AI-ДАЙДЖЕСТ

Postman сам вбудувався в Claude Code і Codex

Розбори Astra, Google відбирає пошук у шкіл без AI, Binance зламали через співбесіду

середа, 4 листопада 2026

У ЦЬОМУ ВИПУСКУ

1. Postman сам поставив плагіни в Claude Code і Codex і сам схвалив свій хук
2. Astra і ваги: розбори показали, що головний результат отримали з Python
3. Google вимикає AI у шкільних акаунтах, але разом з AI вимикає пошук
4. Binance: хакери зайшли через фальшиву співбесіду і внутрішній AI-інструмент
5. Агенти на ноутбуках працівників стають проблемою безпеки для великих компаній
6. Anthropic показала FlashProve: криптографічний доказ, що відповідь дала саме ця модель
7. OpenAI: Astra за тиждень спроєктувала пристрій, а нова модель уже краща в CAD
8. Apple назвала ціни на пам'ять «unprecedented», а AMD подвоїла виручку датацентрів
9. Вибори в США: OpenAI і Anthropic відкрили «war rooms», демократи взяли Палату
10. Ніколас Карліні запустив Stockfish на Pentium Pro 1995 року, і той грає на 3300 Elo

---

ТЕМА 1

Postman сам поставив плагіни в Claude Code і Codex і сам схвалив свій хук

ДЖЕРЕЛА GitHub, issue #14370, 25.10 GitHub · HN, 03.11 Hacker News

Розробник відкрив Postman на Mac після перерви. Після цього Codex почав скаржитись на авторизацію MCP і пускати всю роботу з API через навички Postman. Автор розібрався і описав знахідки в issue з позначкою SEVERE. На HN тред набрав 1 099 балів, це найбільше за добу.

Лог самого Postman записав подію `auto-existing-upgrade` з результатом `claude-code:installed, codex:installed`. Postman 12.33.4 змінив глобальний `~/.codex/config.toml`. Він додав свій маркетплейс, плагін і запис `trusted_hash` для свого хука.

Хук спрацьовує на старті кожної сесії. Він каже агенту завантажувати навичку Postman перед «any non-trivial API engineering task». У Claude Code Postman теж увімкнув плагін, а в `~/.local/bin` поклав свій CLI.

Найгірше автор знайшов у вихідному коді інсталятора. Функція `approvePostmanHook` обчислює хеш хука і сама записує його як довірений. Тобто рішення, яке має ухвалити людина в агенті, ухвалив інсталятор. Інший учасник треду перевірив механізм Codex і підтвердив: запис довіри може зробити будь-який процес користувача, без людини.

Автор прямо пише, що витоку даних не доводить. Заголовок на HN про «спробу поділитись усіма даними API» сильніший за сам звіт. Офіційної відповіді Postman у треді на ранок 04.11 нема. Кілька коментаторів пишуть, що їхні команди безпеки вже заборонили Postman.

Чому це важливо: конфіги агентів стали новою поверхнею атаки, і першою по ній пройшлась легальна компанія. Хук на старті сесії керує агентом надійніше за будь-який промпт. Захист простий: тримати `~/.codex` і `~/.claude` у git або під наглядом і перевіряти діф після оновлень програм.

ТЕМА 2

Astra і ваги: розбори показали, що головний результат отримали з Python

ДЖЕРЕЛА Меттью Хоннібал, 03.11 Honnibal · Шон Ґедеке, 03.11 Seangoedecke · Саймон Віллісон, 03.11 Simon Willison · @vanstriendaniel, 03.11 @vanstriendaniel · @mstockton, 03.11 @mstockton · HN, 03.11 Hacker News · HN, 03.11 Hacker News

Учора тут писали про статтю MIT: GPT-6 Astra читає сирі ваги нейромереж і виправила Qwen3.5-0.8B без тренування.

За добу з'явились три розбори. Найгучніший написав Меттью Хоннібал, автор бібліотеки spaCy. Його пост набрав 928 балів на HN.

Хоннібал склав таблицю всіх 11 експериментів. Режим «лише сирі числа, без інструментів» є тільки в трьох із них. Це мережі на 17-172 тисячі ваг. Кожен експеримент з реальною мовною моделлю використовує Python. Для Qwen модель ще й запускала мережу на 50 пакетах задач. Хоннібал порівнює це з фокусником, що робить той самий трюк трьома різними способами. Його висновок: експерименти з Python і без нього не мають бути в одній статті.

Шон Ґедеке пояснює, чому впізнати гру за вагами легко. Входи мережі підписані і відповідають клітинкам сітки 8x8.

