AI-ДАЙДЖЕСТ

GPT-6 Astra править ваги нейромережі без тренування

Anthropic оновила S-1, Ubuntu досі без патча, хакери гріють холодильники, Marvell купує Lumentum

вівторок, 3 листопада 2026

У ЦЬОМУ ВИПУСКУ

1. GPT-6 Astra читає і пише ваги нейромереж напряму
2. Anthropic оновила S-1: $17 мільярдів за квартал, ріст сповільнився, зобов'язання на $895 мільярдів
3. CVE-2026-98432: Ubuntu досі вразлива, тред «WTF Ubuntu?» набрав 1 375 балів
4. Хакери вдруге нагріли холодильники супермаркетів, тепер у семи країнах Європи
5. OpenAI вчетверо підняла баг-баунті і кличе зовнішню команду ламати свої системи
6. Чипи втратили $3,7 трильйона з 6 жовтня, а Marvell купує Lumentum за $115 мільярдів
7. OpenAI переписала бекенд Git під навантаження від агентів
8. Пітер Левелс переніс Battlefield 1942 у браузер, а Claude тепер відмовляє в клонах ігор
9. Стенфорд перевірив 50 відомих висновків про LLM: 44% уже не працюють
10. Посттренування стає окремим бізнесом: Applied Compute і Trajectory

---

ТЕМА 1

GPT-6 Astra читає і пише ваги нейромереж напряму

ДЖЕРЕЛА блог MIT CSAIL, 02.11 Github · @rikiyatakehi, 02.11 @rikiyatakehi · HN, 02.11 Hacker News

Четверо дослідників з MIT CSAIL, серед них Омар Хаттаб, автор DSPy, перевірили просту річ. Чи може мовна модель зрозуміти нейромережу, якщо їй показати лише числа ваг? Метод вони назвали Weight-Space Programming. Модель не пише цикл тренування і не рахує градієнти. Вона міркує, що обчислює кожна частина мережі, і міняє числа руками.

Результати такі. Astra подивилась на 17 253 ваги і з упевненістю 99% правильно сказала, що мережу тренували на Breakout.

Мережа для гри MinAtar, яку Astra написала з нуля за один прохід, перемогла PPO після 100 мільйонів кроків у всіх трьох іграх. Для машини, що застрягала перед рампою, Astra змінила 3,5% ваг, і успіх виріс з 0% до 100%.

Головний експеримент на Qwen3.5-0.8B. Ця маленька модель часто зациклюється: 93% відповідей на MATH-500 упирались у ліміт 32 768 токенів. Astra знайшла один

напрямок у шарах 15 і 19, який спричиняє петлі, і відняла його з 131 072 ваг.

Це 0,016% моделі. MATH-500 виріс з 7,9% до 48,1%, GSM8K з 13,8% до 67,1%, IFEval з 21% до 62,3%.

Автори чесно показують і провали. На Gemma 3 270M, яка майже не зациклюється, Astra не змогла підняти математику більше ніж на пункт. Висновок самих авторів: правка ваг прибирає перешкоду, що ховала вміння, але нового вміння не додає. Підлабузницьку модель Astra за самими вагами не розпізнала. Вдалось лише тоді, коли їй дали подивитись на патерни уваги.

На HN тред набрав 2 134 бали, це найбільше за добу. Ітан Моллік назвав роботу «pretty big deal», але чекає повторення.

Чому це важливо: досі ваги були територією оптимізаторів типу Adam, бо людина не вміє правити їх руками. Якщо модель уміє знайти і вирізати конкретний дефект, частина тонкого налаштування стає хірургією замість повного перетренування.

Межа теж видна: результат на Qwen це ремонт відомої вади, яку описує сама картка моделі. Незалежного повторення поки нема.

ТЕМА 2

Anthropic оновила S-1: $17 мільярдів за квартал, ріст сповільнився, зобов'язання на $895 мільярдів

ДЖЕРЕЛА CNBC, 02.11 Cnbc · Business Insider, 02.11 Businessinsider · The Wall Street Journal, 02.11 WSJ · @deredleritt3r, 02.11 @deredleritt3r · HN, 02.11 Hacker News

27.10 тут писали про перший S-1 Anthropic з цифрами за пів року. 2 листопада компанія подала оновлений документ з третім кварталом. За дев'ять місяців виручка склала $33,24 мільярда, це приблизно +1 200% до минулого року. За третій квартал вийшло близько $17 мільярдів. Ріст квартал до кварталу впав до 47%, після 144% у другому і 132% у першому кварталі.

