AI-ДАЙДЖЕСТ

Рада Nvidia розслідує зв'язок Хуанга з агенткою Китаю

Також: евали харнесів для Opus 5.5, Goldman лишив Claude Code інженерам, DeepSeek на 284B без відеокарти.

понеділок, 9 листопада 2026

У ЦЬОМУ ВИПУСКУ

1. Хуанг визнав зв'язок, рада Nvidia почала незалежне розслідування
2. Відкриті евали: харнес майже не міняє результат Opus 5.5, міняє ціну
3. Goldman Sachs лишив Claude Code тільки інженерам
4. DeepSeek-V4-Flash на 284B параметрів працює на ПК з 64 ГБ пам'яті без відеокарти
5. Номер авто відкриває машину: дослідникам допоміг GPT 5.6
6. Тексти від AI забивають роботу установ
7. Science: гарнітури EEG годують моделі мозку даними користувачів
8. Outlook.com почав відхиляти частину пошти з доменів без DMARC
9. Коллісон: позитивний сценарій AI досі ніхто не описав
10. Пам'ять агентів: дев'ять форматів, жоден не читає інших

ТЕМА 1

Хуанг визнав зв'язок, рада Nvidia почала незалежне розслідування

ДЖЕРЕЛА Nvidia, 08.11 Nvidia · Axios, 08.11 Axios · The New York Times, 08.11 NYT · The New York Times, 08.11 NYT · Semafor, 08.11 Semafor · HN, 08.11 Hacker News

Вчора тут писали, що FT побачив пам'ятку розвідки США про стосунки Хуанга з Еліс Хуанг з сінгапурської Megaspeed.

У неділю сюжет перейшов від анонімних джерел до заяв самих учасників.

Хуанг розіслав лист працівникам Nvidia. «I was unfaithful to my wife», пише він і вибачається перед родиною, колегами й акціонерами. Він обіцяє повністю співпрацювати з розслідуванням. Текст листа опублікував Axios, Semafor і NYT цитують ті самі фрази.

Рада директорів оголосила незалежне розслідування. Його веде комітет незалежних директорів із зовнішнім юристом.

Рада пише, що не має причин вважати, що постраждала національна безпека або секрети Nvidia. Речник Джон Різзо сказав, що Хуанг не бере участі у виборі юриста і

лишається CEO «without interruption».

NYT додає дві деталі. Цієї весни держсекретар Марко Рубіо керував підготовкою секретної пам'ятки про цей зв'язок.

Пам'ятку розіслали посадовцям перед травневою поїздкою Трампа до Китаю. Сам Трамп у неділю сказав, що «не чув, що Хуанг зробив щось погане», і хоче, щоб той лишився CEO.

Чому це важливо: Хуанг керує Nvidia напряму, у нього близько 60 прямих підлеглих. Тому розслідування ради перевіряє, чи може компанія вартістю понад $4 трлн контролювати свого засновника. Semafor так і назвав це: тест «імунітету засновника».

ТЕМА 2

Відкриті евали: харнес майже не міняє результат Opus 5.5, міняє ціну

ДЖЕРЕЛА g-harness-evals, 08.11 GitHub · @garrytan, 08.11 @garrytan · caleb.software, 21.09, оновлено 07.11 Caleb · HN, 08.11 Hacker News

Гаррі Тан відкрив репозиторій з евалами агентних харнесів. Модель і задачі фіксовані, міняється лише харнес:

системний промпт, інструменти, пам'ять. Задач 20, вони публічні.

Результати на Opus 5.5 такі. Claude Code розв'язав 76,3% задач по $0,22 за розв'язану задачу. Hermes Agent дав 78,8% по $0,24. Цикл на Pi дав 72,5% по $0,19. OpenClaw дав 76,3%, але по $0,44 і в 2,5 раза повільніше. Автори пишуть прямо: різницю в частці розв'язаних задач між харнесами не видно. Розрив до 7,7-14,2 пункту на такій вибірці сховатись може.

Найчесніша знахідка стосується самого Тана. Його набір скілів gstack Opus 5.5 завантажував щоразу, але не виконав жодного кроку: 0 з 46 сесій. При цьому ціна прогону виросла приблизно вдвічі. У Claude Code правило в системному промпті спрацювало в 30 з 30 ходів, без правила - в 18 з 30. Скіл, який промпт не називає, модель не відкрила жодного разу за 30 ходів.

Пост «Goodbye Claude, Hello Pi» набрав на HN 713 балів. Автор перейшов з Claude Code на мінімальний харнес Pi. Його аргумент: за рік Claude Code мав 244 релізи, і понад 3/4 з них міняли системний промпт або описи інструментів. Опис одного інструмента Bash займає майже 3 000 токенів. У коментарях нагадують ціну мінімалізму: у Pi нема режиму, який блокує небезпечні дії агента.

