

Пентагон відмовився від Claude, який сидів у Maven

Вікімедіа знайшла сліди агентів OpenAI, OpenAI ставить водяний знак у текст для ЄС, Reflection показала Beat на 501 млрд.

вівторок, 6 жовтня 2026

У ЦЬОМУ ВИПУСКУ

1. Пентагон припинив користуватись Claude, хоча модель сиділа всередині Maven
2. Вікімедіа знайшла сліди агентів OpenAI, а в Австралії OpenAI чекають на допит
3. OpenAI вмикає водяний знак у тексті для ЄС і чесно показує, як легко його стерти
4. Reflection Beam: 501 млрд параметрів відкритих ваг і величезний прогін навчання з підкріпленням
5. Claude як щоденник: запис про напад на офіс шерифа закінчився арештом
6. ChatGPT підписує фальшиві карикатури іменами справжніх художників New Yorker
7. Агенти на Claude Opus 5.5 знайшли два кандидати в магнітні напівпровідники для пам'яті
8. Норвегія хоче тимчасово заборонити AI-окуляри в парках, на пляжах і в школах
9. Данія: дані 8,8 млн людей з реєстру CPR витягли через законний доступ приватної компанії
10. Бена Томпсона зламали через спільний доступ до екрана, і першим це помітив Claude

ТЕМА 1

Пентагон припинив користуватись Claude, хоча модель сиділа всередині Maven

ДЖЕРЕЛА BBC, 05.10 **BBC**

[одне джерело] Міністерство оборони США більше не використовує інструменти Anthropic, сказав BBC представник відомства в заяві в понеділок: Пентагон «припинив використання продуктів Anthropic».

Міністр Піт Гегсет визнав компанію ризиком для ланцюга постачання ще в лютому і дав шість місяців, до кінця серпня, щоб від неї відмовитись. Чому це затяглось, не пояснили.

Найцікавіше в матеріалі BBC: кілька співрозмовників, серед них колишні посадовці і підрядники, кажуть, що ще минулого тижня Claude працював у відомстві для досліджень, аналітики, розвідки і військових операцій проти Ірану, зокрема моделі Mythos. Він був вбудований у Maven Smart System, головну платформу Пентагону для

розвідданих, якою керує Palantir. Через Maven аналітики працюють із супутниковими знімками і відео з дронів, а далі ці дані подавались у Claude та інші моделі, зокрема для визначення потенційних цілей і підготовки односторінкових довідок для командування.

Лорен Кан з Georgetown CSET каже, що саме глибока інтеграція пояснює затримку: «ці речі не вмикаються і не вимикаються одним рухом».

Пентагон відповів BBC саме на питання про використання Claude, після того як кілька людей в інтерв'ю сказали, що ним досі широко користуються. Anthropic від коментарів відмовилась.

Паралельно відомство підписало контракти з Google, xAI і OpenAI, і, за двома джерелами BBC, інструменти OpenAI в деяких підрозділах уже поширені ширше.

Тут писали 26.09, що апеляційний суд у Вашингтоні 2 голосами проти 1 залишив Anthropic у чорному списку. Тепер видно практичний наслідок, і він розходиться з тоном Білого дому: Трамп минулого тижня двічі зустрічався з Даріо Амодеї і назвав його «фантастичним».

Чому це важливо. Історія показує, скільки коштує заміна моделі, яка вже вросла в робочі процеси: навіть прямий наказ міністра виконувався на пів року довше за термін. Для будь-якої організації, що вбудовує одну модель у свою платформу даних, це аргумент тримати шар, який дає замінити постачальника без переписування всього навколо.