Чистий збиток за дев'ять місяців склав $104 мільярди. CNBC пише, що більшість цієї суми дала переоцінка конвертованих нот і похідних зобов'язань. Витрати на обчислення за дев'ять місяців зросли на $14,4 мільярда, тобто на 329%. Загальна сума угод про закупівлю і оренду становить близько $895,4 мільярда. Майже вся вона припадає на обчислення та інфраструктуру.

У новій версії з'явилось і речення про безпеку. Anthropic пише, що за внутрішніми оцінками майбутніх моделей вона може «imminently» досягти порогу своєї політики RSP про «dramatically accelerated progress attributable to AI». Аналітик @deredleritt3r пояснює поріг: удвічі швидший прогрес можливостей, ніж компанія очікувала б без внеску AI.

Business Insider знайшов у документі ще одну цифру. З квітня внутрішнє використання Claude за роздрібними цінами перевищує місячний фонд оплати праці без акцій. Це

оцінка за прайсом, Anthropic не платить ці гроші живими коштами.

Понад 90% коду, доданого у вересні, написав Claude.

WSJ пише, що після обвалу техакцій інституційні інвестори радять оцінку IPO $1,5-1,6 трильйона. Раніше звучали цифри ближче до $2 трильйонів.

Чому це важливо: S-1 тепер дає перші публічні цифри про те, скільки коштує фронтирна модель у масштабі. Ріст лишається величезним, але сповільнення з 144% до 47% ринок побачить раніше за абсолютні суми. А речення про поріг RSP перетворює внутрішню політику безпеки на розкриття ризику для інвестора.

ТЕМА 3

CVE-2026-98432: Ubuntu досі вразлива, тред «WTF Ubuntu?» набрав 1 375 балів

ДЖЕРЕЛА chalmovsky.com, 02.11 Chalmovsky · трекер Ubuntu Ubuntu · @levelsio, 02.11 @levelsio · HN, 02.11 Hacker News · HN, 02.11 Hacker News

Учора тут писали, що Red Hat виправив ядра, а Ubuntu ще ні. Станом на ранок 03.11 трекер Ubuntu досі показує «Vulnerable, work in progress» для 22.04, 24.04, 26.04 і 26.10. Помилку в TCP-коді ядра оцінили на CVSS 9.8, вона живе в коді з 2016 року.

Автор поста «WTF Ubuntu?» нагадує хронологію. Debian випустив виправлення 16 жовтня. Arch і Alpine теж. Ubuntu обіцяє на своєму сайті «timely patches for critical vulnerabilities», але за два з половиною тижні патча нема. Автор пише, що всі сервери на Ubuntu довелось вимкнути: на переїзд забракло часу.

Пост став другим на HN за добу з 1 375 балами і 594 коментарями. Окремий тред про ту саму затримку набрав 381 бал.

@levelsio пише, що його команді, можливо, доведеться перевести всі сервери на Debian.

Чому це важливо: дистрибутив обирають за швидкістю патчів, і тепер цю швидкість видно в днях. Волонтерський Debian виправив проблему на 18 днів раніше за компанію, що продає підписки на безпеку. Для серверу на Ubuntu поради не змінились: мінімум відкритих портів і фільтр nftables, поки не вийде нове ядро.

ТЕМА 4

Хакери вдруге нагріли холодильники супермаркетів, тепер у семи країнах Європи

ДЖЕРЕЛА The Guardian, 02.11 Guardian · The New York Times, 02.11 NYT · The Guardian, 03.11 Guardian · HN, 02.11 Hacker News

13.10 тут писали про атаку на холодильники в сотнях французьких супермаркетів. У ніч на 2 листопада це повторилось у Бельгії, Естонії, Ірландії, Фінляндії, Люксембурзі,

Нідерландах і Швеції. Температура в холодильниках і морозилках піднялась до 50 °C.