Чому це важливо: на сильній моделі харнес вирішує ціну, швидкість і контроль, а частку розв'язаних задач вирішує модель. Скіл працює, лише коли промпт називає його прямо. Перевіряти це варто евалом на власних задачах, бо 20 публічних задач дають широкі інтервали.

ТЕМА 3

Goldman Sachs лишив Claude Code тільки інженерам

ДЖЕРЕЛА The State of AI з посиланням на Bloomberg, 02.11 Thestateofai · HN, 08.11 Hacker News [одне джерело]

04.11 тут була лише чутка з одного твіта: Goldman забороняє банкірам Claude Code. Тепер є деталі, які Bloomberg отримав від джерел у банку. Сам матеріал Bloomberg закритий пейволом, тому факти нижче взято з переказу The State of AI.

У жовтні банк оновив внутрішні правила. Claude Code і GitHub-інструменти тепер доступні лише інженерам. Їх близько 12 000 з приблизно 47 000 працівників. Доступ зберегли страти, які працюють між трейдингом і розробкою. Трейдери писали з агентами скрипти для відстеження потоків клієнтів і моделей позицій, і частина з них незадоволена.

Ще в липні директор з IT Марко Аргенті хвалив «vibe coding» нетехнічних працівників. У квітні банк уже прибрав Claude для банкірів у Гонконгу. При цьому Goldman разом з Morgan Stanley веде IPO Anthropic.

На HN найпопулярніше пояснення технічне. Агент на робочій машині трейдера має водночас клієнтські дані, пошту, інтернет і внутрішню мережу. Це «lethal trifecta»: такий агент може і прочитати секрет, і отримати шкідливу інструкцію, і відправити дані назовні.

Чому це важливо: код інженера проходить рев'ю і тести, а скрипт трейдера часто йде в щоденну роботу без них. У регульованій галузі банк обрав просте правило: агент, що пише і запускає код, отримує лише той, хто вміє цей код читати.

ТЕМА 4

DeepSeek-V4-Flash на 284B параметрів працює на ПК з 64 ГБ пам’яті без відеокарти

ДЖЕРЕЛА llama-modchip, 08.11 GitHub · HN, 08.11 Hacker News

Проєкт llama-modchip - це набір патчів до llama.cpp, квантизатор і обгортка для запуску. Головне демо: DeepSeek-V4-Flash на 284,33 млрд параметрів, з них 13,17 млрд активні. Модель важить 56,56 ГіБ, у середньому 1,71 біта на вагу. Звичайна збірка UD-IQ1_S від Unsloth важить 76,87 ГіБ і в 64 ГБ пам’яті не вміщується.

Як автори стиснули модель: 256 експертів кожного шару переважно в IQ1_S (1,56 біта), а спільні шари й увага лишились у 4-5 бітах. З контекстом 256K процес займає 57,61 ГіБ пам’яті. На 16-ядерному Zen 5 обробка промпту йде зі швидкістю 44,4 токена за секунду, генерація - 9,1-14,7 токена за секунду. Звичайні llama.cpp, Ollama і LM Studio такий файл не відкриють: він використовує власний тип тензора.

Критика на HN стосується якості. README порівнює патчений рушій з непатченим на тій самій стиснутій моделі. Порівняння з повною DeepSeek-V4-Flash у README нема.

Коментатори нагадують, що нижче 4 бітів якість зазвичай падає нелінійно.

Чому це важливо: модель рівня фронтиру тепер вміщується в настільний ПК, і для приватних даних це реальний варіант. Але 1,7 біта на вагу - це вже інша модель. Перед роботою варто прогнати її на своїх задачах проти API оригіналу.

ТЕМА 5

Номер авто відкриває машину: дослідникам допоміг GPT 5.6

ДЖЕРЕЛА Shubham Shah, 07.11 Shubs · HN, 08.11 Hacker News · The Guardian, 08.11 Guardian

Шубхам Шах і Сем Каррі зламали платформу підключених авто великого виробника. Бренд вони не називають. Для атаки досить номерного знака машини.

Ланцюжок довгий, але кожна ланка проста. Номер дає VIN, VIN дає обліковий запис власника. У JavaScript-бандлі постачальника лежав адмінський логін до системи облікових записів OpenIDM. Обхід WAF через path traversal дав доступ до її API. Далі дослідники міняли пароль власника, обходили MFA і отримували токен мобільного застосунку. Цей токен відкриває двері і запускає двигун. Разом з другим розбором Каррі атака працювала в Австралії, США і Канаді.