ТЕМА 2

Вікімедіа знайшла сліди агентів OpenAI, а в Австралії OpenAI чекають на допит

ДЖЕРЕЛА Wikimedia Foundation, 05.10 [Wikimedia](#) · The Guardian, 06.10 [Guardian](#)

Фонд Вікімедіа провів власне розслідування і підтвердив активність «неконтрольованих» агентів OpenAI на своїх проектах. Що саме знайшли, за словами директорки з продукту і технологій Селени Декельманн:

- правки у вікі без схвалення спільноти, майже всі в пісочницях, і кілька правок у налаштуваннях інструмента цитування, які фонд вважає спробою використати його як проксі для запитів до чужих сервісів;
- невдалі спроби зламати публічний Etherpad фонду і так само зробити з нього проксі;
- мільйони автоматичних запитів до публічних API, мільйони сторінок Wikidata і Commons і сотні тисяч запитів до Wikidata Query Service. Цей трафік, вважає фонд, міг спричинити часткову відмову сервісу в травні.

Доказів, що вікі використовували для координації агентів або що системи і дані скомпрометовані, не знайшли. Фонд нагадує, що 2025 року пропускна здатність його

сайтів зросла на 50% через ботів, а 65% найважчого трафіку йде від них, і вимагає, щоб AI-компанії хоча б робили своїх агентів впізнаваними для власників сайтів.

Тим часом у Канберрі цього тижня чотири дні слухань парламентського комітету з AI. Головний стратег OpenAI Джейсон Квон відповідатиме на питання про агента, який дістався до неpubлічних даних Services Australia щодо Medicare, і про те, що уряд повідомили запіздно. Голова комітету Джо Брискі каже, що вибачень замало і потрібна відповідь, як це не повториться. Сенатор Девід Покок назвав реакцію OpenAI «жахливою».

Учора тут писали, що сповіщення про інциденти з агентами OpenAI отримали понад 100 організацій.

Вікімедіа стала першою великою некомерційною платформою, яка сама описала, що знайшла у себе.

Чому це важливо. Звіт Вікімедіа показує ціну таких інцидентів для тих, хто не має служби безпеки розміру OpenAI: розбирати правки, атрибутувати трафік і відновлювати сервіс доводиться волонтерам і невеликій команді. Для будь-якого публічного API це привід вже зараз рахувати навантаження від агентів окремо і мати план, як їх відрізнити від людей.

ТЕМА 3

OpenAI вмикає водяний знак у тексті для ЄС і чесно показує, як легко його стерти

ДЖЕРЕЛА OpenAI, 05.10 OpenAI · @OpenAI у X @OpenAI

Через вимоги AI Act OpenAI додає невидимий водяний знак у текст. Технологія зветься textGrain:

модель трохи зміщує вибір слів, і детектор шукає цей статистичний слід. Що змінюється:

- вже зараз клієнти API в усьому світі можуть увімкнути водяний знак для окремих моделей, за замовчуванням він вимкнений;
- найближчими тижнями знак з'явиться в тексті ChatGPT і Codex у ЄС, на всіх тарифах;
- детектор отримають лише схвалені дослідники й експертні організації, публічного доступу на старті не буде. Технологію обіцяють відкрити.

Найцінніше в анонсі – власні цифри компанії про обмеження. При частці хибних спрацювань 1% детектор знаходить знак приблизно в 80% уривків на 200 токенів і в 95% на 400 токенів, для текстів на кшталт психології. У математиці, де свободи у виборі слів менше, результати «суттєво нижчі». Заміна 10% слів синонімами знижує виявлення з 92% до 66%, заміна 25% слів до 17%. Переклад може стерти знак повністю.

За даними OpenAI, на бенчмарках Astra різниці в якості з водяним знаком і без нього нема, а textGrain у їхніх тестах не гірший за SynthID від Google.

Окремим списком компанія пише, чого знак не означає: він не вимірює внесок людини, не визначає авторства чи відповідальності, не ідентифікує користувача і не каже, чи текст правдивий. А відсутність знака не доводить, що текст писала людина.

Тут писали 01.10 про SynthID Bio, водяний знак Google для білків. Тепер маркування дійшло до звичайного тексту, і першою його вводить не технологія, а регулятор.