The Guardian перелічує постраждалих. Ahold Delhaize повідомила про десятки магазинів у Бельгії й Люксембурзі. SuperValu в Ірландії має 13 уражених магазинів з 223, проблема лише в одній моделі обладнання. Coop у Швеції має шість. Компанія, що обслуговує холодильники Delhaize, каже, що випарники перевели в режим розморожування. Менеджер одного магазину оцінив втрати в €100 000.

NYT додає американський слід. У серпні морозилки на військових базах у США раптом переходили в режим розморожування.

Слідчий відділ армії США підтвердив, що вивчає ці випадки і працює з французькою поліцією. Ніхто відповідальності не взяв.

Викупу ніхто не вимагав, тому експерт SentinelOne бачить у цьому саботаж.

Того ж дня британський NCSC опублікував річний огляд. Агентство вважає «highly likely», що до 2028 року AI-шкідник вдарить по критичній інфраструктурі Британії. Голова NCSC Річард Горн прямо згадує атаки на холодильники як попередження.

Чому це важливо: атака дешева, а шкода миттєва: кілька команд контролеру температури знищують товар за ніч.

Експерт Tributech каже, що тисячі таких контролерів лишались доступними з інтернету і після французької атаки. Будь-яке обладнання з віддаленим керуванням тепер треба рахувати як сервер, з патчами і закритими портами.

ТЕМА 5

OpenAI вчетверо підняла баг-баунті і кличе зовнішню команду ламати свої системи

ДЖЕРЕЛА OpenAI, 02.11 OpenAI · OpenAI, 02.11 OpenAI · @OpenAI, 03.11 @OpenAI

OpenAI готує наступний внутрішній спринт безпеки. Цього разу компанія запрошує дослідників Cathedral Systems. Вони атакуватимуть захист OpenAI разом з її командами і її ж моделями. Нагороди за критичні знахідки в програмі баг-баунті компанія підняла більш ніж учетверо.

Нові ставки діють і для звітів, які ще чекають виплати. За нагороди останніх 30 днів OpenAI доплатить різницю. У минулому спринті OpenAI віддала на пошук і виправлення вразливостей 25% продуктових інженерів. Вони працювали з Codex.

Компанія також виділила $1 мільярд пільгового доступу на пів року для тих, хто захищає критичні сервіси.

Чому це важливо: одна доба дала три сигнали: Ubuntu без патча, холодильники в семи країнах і прогноз NCSC. Логіка OpenAI проста: моделі вже добре шукають вразливості, тому захисник має застосувати їх раніше за нападника. Вища нагорода також змагається з ціною, яку за ту саму знахідку дає чорний ринок.

ТЕМА 6

Чипи втратили $3,7 трильйона з 6 жовтня, а Marvell купує Lumentum за $115 мільярдів

ДЖЕРЕЛА Yahoo Finance, 02.11 Yahoo · The New York Times, 02.11 NYT · CNBC, 02.11 Cnbc · HN, 03.11 Hacker News

Учора тут писали про падіння Intel на 29% за день. У понеділок біржа загалом відновилась: за даними Yahoo Finance, виросли близько 70% акцій S&P 500. Але чипи падали далі й за одну сесію втратили понад $400 мільярдів. Intel впав ще приблизно на 9% і тепер коштує менше половини вартості кінця вересня. З 6 жовтня чипи і пам'ять втратили близько $3,7 трильйона.

Того ж дня Marvell оголосила обмін акціями з Lumentum. NYT оцінює угоду в $115 мільярдів разом з боргом Lumentum, WSJ у заголовку пише $80 мільярдів. Lumentum робить лазери й оптичні компоненти. Ставка угоди: у датацентрах для AI мідні з'єднання між стійками замінить світло поруч із чипом. Аналітик Rosenblatt чекає антимонопольної перевірки на 12-18 місяців, зокрема в Китаї.

Oracle подала документи на продаж $10 мільярдів в акціях і $10 мільярдів в облігаціях, і її акції впали на 8,8%. CNBC пише, що борг компанії становить $125,34 мільярда. Спред п'ятирічних CDS Oracle виріс на 265% за рік і тепер ближчий до рівня сміттєвих облігацій.

