують лише звичайну ділову інформацію.

Чому це важливо: після виборів Вашингтон став жорсткішим до великих технологічних компаній, і Nvidia тепер під тиском з двох боків: експортний контроль і особиста безпека керівника. Пункт 4 показує, як швидко політика тепер потрапляє в ціну компаній з AI.

---

misc

- **Claude Code і пошта, продовження:** тред про пошту в User-Agent став найпопулярнішою сторі HN за добу, 1 559 балів. Issue має 699 реакцій 👍, коментаря від Anthropic у ньому досі нема. GitHub · HN Hacker News
- **Шахрайство з AI-голосом:** у США висунули обвинувачення трьом чоловікам. Вони голосом і відео з AI видавали себе за керівників і юристів і вкрали понад $129 млн. Guardian
- **libghostty 1.0:** Мітчелл Хашимото випустив стабільну C-бібліотеку терміналу з Ghostty. Нею вже працює термінал у JetBrains IDE, понад 12,5 млн користувачів. Mitchellh · HN Hacker News
- **FAA без дискет:** FAA до кінця 2028 року замінить системи на дискетах у 435 об'єктах управління повітряним рухом. Ars Technica · HN Hacker News
- **Конектори ChatGPT:** розробник описав, як ChatGPT мовчки відправив наступний виклик через інший конектор з одноразовим кодом від першого. Openai · HN Hacker News
- **BioBasicsBench:** @OpenAI разом з Ligo випустила тест зі 100 задач з реальних лабораторних експериментів для агентів. @OpenAI
- **DeepFinance:** @0interestrates випустив тест з 21 задачі рівня PE і інвестбанків на реальних даних угод. @0interestrates
- **AI-спам у відповідях:** @levelsio заблокував понад 10 000 акаунтів з AI-відповідями і обіцяє опублікувати список. @levelsio

---