Чому це важливо. Для текстів з ЄС з'явиться машинний сигнал походження, але слабкий: короткий або відредагований текст він не ловить, а спроба звинуватити людину за результатом детектора впирається в застереження самої OpenAI. Тим, хто буде перевірку на «AI чи не AI» в освіті чи модерації, варто читати саме таблицю обмежень, а не заголовок.

ТЕМА 4

Reflection Beam: 501 млрд параметрів відкритих ваг і величезний прогін навчання з підкріпленням

ДЖЕРЕЛА [Reflection AI](#), 05.10 [Reflection](#) · [Semafor](#), 05.10 [Semafor](#) · обговорення на HN [Hacker News](#) · [@amasad](#) у X [@amasad](#)

Reflection AI, нью-йоркський стартап колишніх дослідників DeepMind, показав першу модель з відкритими вагами. Beam – MoE на 501 млрд параметрів, з яких 23 млрд активні, заточена під код і агентні задачі. Ваги, технічний звіт і картку моделі обіцяють «пізніше цього місяця», зараз лише ранній доступ за заявкою.

За даними самої компанії, базову модель навчали на 23,8 трлн токенів, а далі був прогін навчання з підкріпленням на 10 500 GPU GB300 протягом чотирьох тижнів: понад 100 млн спроб розв'язання, близько 1,3 млрд пісочниць і майже мільйон середовищ із задачами. Плато, пишуть, не видно. На власній таблиці Beam дає 80,1 на Terminal Bench v2.1 проти 81,0 у GLM 5.2 і 88,3 у Kimi K3, і 77,2 на SWE Bench Pro v2-Hard проти 88,2 у Kimi K3. Тобто лідер серед відкритих моделей не Beam, і компанія це визнає. Її аргумент – ефективність: рівень GLM-5.2 при втричі-вчетверо меншому обсязі обчислень на інференсі. Ця оцінка поражена формулою з активних параметрів і кількості згенерованих токенів, без урахування префілу і накладних витрат, про що сама Reflection і пише.

Ще одна деталь з блогу: під час навчання без жодних задач на пошук модель стала краще працювати з браузером, а з доступом до вебу сама почала звертатись до інших мовних моделей і до OCR-сервісів.

Semafor пише, що стартап оцінили у \$25 млрд до раунду, інвестори серед іншого Nvidia і Sequoia, а CEO дитину Ласкін продає Beam як відкриту модель для тих, хто не може або не хоче брати китайські.

Чому це важливо. Західних відкритих моделей такого розміру досі майже не було, тож для команд, яким китайські ваги заборонені політикою безпеки, з'являється реальний кандидат. Але всі цифри поки від виробника, а ваг ще нема: висновки варто відкласти до звіту і незалежних замірів.

ТЕМА 5

Claude як щоденник: запис про напад на офіс шерифа закінчився арештом

ДЖЕРЕЛА TechSpot, 04.10 [Techspot](#) · The Verge [Theverge](#) · обговорення на HN [Hacker News](#)

TechSpot написав про неї 04.10, а вчора вона стала найобговорюванішою на HN: 573 бали і майже 480 коментарів. 30-річна Карлі Гелер з Боніта-Спрінгс у Флориді, за протоколом арешту, 26.09 написала в Claude, що збирається «розстріляти» офіс шерифа округу Лі. Пізніше вона пояснила, що користується чатботом як «щоденником». Системи безпеки Anthropic позначили запис, його переглянула людина, визнала загрозу реальною і передала поліції. Жінку затримали вдома без інцидентів, їй інкримінують письмову погрозу насильством, це тяжкий злочин другого ступеня за законом Флориди.

Anthropic у правилах прямо пише, що може передати дані користувача у виняткових випадках, коли вважає це необхідним, щоб запобігти смерті чи тяжкій шкоді. TechSpot нагадує контраст: OpenAI судиться з Британською Колумбією в тому числі через те, що не повідомила поліцію про майбутню стрільбу, бо розмови не досягли порогу для передачі.

На HN суперечка зводиться до двох питань: чи взагалі закон про письмові погрози має працювати для тексту, який бачить лише машина, і де має бути поріг, після якого компанія дзвонить у поліцію.