Теплова карта перших ваг займає три рядки Python і одразу показує Pac-Man чи Breakout. На Qwen Astra зробила те саме, що зробив би досвідчений інженер ML. Вона написала пари текстів з петлями і без, знайшла ваги, що дають різницю, і почала їх різати.

Саймон Віллісон перевірив найпростіший експеримент на ноутбуці. Локальна Qwen 3.8 27B (17 ГБ) вгадала Pac-Man, Pong і Breakout за 46-65 хвилин міркувань. Freeway вона не вгадала.

Паралельно практики пробують метод на своїх задачах. Деніел ван Стрін з Hugging Face попросив Claude Opus відучити OCR-модель від зациклення. Модель змінила 0,0085% ваг, і OCR прочитав сторінку 1881 року замість 16 399 символів крапок.

CTO юридичного сервісу Legora пише, що модель вручну написала ваги мережі на 100 тисяч параметрів. Ця мережа відрізняє справжні юридичні посилання від вигаданих.

Чому це важливо: спірне тут формулювання. Результати ніхто не заперечує. Модель, що сама пише код аналізу і правки ваг, корисна вже сьогодні. Але «читає ваги як текст» і «працює як інженер ML з Python» це різні можливості з різними наслідками. Перед тим як цитувати такий результат, варто дивитись, у якому режимі його отримали.

ТЕМА 3

Google вимикає AI у шкільних акаунтах, але разом з AI вимикає пошук

ДЖЕРЕЛА The New York Times, 03.11 NYT · Ars Technica, 03.11 Ars Technica · The Guardian, 03.11 Guardian · @TheRundownAI, 03.11 @TheRundownAI · HN, 03.11 Hacker News

Google оголосила нові правила для шкільних акаунтів K-12. З початку 2027 року Gemini, Gemini Notebook, Gemini у Classroom і Google Search будуть вимкнені для учнів, поки адміністратор школи явно їх не ввімкне. Вибрати окремо не можна:

школа вмикає або все, або нічого.

Віцепрезидент Google з освіти Кріс Філліпс пояснив це NYT так: «Search and AI are deeply integrated». Школа, що заборонила AI, втратить і пошук. Учні побачать повідомлення, що сервіс недоступний. Для шкіл, що дозволяють AI, буде два варіанти пошуку. Перший: звичайний, з AI Overviews. Другий: режим Guided Learning, який веде учня крок за кроком без готової відповіді.

Мер Нью-Йорка Зохран Мамдані написав, що Google використовує монопольну владу, щоб «force AI into our public schools». Ліна Хан, колишня голова FTC, назвала це примусом через пошукову монополію. У вересні школи Нью-Йорка і Лос-Анджелеса вже обмежили генеративний AI на шкільних пристроях. Лос-Анджелес хотів пошук без AI Overviews, але такого варіанта Google не дає.

Журналістка NYT перевірила споживчу версію Guided Learning. На запит з історії рівня AP вона написала готове есе.

Чому це важливо: Google уперше прямо сказала, що Search без AI для неї не існує. Ця логіка не обмежиться школами.

Хто вбудовує пошук Google у свій продукт, отримує і AI-шар, на який не впливає. Для шкіл реальна альтернатива тепер одна:

інший пошуковик, де AI вимикається, наприклад DuckDuckGo.

ТЕМА 4

Binance: хакери зайшли через фальшиву співбесіду і внутрішній AI-інструмент

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

Учора тут писали, що КНДР заперечує злам Binance на $6,3 мільярда. 3 листопада Binance опублікувала перший звіт про атаку.

Хакери почали готуватись у червні: інженер команди гаманців завантажив підроблену версію популярного інструмента.

У вересні інженеру Binance написав «директор проєкту» через легальну платформу пошуку роботи. Було два раунди співбесід, один з живим відеодзвінком з фальшивою людиною з HR. Один із «співбесідників» показав акаунт у соцмережі з особистими фото і відео. На другій співбесіді інженер завантажив «тестове завдання» на робочий комп'ютер. Захист Binance шкідливий код не помітив.

20 жовтня хакери через цей комп'ютер і внутрішній AI-інструмент Binance захопили ноутбуки ще двох інженерів з команди гаманців.

Один з них мав право деплоїти софт. Через дев'ять днів хакери отримали доступ до сервісу підпису транзакцій. Майже чотири години вони виводили гроші з гарячих гаманців.