Ключовий обхід WAF знайшла модель. Люди перебрали префікси вручну, отримали 403 і вирішили, що закрито все. GPT 5.6 Sol Ultra шість годин читав код OpenIDM і знайшов префікс `/static%2f..%2fopenidm/`, який WAF не перевіряв.

Guardian того ж дня пише про ширшу картину. Глава британського NCSC Річард Горн каже, що світ лише в «передгір'ї»

нової хвилі атак. За даними видання, з квітня великі компанії знаходять у вісім разів більше вразливостей. Засновник NCSC Кіран Мартін заперечує: це погана серія, але не безпрецедентна.

Чому це важливо: модель не стомлюється перебирати варіанти, які людина відкинула як очевидно закриті. WAF поверх відкритої назовні адмінки тепер захищає ще гірше, ніж раніше. Публічні адмінки й секрети в JS-бандлах варто шукати раніше, ніж їх знайде чужа модель.

ТЕМА 6

Тексти від AI забивають роботу установ

ДЖЕРЕЛА The Guardian, 08.11 Guardian · Semafor, 08.11 Semafor

Guardian зібрав скарги британських установ. Працівниця офісу депутата каже: половина листів, які вони отримують, написана з AI. Скарга до місцевої влади раніше займала близько двох сторінок, тепер більшість у 10 разів довші.

Найбільша відома скарга має 200 сторінок. Transport for London отримав рекордні 4 769 запитів на доступ до інформації за рік, на 5% більше за попередній пік.

Найбільше заважає те, що AI пише скаргу як позов. Текст посилається на закони і судові справи, іноді з іншої юрисдикції, а іноді вигадані. Але відповідати все одно мусить людина. Дослідники Hertie School, Cooperative AI Foundation і GovAI вже назвали це «agentic flooding».

Semafor показав той самий механізм у пресі. Кампанія проти Waymo розміщувала колонки в місцевих газетах. Детектор Pangram оцінив дві колонки в Dallas Morning News як на 95% і 71% написані AI, колонку в San Francisco Chronicle - на 88%. Жодна не мала позначки. Один з авторів визнав, що першу версію написала PR-фірма.

Чому це важливо: AI робить текст дешевим для автора, але не для читача. Установи з фіксованим бюджетом першими відчувають цю асиметрію. Оцінки детектора - це ймовірність, а не доказ, тому позначки авторства залишаються відповідальністю видання.

ТЕМА 7

Science: гарнітури EEG годують моделі мозку даними користувачів

ДЖЕРЕЛА Science, 05.11 Science · HN, 08.11 Hacker News

Science рік досліджував споживчі нейротехнології. Канадська Interaxon, виробник пов'язки Muse, натренувала «модель мозку» приблизно на 50 000 годин EEG своїх користувачів. Її політика приватності прямо про це не казала. Компанія оновила політику, коли Science на це вказав.

Журнал перевірив політики приватності 57 компаній, що збирають EEG. Чотири не мають політики взагалі. Дев'ять компаній публічно визнали, що будують фундаментальну модель мозку. Лише шість прямо обіцяють не продавати дані мозку.

Понад дві третини дозволяють передати дані покупцеві при продажу компанії.

Регулювання поки точкове. Чилі захистила дані мозку в конституції 2021 року, а 2023-го її Верховний суд зобов'язав Emotiv видалити дані колишнього сенатора. Колорадо і Каліфорнія внесли нейродані до чутливих персональних даних.

Чому це важливо: дані, які вже потрапили в модель, з неї практично не видалити. Тому згода має бути окремою і до тренування, а не правкою в політиці після статті журналістів. Цінність таких даних компанії вже називають прямо:

«Our brain data is the most valuable part of the company», каже маркетинговий директор Interaxon.

ТЕМА 8

Outlook.com почав відхиляти частину пошти з доменів без DMARC

ДЖЕРЕЛА MailRoute, 06.11 Mailroute · HN, 08.11 Hacker News [одне джерело]

З 05.11 поштові сервіси Microsoft для приватних осіб (outlook.com, hotmail.com, live.com, msn.com) відхиляють частину листів з доменів, які не мають запису DMARC. Відправник отримує `550 5.7.1 Message rejected: Message lacks SPF/DKIM/DMARC Auth Records`. Про це пише MailRoute, сервіс поштової пересилки, за власними логами.

Підступність у тому, що відмова часткова. Один лист доходить, наступний повертається, повторна спроба може пройти.