Чому це важливо: ринок розділяє AI-історію на частини. Софт і більшість S&P відновились, а все, що тримається на позиках під майбутній попит, дешевшає. Угода Marvell і Lumentum показує, куди ставлять навіть у такий тиждень:

гроші йдуть у мережу між чипами.

ТЕМА 7

OpenAI переписала бекенд Git під навантаження від агентів

ДЖЕРЕЛА OpenAI, 02.11 OpenAI · @gdb, 03.11 @gdb · HN, 02.11 Hacker News

OpenAI розповіла про Pando, новий бекенд для свого монорепо. Причина в Codex: один інженер тепер має кілька змін у CI одночасно, і кожен пуш запускає кілька задач, що тягнуть той самий коміт. З кінця минулого року надійність Git у компанії часом падала нижче 90%.

Pando лишається сумісним з протоколом і клієнтом Git, тому інженери й агенти нічого не міняли. Сервер обробляє кожну гілку окремо, і пуші в різні гілки не чекають один одного. Ціна такого рішення: клієнт, що читає кілька гілок, не бачить їх в один момент часу. Ще одне рішення: сервер віддає різним клієнтам ті самі готові packfile замість окремого для кожного.

Медіанна затримка пушу впала майже вп'ятеро. Pando тепер віддає майже петабайт даних Git на добу. Ґреґ Брокман пише, що OpenAI думає відкрити Pando для інших

компаній.

Чому це важливо: агенти міняють навантаження на інструменти, які роками вважались розв'язаною задачею. Хто запускає багато агентів на одному репо, впреться в ту саму стіну. Вузьким місцем стане сервер Git.

Компроміс Pando корисний як зразок: команда відмовилась від глобальної узгодженості, яка на практиці нікому не потрібна.

ТЕМА 8

Пітер Левелс переніс Battlefield 1942 у браузер, а Claude тепер відмовляє в клонах ігор

ДЖЕРЕЛА @levelsio, 02.11 @levelsio · @levelsio, 02.11 @levelsio · Ars Technica, 02.11 Ars Technica

Пітер Левелс за один день переніс у браузер демо Battlefield 1942 з мультиплеєром. Для цієї гри нема відкритого рушія, і пів року тому він здався. Цього разу Claude Code з Opus 5.5 дизасемблював 57 000 функцій x86 і переклав кожну в C.

Потім агент зібрав усе в WebAssembly через Emscripten. Близько 300 функцій Windows і Direct3D довелось переписати на WebGL 2. Для мультиплеєра агент зробив власний мережевий стек на WebSockets.

За словами Левелса, робота коштувала 1 мільйон вихідних токенів, близько $100 за цінами API. Він платить $100 на місяць за підписку.

Того ж дня Ars Technica написала про протилежний бік. Близько 27 жовтня Anthropic додала в Claude Code два речення:

будуй оригінальні ігри та ігри в жанрі, але не відтворюй конкретну існуючу гру. Інженер Anthropic Едді Османі пояснив:

порушувати IP через Claude забороняє політика використання. Спільнота decomp.dev, яка роками розбирала ігри вручну, у жовтні закрилась і прямо назвала причиною потік AI-портів.

Чому це важливо: дизасемблювання і перенесення старого бінарника раніше займало роки ручної роботи. Тепер це день роботи агента за ціною підписки, і це стосується будь-якого легасі-коду без вихідників. Правило з двох речень у промпті ця можливість не скасовує, бо інші моделі таких обмежень не мають.

ТЕМА 9

Стенфорд перевірив 50 відомих висновків про LLM: 44% уже не працюють

ДЖЕРЕЛА Stanford, 02.11 Github · @emollick, 03.11 @emollick

Оуен Квін і Джеймс Цзоу зі Стенфорда взяли 50 часто цитованих висновків про LLM і повторили кожен експеримент на 20 моделях. Моделі вийшли з липня 2024 до вересня 2026 у восьми розробників. Висновок класифікували як «expired», якщо від ефекту лишилась п'ята частина або менше.

44% висновків уже expired, ще 34% помітно ослабли. Лише 16% тримаються, а 6% посилились. Приклад: у 2023 році GPT-3.5 вигадував 55% посилань в оглядах літератури. Сучасні моделі майже не вигадують. Зникли й провали з підрахунком літер, фактами в середині довгого контексту і відмовою від правильної відповіді під тиском.