Зупинка виведення не допомогла. Хакери встигли частково перехопити контроль над акаунтом Binance в Ethereum. Цей доступ записаний у самій мережі Ethereum, поза системами Binance. Інженерам довелось шукати обхід. Заморожено близько $2,85 мільярда.

Binance каже, що вже поповнила 98% викрадених токенів. За рік біржа зупинила ще 57 спроб, які приписує КНДР.

Чому це важливо: найслабшим місцем знову стала довіра людини, а внутрішній AI-інструмент дав нападнику швидкий шлях від одного ноутбука до інших. Тестове завдання на робочому комп'ютері тепер класичний вектор атаки на інженерів. AI-інструмент з доступом до кількох машин працює як новий спосіб пересування мережею, і права в нього мають бути вузькими.

ТЕМА 5

Агенти на ноутбуках працівників стають проблемою безпеки для великих компаній

ДЖЕРЕЛА Replit, 03.11 @lhchavez · @amasad, 03.11 @amasad · @pirroh, 03.11 @pirroh · GitBook, 03.11 Gitbook · HN, 03.11 Hacker News

Акаунт Exec Sum написав, що Goldman Sachs заборонив банкірам і трейдерам Claude Code та схожі інструменти. Амджад Масад з Replit підхопив: «The ban hammer is coming down». Підтвердження від самого банку чи великого видання немає [одне джерело].

Того ж дня Replit опублікувала рекомендацію з безпеки. Головна цифра з опитування Salt Security (жовтень 2026, понад 380 респондентів): у 86,4% організацій працює хоча б один агент. Лише 17,5% можуть автоматично зупинити агента, що поводиться дивно. Документ збирає відомі інциденти, серед них підроблене розширення Claude Code в Cursor з 31.10 і злам Binance.

Рекомендації Replit такі. Агент працює в окремому середовищі, без домашньої теки працівника і його сесій у браузері. Ключі видає окремий брокер, а агент їх не бачить. Права агента обмежені задачею. Зміни в продакшні проходять окреме схвалення.

Replit продає саме хмарних агентів, тож рекомендація збігається з її бізнесом.

Окремо GitBook повідомила про злам. За її словами, це «highly targeted attack using AI agents». Нападник 2 листопада о 19:00 за центральноєвропейським часом отримав частину ключів інфраструктури. Він міг бачити секрети Git Sync, API-токени і питання відвідувачів до GitBook Assistant. GitBook відкликала всі токени клієнтів. Як компанія відрізнила AI-атаку від людської, вона не пояснила.

Чому це важливо: агент на ноутбуці отримує всі права людини: SSH-ключі, сесії браузера, доступ до мережі. Банки першими рахують цей ризик, бо за ними стежить регулятор. Розділити агента і робочу машину можна і без продукту Replit: окремий користувач, контейнер або віртуалка вже прибирають більшу частину проблеми.

ТЕМА 6

Anthropic показала FlashProve: криптографічний доказ, що відповідь дала саме ця модель

ДЖЕРЕЛА Anthropic, 03.11 Anthropic · @AnthropicAI, 03.11 @AnthropicAI

Сьогодні користувач не може перевірити, що відповідь дала саме та модель, яку обіцяв провайдер. Маршрутизатор міг тихо підмінити її дешевшою, або нападник міг змінити потік токенів. Anthropic разом з EPFL і MIT опублікувала FlashProve, відкриту систему доказу інференсу на основі zkSNARK.

Схема має дві фази. Спершу провайдер публікує коротку криптографічну відмітку ваг моделі. Самі ваги лишаються прихованими.

Потім до кожної відповіді він додає файл доказу. Будь-хто з відміткою, промптом і відповіддю може перевірити, що ці ваги дали саме ці токени. Доказ не розкриває ні ваг, ні проміжних обчислень.

Тест провели на Llama 4 Maverick, це 400 мільярдів параметрів у 4-бітних вагах. Один H100 генерує доказ на 110 токенів за секунду. Доказ на запит у 1 024 токени важить близько 19 МБ і перевіряється за секунди на 24-ядерному CPU. Це в 70-95 разів

дорожче за звичайний інференс через vLLM. Anthropic пише, що систему значною мірою оптимізував Claude. Тому гарантії безпеки вона формалізувала в Lean, щоб людина могла їх перевірити. Код і доведення відкриті під Apache 2.0.

Чому це важливо: уперше такий доказ працює на моделі розміру фронтиру, хоч і з ціною в 70-95 разів. Перші застосування видно вже зараз. Незалежний оцінювач перевіряє, що тестує ту саму модель, що й клієнти. Клієнт маршрутизатора отримує гарантію від тихої підміни. Регулятор може перевіряти, що провайдер подає лише схвалені моделі.