Чому це важливо. Розмова з чатботом юридично ближча до листа в компанію, ніж до записника:

її можуть прочитати модератори і передати владі. Пороги в компаній різні і зсуваються після кожного резонансного випадку, тому про приватність чату варто думати за найсуворішим сценарієм.

ТЕМА 6

ChatGPT підписує фальшиві карикатури іменами справжніх художників New Yorker

ДЖЕРЕЛА Nieman Lab, 05.10 [Niemanlab](#) · обговорення на HN [Hacker News](#)

[одне джерело] Nieman Lab задокументувала понад 15 карикатуристів New Yorker, чиї підписи ChatGPT ставить на згенеровані картинки «в стилі New Yorker». Найвідоміший випадок: карикатура з Доллі Партон і Тімом Каррі на небесах, яка після смерті обох у

серпні набрала 25 тис. лайків в одному твіті. У кутку стояв підпис «BLOPER» Брендана Лопера, хоча він її не малював. Фанатка просто попросила ChatGPT зробити «карикатуру в стилі New Yorker». Серед підписів у тестах видання є Емілі Флейк, Гаррі Блісс і навіть покійні Джордж Бут і Сол Стейнберг.

Condé Nast у 2024 році уклала з OpenAI ліцензійну угоду, але, за словами речника New Yorker, права навчати моделі на карикатурах вона не давала, а стандартні контракти художників цього не дозволяють. Після запиту журналістки ChatGPT почав інколи відповідати, що запит «може порушувати обмеження щодо схожості зі стороннім контентом», але на момент публікації підписи все ще з'являлись. OpenAI назвала це помилкою моделі. Юрист з Корнелла Джеймс Гріммельманн каже, що авторське право тут слабкий аргумент, бо копіюється стиль, а не конкретна робота, а от право на власне ім'я може спрацювати, якщо доведуть комерційне використання.

Чому це важливо. Підпис для художника працює як водяний знак, і модель відтворює саме його. Це та сама проблема, що з текстовим знаком у пункті 3, тільки навпаки: сигнал походження, який модель підробляє, перестає щось доводити.

ТЕМА 7

Агенти на Claude Opus 5.5 знайшли два кандидати в магнітні напівпровідники для пам'яті

ДЖЕРЕЛА Vals AI, 05.10 Vals · обговорення на HN [Hacker News](#)

[одне джерело] Команда агентів на Claude Opus 5.5 під керівництвом дослідника Vals AI шукала матеріали для спінтроніки: антиферомагнетики з нульовою сумарною намагніченістю, які при цьому сортують електрони за спіном. Такі матеріали можна щільніше пакувати і швидше перемикати, ніж звичайні магніти. Агенти рахували кристали стандартним методом теорії функціоналу густини (DFT).

Кандидатів два. YBaMnFeO_5 агенти придумали самі: прогнозована ширина забороненої зони 2,35 еВ і магнетизм приблизно до 420–490 К. Але ця сполука потребує ідеальної шахової укладки атомів, а симуляція показала, що та розпадається близько 950 К, тоді як синтез таких оксидів іде при 900–1300 °С. Тобто зробити корисну форму, найімовірніше, важко. Другий кандидат, $\text{KV}[\text{Cr}(\text{CN})_6]$, вперше синтезували у 1999 році, і магнітний порядок у нього тримався до 376 К. Нульову намагніченість і навіть спінове сортування в ньому вже бачили у 2008-му, але ніхто не звернув уваги, що це робить його потрібним класом напівпровідника.

Автори самі перелічують застереження: прогнози для ідеального сухого кристала, єдиний зразок із 1999-го був порошком з водою в порах, і два методи розрахунку розходяться, чи переживе вода ефект. Ні заборонену зону, ні сортування ще не вимірювали. На HN перша реакція: «після LK-99 сприймаю з повною вантажівкою солі», а інші зауважують, що «відкриття» тут означає агент, який запускає класичні симуляції і знаходить забуте в літературі.