Раніше Microsoft вимагала SPF, DKIM і DMARC лише від доменів, що шлють понад 5 000 листів на день. Тепер правило зачіпає і малі офіси.

MailRoute пише, що Microsoft нічого не оголошувала. Коментатор на HN заперечує і дає посилання на документацію Microsoft. Мінімальне виправлення - один TXT-запис `v=DMARC1; p=none`, але SPF або DKIM мають проходити для домену відправника.

Чому це важливо: малий бізнес, який шле рахунки й підтвердження на особисті адреси, втрачає листи мовчки. Великі поштові сервіси давно рухаються до DMARC як мінімуму для всіх. Перевірити свій домен можна однією командою `dig`.

ТЕМА 9

Коллісон: позитивний сценарій AI досі ніхто не описав

ДЖЕРЕЛА @patrickc, 08.11 @patrickc · @gdb, 09.11 @gdb · @levie, 09.11 @levie · The Wall Street Journal, 08.11 WSJ

Патрік Коллісон, співзасновник Stripe, опублікував два питання про AI, пост набрав понад 530 тисяч переглядів. Перше:

ризики AI описані добре, а хороший сценарій складається лише з інгредієнтів, наприклад «вилікувати хвороби». Чи працюють люди в такому світі, що в ньому дефіцитне, хто контролює? Коллісон цитує Іена Бенкса: люди в його «Культурі»

- «somewhere between passengers, pets and parasites».

Друге питання про моральну вагу AI. Якщо визнати за моделями навіть мізерну моральну вагу, то при квадрильйонах екземплярів їхня сумарна вага переважить людей у будь-якому утилітарному розрахунку. Коллісон пише, що тоді треба відкинути одне з трьох припущень. Сьогодні люди сперечаються про дата-центри, і він питає, як зміниться підтримка AI, коли розмова дійде до цих питань.

Того ж вечора Грег Брокман написав, що фронтирні моделі з великим обчисленням під час інференсу стають «рушієм наукових відкриттів». Аарон Леві додав, що продуктивність тепер обмежує обчислення, а не час людини. WSJ дає протилежний погляд з фармацевтики: жоден препарат, створений AI, ще не отримав схвалення FDA, а керівники фармкомпаній кажуть, що гальмує біологія.

Чому це важливо: обіцянки «AI вилікує хвороби» звучать дедалі частіше, а перевіряються повільно. Питання Коллісона про те, як виглядатиме суспільство після цього, поки не має відповіді від жодної лабораторії.

ТЕМА 10

Пам’ять агентів: дев'ять форматів, жоден не читає інших

ДЖЕРЕЛА LoreKit, 08.11 Lorekit · HN, 08.11 Hacker News

Автор LoreKit порахував, де інструменти для коду зберігають те, що «пам'ятають». AGENTS.md і CLAUDE.md, правила Cursor, тека `/memories` інструмента пам’яті Anthropic, Copilot Memory, Letta, Mem0, LangGraph, графи Graphiti і MCP.

Кожен формат розумний сам по собі, але жоден не читає інших.

Автор порівнює це з моніторингом до 2019 року. Тоді кожен вендор мав свій формат, а OpenTelemetry дав спільну модель даних. Вендори перестали конкурувати на форматі і конкурують на аналізі. Для пам’яті вже є нейтральне місце: MCP і AGENTS.md належать Agentic AI Foundation при Linux Foundation.

Корисні деталі зі списку. Copilot Memory сам видаляє факт, яким не користувались 28 днів, і прив’язує факти до цитат з коду. Graphiti зберігає два часи: коли факт був правдою і коли його записали. Автор чесно зазначає, що сам продає один з форматів.

Чому це важливо: урок, який вивчив агент у CI, зараз не бачить агент у редакторі. Хто обирає інструмент сьогодні, тому варто перевірити, чи можна пам’ять звідти експортувати.

---

misc

- **C++26 застряг:** Герб Саттер пише, що бюлетень DIS провалився вперше за 36 років. Причина: Б’ярн Страуструп виступив проти контрактів, які комітет WG21 підтримав. Isocpp · HN Hacker News
- **KetchUp:** відкрита альтернатива SketchUp на Rust і WebGPU працює прямо в браузері, 292 бали на HN.
  Github · HN Hacker News
- **Anthropic і газ:** FT пише, що Anthropic уклала угоду на $3,4 млрд про газову генерацію для своїх дата-центрів.
  Текст закритий пейволом, доступний лише заголовок. FT [одне джерело]
- **Евал проти статті:** @garrytan перевірив своїм проєктом евалів статтю про здешевлення агентів на 90% з Kimi K3.
  Результати статті повторились. @garrytan