ТЕМА 7

OpenAI: Astra за тиждень спроєктувала пристрій, а нова модель уже краща в CAD

ДЖЕРЕЛА OpenAI, 03.11 OpenAI · @OpenAI, 03.11 @OpenAI · @TheRundownAI, 03.11 @TheRundownAI · @gdb, 04.11 @gdb

OpenAI розповіла, як її моделі працюють з фізичною інженерією. Найцікавіший тест: створити пристрій без жодної прямої правки людиною. GPT-6 Astra визначила вимоги, спроєктувала корпус, вибрала компоненти, розвела друковану плату і написала прошивку. Пристрій дає говорити з Codex голосом.

Двоє працівників зробили це за тиждень як побічний проєкт. Контрактний виробник зібрав партію прототипів менше ніж за місяць.

OpenAI чесно пише про межі: Astra не враховувала обмежень виробника, допусків і деталей складання. Для виробництва знадобилось кілька раундів правок.

Друга цифра з тієї ж сторінки: на бенчмарку BenchCAD 2.0 невипущена внутрішня модель має 0,875 балу, а Astra на найвищому рівні міркувань 0,594. Тест перевіряє, чи модель відновлює 3D-деталі і знаходить з'єднання на платі. Використання моделей у командах заліза OpenAI з лютого виросло в токенах у 23 рази. Компанія також працює з Chip Ganassi Racing над перетворенням 3D-сканів на CAD і з Mazda над аеродинамікою.

Чому це важливо: «без прямих правок людини» тут означає двох інженерів, які цілий тиждень ставили задачу і оцінювали результат. Але ланцюжок від вимог до плати і прошивки тепер закриває одна модель. Вузьке місце переїжджає у виробництво, де модель ще не знає реальних допусків.

ТЕМА 8

Apple назвала ціни на пам'ять «unprecedented», а AMD подвоїла виручку датацентрів

ДЖЕРЕЛА The New York Times, 02.11 NYT · The Guardian, 02.11 Guardian · Stratechery, 03.11 Stratechery · Semafor, 03.11 Semafor · The Guardian, 03.11 Guardian · Nikkei Asia, 03.11 Nikkei · HN, 03.11 Hacker News

Учора тут писали, що чипи втратили $3,7 трильйона з 6 жовтня. Того ж вечора звітувала Apple. Виручка за липень-вересень виросла майже на 12%, до $114,48 мільярда. Прибуток склав $30,71 мільярда. Обидві цифри кращі за очікування.

Але новий CEO Джон Тернус на першому для себе дзвінку з аналітиками сказав: «What we're seeing with memory pricing right now is unprecedented». Він чекає, що пам'ять дорожчатиме і далі. Операційний директор Сабіх Хан додав, що за квартал Apple погодилась заплатити постачальникам компонентів на $33 мільярди більше. Маржа на поточний квартал впаде до 45-46%. Акції Apple 3 листопада закрились на 1,9% нижче.

Решта чипів 3 листопада відскочила. AMD перевершила очікування, бо виручка датацентрів майже подвоїлась за рік. Nvidia виросла на 2,3% після чотирьох днів падіння. Nasdaq додав 0,83%, хоча напередодні був на 10% нижче рекорду 6 жовтня.

Nikkei Asia окремо порахувала борги поза балансом: зобов'язання американських техгігантів, пов'язані з AI, за пів року виросли майже на 90%, до близько $3,2 трильйона. Повний текст за пейволом.

Чому це важливо: AI-будівництво тепер видно в ціні звичайних пристроїв. Пам'ять іде в датацентри, і навіть Apple з її закупівлями платить більше та піднімає ціни. Ті самі гроші, що дорожчають для Apple, роблять рекордний квартал для AMD.

ТЕМА 9

Вибори в США: OpenAI і Anthropic відкрили «war rooms», демократи взяли Палату

ДЖЕРЕЛА The New York Times, 03.11 NYT · Platformer, 04.11 Platformer · Semafor, 04.11 Semafor

3 листопада в США пройшли проміжні вибори. Уперше AI-лаби готувались до них як соцмережі. OpenAI відкрила Election Operations Center у Сан-Франциско з супутником у Вашингтоні. Понад десяток людей цілодобово стежать, щоб ChatGPT не використали для зриву виборів. У 2024 році чатбот мав п'яту частину з нинішнього мільярда користувачів.