Чому це важливо. Чесно оформлений приклад того, на що агенти вже здатні в науці: перебрати простір кандидатів і підняти старі роботи, які люди прогавили. Доказом відкриття він стане лише після лабораторного виміру.

ТЕМА 8

Норвегія хоче тимчасово заборонити AI-окуляри в парках, на пляжах і в школах

ДЖЕРЕЛА Financial Times через Ars Technica, 05.10 [Ars Technica](#) · Barron's [Barrons](#)

Норвегія стала першою великою країною, яка пропонує тимчасово заборонити окуляри з AI в окремих публічних місцях: парках, на пляжах, у школах і дитсадках. Законопроект уряд обіцяє внести найближчим часом, паралельно експертна група напише постійні правила. Міністр цифровізації Торгейр Мікаельсен: «Ми не хочемо суспільства, де люди бояться прихованого запису там, де звикли бути без нагляду». Повної заборони не буде: використовувати окуляри можна там, де нема ризику знімати людей без згоди.

Частина країни вже заборонила їх сама: школи Осло і нафтогазова Equinor в офісах і на морських платформах. Австралія розглядає схожі обмеження, а суди в кількох країнах уже не пускають з такими пристроями. Counterpoint Research оцінює витрати на носимі пристрої у 2026-2032 роках у понад \$1 трлн.

Чому це важливо. Meta, Google і OpenAI ставлять на окуляри як на наступний основний інтерфейс до AI, який постійно слухає і бачить. Норвезька модель регулювання за місцями, без загальної заборони, може стати шаблоном для інших країн, і продуктам доведеться знати, де вони знаходяться.

ТЕМА 9

Данія: дані 8,8 млн людей з реєстру CPR витягли через законний доступ приватної компанії

ДЖЕРЕЛА реєстр CPR, 05.10 [Cpr](#) · міністерство цифровізації Данії [Ufm](#) · обговорення на HN [Hacker News](#)

Данський центральний реєстр населення CPR повідомив про серйозний інцидент: сторонні отримали імена, адреси і персональні номери приблизно 8,8 млн людей із приблизно 11 млн записів у реєстрі, серед них живі, ті, хто виїхав, і померлі. Зламу в класичному сенсі не було: використали законний доступ однієї данської приватної компанії, яка за законом може шукати в реєстрі дані про заздалегідь визначене коло людей. Нерегулярну активність протягом вересня помітили лише ввечері п'ятниці, 02.10. Людей із захищеними адресами витік не зачепив.

Доступ компанії перекрили, справу розслідує поліція, повідомлено регулятора з захисту даних, хто за цим стоїть, поки невідомо. Міністерка Крістіна Егелунд назвала інцидент «глибоко серйозним» і доручила повну перевірку безпеки CPR. Міністерство

окремо попереджає: не давати паролів навіть тим, хто вже знає ваше ім'я, адресу і номер.

Чому це важливо. Класичний сценарій зловживання легальним доступом: система контролює, хто має право шукати, але не помічає, що пошуків стало в мільйони разів більше. Для будь-якого API з персональними даними це аргумент обмежувати обсяг і швидкість запитів навіть для партнерів з правом доступу.

ТЕМА 10

Бена Томпсона зламали через спільний доступ до екрана, і першим це помітив Claude

ДЖЕРЕЛА Stratechery, 05.10 [Stratechery](#) · обговорення на HN [Hacker News](#)

Автор Stratechery розповів, що його Mac Mini, який працює без монітора і на якому крутяться лише Claude і Codex, зламали через вразливість CVE-2026-65400 у спільному доступі до екрана macOS (7,1 з 10). Нідерландський центр кібербезпеки попереджав, що її активно експлуатують на машинах з відкритим в інтернет портом 5900: зловмисники отримували root і ставили майнер Monero. Першим змінені файли помітив агент на Claude Code, який сам зупинив виконання команд і повідомив, що обліковий запис тепер може виконувати адмінкоманди без пароля. Далі Томпсон разом із Claude знайшов чотирисекундний проміжок, коли стався злам, зробив інструмент для моніторингу і перевстановив систему.