Посилились три висновки. Моделі-судді частіше обирають власні відповіді. Моделі хвалять те, що подобається користувачу. Моделі й далі генерують випадкові числа нерівномірно. Тримаються підлабузництво і лівий політичний нахил.

За оцінкою авторів, через рік після публікації expired 16% висновків, через два роки 39%.

Чому це важливо: багато правил роботи з LLM походять зі статей про моделі 2023 року. Перевіряти своїх суддів і підлабузництво варто й далі, бо саме ці ефекти тримаються. А «моделі вигадують посилання» вже переважно історія. Будь-яку перевірку якості варто повторювати на кожній новій моделі.

ТЕМА 10

Посттренування стає окремим бізнесом: Applied Compute і Trajectory

ДЖЕРЕЛА @eladgil, 02.11 @eladgil · @rronak_, 02.11 @rronak_ · @deedydas, 02.11 @deedydas

Applied Compute, команда посттренування з колишніх працівників OpenAI, залучила $350 мільйонів за оцінкою $3,25 мільярда. Раунд очолив Елад Гіл. За даними компанії, річна виручка перевищила $100 мільйонів, а 18 місяців тому була нуль. 40% клієнтів і тренують у неї власні моделі, і запускають їх у продакшні.

Того ж дня Trajectory відкрила платформу для посттренування власних моделей через кілька промптів агенту. Компанія залучила загалом $60 мільйонів за оцінкою $300 мільйонів. Вона показала два прев'ю: модель вчиться на оцінках смаку замовника і на власних діях у продакшні.

Окремо OpenRouter, за даними інвестора Діді Даса, перейшла річний темп у 10 квадрильйонів токенів. Обсяг подвоюється кожні 10,5 тижня.

Чому це важливо: цифри інвесторів кажуть, що компанії хочуть власну модель на своїх даних, а базова модель стає сировиною. Для тих, хто вже має логи продакшну, посттренування дешевшає швидше за базові моделі. Цифри виручки тут названі самими компаніями, незалежної перевірки нема.

misc

- **LifeVal:** OpenAI випустила бенчмарк побутових задач: пошук садочка, держпослуги, особисті фінанси. За переказом The Rundown, GPT-6 Astra має в середньому 80,6%.

  @OpenAI · @TheRundownAI
- **GDPx:** SemiAnalysis випустила бенчмарк агентної роботи в дорогих галузях, найкраща модель має 12%. Ітан Моллік радить додати людський базовий рівень і пише, що GPT-6 Pro знайшов помилки в оцінюванні.

  @emollick
- **Перепис США:** Саймон Віллісон оновив графік використання AI за опитуванням Census Bureau. Після зміни питання великі компанії відповідають «так» приблизно в 40% випадків, рік тому було близько 12%.

  Simon Willison · Hacker News
- **Тао, день другий:** учора тут був гостьовий пост Калегарі. Тепер у блозі Тао вийшов текст трьох математиків про «математику без его»: цінність істини не залежить від того, хто її знайшов, людина чи машина.

  Wordpress · Hacker News
- **Отруєння претрейну:** Ніколас Карліні з колегами показали, що 0,1% отруєних даних у претрейні вшиває бекдор в агента. Модель на 4B виконує команду нападника приблизно в 92% випадків.

  arXiv · Hacker News
- **Subvocal:** стартап з YC показав окуляри, які радіохвилями читають беззвучну артикуляцію і перетворюють її на текст. Тред набрав 594 бали.

  Hacker News
- **Trigger.dev:** сервіс повідомив про витік змінних середовища частини клієнтів. Експлойт компанія заблокувала.

  Trigger · Hacker News
- **Binance:** Кім Йо Чжон назвала «absurd rhetoric» звинувачення Кндр у зламі Binance на $6,3 мільярда. ФБР приписує атаку групі TraderTraitor.

  BBC
- **Резидентура Anthropic:** компанія відкрила оплачувану дослідницьку програму для науковців, засновників та інженерів. ML-досвід не потрібен, заявки до 22 листопада.

  @AnthropicAI