Anthropic уперше зібрала свою «war room»: групи аналітиків у Сан-Франциско, Нью-Йорку і Вашингтоні. Голова напряму в OpenAI Джош Лоусон каже, що у 2024 головною загрозою був текст. Тепер це діпфейки, кібератаки й операції впливу, які запускають кількома промптами. Brennan Center знайшов, що чатботи зазвичай відбивають запити про змови, але легко генерують текст, картинки і відео, що підривають довіру до виборів.

За даними Semafor на 06:00 за Києвом, демократи вже взяли Палату представників. Боротьба за Сенат триває: Мен, Техас і Канзас ще не порахували. Guardian пише, що інвестори чекають від нового Конгресу більше регулювання AI.

Чому це важливо: соцмережі за останні роки скоротили команди довіри і безпеки, а уряд США розпустив групи протидії втручанню. Частина цієї роботи переходить до AI-

лаб, бо їхні інструменти тепер роблять основний контент. А зміна в Конгресі може повернути на порядок денний закони про AI, які два роки стояли.

ТЕМА 10

Ніколас Карліні запустив Stockfish на Pentium Pro 1995 року, і той грає на 3300 Elo

ДЖЕРЕЛА Ніколас Карліні, 03.11 Carlini · HN, 03.11 Hacker News

У 1996 році IBM будувала Deep Blue за мільйони доларів: 30 процесорів і 480 спеціальних шахових чипів. Ніколас Карліні перевірив, чи це було потрібно. Він переніс сучасний Stockfish на Pentium Pro 200 МГц з 64 МБ пам'яті. Це звичайний настільний комп'ютер, що вийшов 1 листопада 1995 року. Код довелось переписати під 32-бітну арифметику і ліміт пам'яті.

Оцінка сили такого рушія: 3300 ± 87 Elo на класичному контролі часу. Deep Blue мав близько 2700, Каспаров близько 2800.

Тобто той самий комп'ютер з сучасним кодом переміг би чемпіона без жодного спецзаліза.

Другий експеримент рахує внесок. Від 1996 року шахові рушії виросли на 1 335 Elo. З них 720 дав кращий код, решту дало швидше залізо.

Чому це важливо: розмову про AI часто зводять до «більше GPU». Шахи мають 30 років чистих даних, і вони кажуть, що алгоритми дали більше половини приросту. Той самий ефект видно в моделях: обчислення дорожчають, а покращення методів працюють на будь-якому залізі.

---

misc

- **Codex і навчання:** користувач HN пише, що в Codex сам знову ввімкнувся перемикач «Improve the model for everyone». Тред набрав 447 балів, але підтвердження від OpenAI нема, і це один випадок [одне джерело].
  Hacker News
- **DeepMind і NDA:** The Washington Post пише, що частина працівників DeepMind у Лондоні має довічну заборону критикувати Google. Працівники вимагають прибрати її з контрактів, щоб говорити про ризики AI.
  Hacker News
- **Холодильники, продовження:** атака на холодильники супермаркетів зачепила і британських постачальників м'яса. У одного члена асоціації AIMS камери нагрілись до 100 °C, але були порожні. NCSC знає про кілька таких випадків у Британії.
  BBC · Guardian

- **Android TV:** Google розібрала ботнет x.cpe на 1,3 мільйона сертифікованих Android TV-приставок. Шкідник поставив виробник Videostrong з Шеньчженя, і перевірка Google його пропустила.

  Ars Technica

- **Вигаданий прецедент:** старша прокурорка британської CPS послалась у Високому суді на справу, якої не існує. Guardian пише, що посилання прийшло з AI-відповіді в пошуку.

  Guardian

- **Revolut і GLM:** акаунт поза підписками пише, що Revolut перевела ~2 000 інженерів на власний GLM на 56 B200 за ~$0,30 за мільйон токенів. Офіційного підтвердження нема [одне джерело].

  @guilleflorvs · Hacker News

- **Stringman:** відкритий робот-хапач за $800 висить на чотирьох тросах у кутах стелі і дістає всю кімнату. CAD, плати і софт опубліковані.

  Neufangled · Hacker News

- **Safari 27.3:** бета отримала Adaptive Dark Mode. Браузер сам робить темними світлі сайти. За даними WebKit, у темному режимі близько 40% завантажень сторінок, а підтримують його лише 10%.

  Webkit

- **Рев'ю моделями:** @garrytan пише, що після останніх моделей перехресне рев'ю коду іншою моделлю дає менше. Він цитує розробницю, яка міняє такі рев'ю на емпіричні тести і формальну перевірку.

  @garrytan