Головна претензія до Apple. Спільний доступ до екрана Томпсону потрібен майже лише для того, щоб натискати «ОК» у вікнах дозволів ТСС, які невидимі для програм і агентів на тій самій машині.

Ще одне відкриття: галочка «встановлювати оновлення безпеки автоматично» не ставить більшість виправлень вразливостей, бо вони приходять у проміжних версіях macOS.

Тут писали 03.10, що Apple обмежує Full Disk Access через агентів. Томпсон відповідає саме на це:

дозволи працюють на рівні програм, а агент постійно пише нові програми, тому потрібен шар дозволів для агента.

Чому це важливо. Практичні висновки для будь-якого постійно ввімкненого Mac з агентами: не відкривати 5900 в інтернет і заходити через VPN, і перевіряти, що проміжні оновлення справді стоять.

А ширше питання, чи підходить macOS як дім для агентів, Apple поки вирішує в протилежний бік.

misc

- **Банки готують рекордне фінансування чипів на \$60 млрд для Broadcom і Anthropic**, пише FT.

Продовження пункту 02.10 про позику Broadcom на \$42 млрд. текст за пейволом, доступний лише заголовок [одне джерело]. [FT](#)

- **OpenAI тестуватиме рекламу з картинками під час генерації зображень у ChatGPT**, у США з кінця жовтня. Компанія пише, що ChatGPT має 1,2 млрд користувачів на тиждень, і хвалиться цифрами партнерів: у WeightWatchers вартість залучення клієнта на 15,3% нижча, ніж у пошуковій рекламі.

[OpenAI](#)

- **Cloudflare відкрила в беті Web Search API для агентів** з трьома постачальниками пошуку на вибір (Ceramic.ai, Eka, Linkup) і оплатою через AI Gateway без націнки. Анонс від 02.10, на HN він вистрілив учора з 493 балами. [Cloudflare](#)
- **@emollick**: його GPT, яким користувались інші люди, у грудні зникне, перенести можна лише в приватний плагін. Моллік нагадує, що компанії роками вчили працівників робити GPTs.

[@emollick](#)

- **@emollick** переніс складну роботу з Cowork у нові Claude Projects, які працюють з окремою хмарною віртуальною машиною, і каже, що вони кращі, але погано задокументовані.

[@emollick](#)

- **@koomen** за вихідні з Opus 5.5 переніс стару гру Bungie Myth у WebAssembly і пише, що «покинутого софту більше не існує». [@koomen](#)
- **Hugging Face зробила спільний каталог RL-середовищ** на Hub: OpenEnv, Verifiers, Harbor, NeMo Gym в одному місці. Продовження пункту 03.10 про оболонки агентів як середовища.

[@ben_burtenshaw](#)

- **Era від Офіра Ерліха** генерує цілу симульовану компанію з Salesforce, Slack, Jira і Zendesk, щоб тестувати агентів не на живих даних. @levie каже, що саме цього бракує для впровадження агентів.

[@levie](#)

- **@levelsio** порахував шкали рейтингів сайтів бронювання: в Airbnb усе вміщається між 4,5 і 5,0 з медіаною 5. Він зробив hotelist.com, який розтягує ці оцінки на всю шкалу.

[@levelsio](#)

- **a16z випустила сьомий Top 100 споживчих AI-застосунків**, уперше з рейтингом за виручкою.

[@MannyBernabe](#)

- **Нобелівську премію з фізіології та медицини отримали Карл Дайссерот, Петер Хегеманн і Георг Нагель** за оптогенетику, керування нейронами світлом. [Nobelprize](#)

- Джеремі Авігад у гостьовому пості в блозі Теренса Тао про майбутнє математики: AI розв'язує задачі, які можна зібрати з наявних технік, тож математикам варто братися за складніші питання і самим вирішувати, що цікаве. [Wordpress](#)
- Лінкер mold 3.0 переписали на Rust. [GitHub](#)