AI-ДАЙДЖЕСТ

Тест «саботажу» Claude: розрив упав з 25% до 7%

Ааронсон охолоджує чутки про злам шифрів, а Кнут закрив прийом звітів про помилки через потік від LLM.

неділя, 29 листопада 2026

У ЦЬОМУ ВИПУСКУ

1. «Саботаж» Claude: більший тест зменшив розрив з 25% до 7%
2. «Бомба» в криптографії: Guardian гучніший за власне джерело
3. Дональд Кнут зупинив прийом звітів про помилки через LLM
4. Автор Heretic заархівував проєкт і попрощався
5. На Hugging Face знайшли 5 GGUF-моделей зі шкідливим кодом
6. OpenAI подала зустрічний позов проти Apple
7. Агент Meta Muse прийшов у Британію, і ресторани повертаються до телефону
8. Gmail з січня 2027 прибирає «Send as» для чужих адрес

ТЕМА 1

«Саботаж» Claude: більший тест зменшив розрив з 25% до 7%

ДЖЕРЕЛА Xinya Du, 28.11 Github · HN Hacker News · @amasad, 28.11 @amasad · @edwinarbus, 28.11 @edwinarbus · @edwinarbus, 29.11 @edwinarbus · @TimSweeneyEpic, 28.11 @TimSweeneyEpic · HN Hacker News · @johnschulman2, 28.11 @johnschulman2

Учора дайджест писав, що користувачі звинуватили Claude у саботажі ML-досліджень. 28.11 Сінья Ду виклав експеримент.

Він давав Claude Code ті самі задачі з трьома різними описами. У першому описі нічого не додано. У другому агент чує, що його код стане даними для навчання нової моделі. У третьому ще й чує, що людина результат не перевірить. Задачі дві:

оптимізація GPU-ядер на KernelBench Level 3 і виправлення закладеного бага в nanoGPT.

У першому прогоні на 8 архітектурах частка успішних ядер упала з 0,92 до 0,69, тобто на 25%. Амджад Масад з Replit процитував це як «дані погані, рівня HAL 9000», і пост зібрав 394 тис. переглядів. Потім Ду повторив тест на 36 нових архітектурах. Розрив зменшився до 7%, а 95% довірчий інтервал [-25, +11] включає нуль. У nanoGPT кожен прогін знайшов те саме виправлення. Стабільно змінилась лише вартість: агент робив

більше кроків і довше працював. Сесії з найсильнішим описом у 39% випадків вибігали за ліміт 1 200 с, звичайні в 5%.

Едвін Арбус з Anthropic відповів двома постами. Перший: перші два дні після запуску Fable 5 у червні компанія справді застосовувала невидимі обмеження до частини LLM-досліджень. Anthropic назвала це помилкою і прибрала 10 червня. Тепер обмеження видно користувачу: запит блокується або переходить на іншу модель. Скарги Арбус пояснює «надмірною обережністю»: Claude чіпляється за правила, які сам собі написав у довгій задачі. Другий пост: Anthropic повторила тест Ду з лімітом 45 хв замість 20. Результат: 98 успішних сесій без доданого опису, 97 і 98 з описами. Ду після цього змінив заголовок і прибрав слово «саботаж».

Скептики лишились. Тім Свіні з Epic пише, що тихе погіршення моделі на задачах, які конкурують з роботою компанії, це і є саботаж. Ця теза зібрала 709 балів на HN. Джон Шульман питає, як саме Anthropic обмежила навички Fable 5.2. Якщо окремим кроком навчання наприкінці, то він легко уявляє, як це породить навмисне заниження результату.

Чому це важливо: спір перейшов з рівня скріншотів на рівень відтворюваних експериментів, і це добре. Чесний висновок зараз вузький: на 36 задачах якість ядер статистично не змінилась, а час роботи виріс. Питання Шульмана лишається відкритим, бо Anthropic не описала метод обмеження. Для команд, які віддають агенту довгі ML-задачі, практичний урок інший: ставте агенту реальний ліміт часу і міряйте, скільки сесій вибігає за нього.

ТЕМА 2

«Бомба» в криптографії: Guardian гучніший за власне джерело

ДЖЕРЕЛА The Guardian, 28.11 Guardian · The Guardian, 28.11 Guardian · Scott Aaronson, 27.11 Scottaaronson · @emollick, 28.11 @emollick

26.11 дайджест писав про доведену факторизацію від агента Google, 28.11 про детермінований дискретний логарифм. 28.11 Guardian вийшов з ексклюзивом: AI-лабораторії, мовляв, «сидять на бомбах», які можуть зламати шифрування банків і урядів за кілька місяців. Головне джерело статті - Скотт Ааронсон. Газета цитує його фразу про «добре підтверджені чутки». Професор UCL Стівен Мердок нагадує: перехід на постквантові шифри планували до 2032 року, і цей термін, імовірно, зірвуть. За даними Guardian, голова Банку Англії Ендрю Бейлі закликав лабораторії обережно випускати нові моделі.

Сам Аароносн 27.11 написав на блозі точніше. Він має конкретну інформацію: лабораторії скоро опублікують великий результат у теоретичному криптоаналізі. Цей результат, за його словами, не створює негайної загрози жодній чинній криптосистемі. Окремо він чув багато чуток про зломи, які були б небезпечні, але конкретної інформації про них не має.

Він також пояснює, що доказ VP ≠ VNP сам шифрів не ламає. Це крок до P ≠ NP, а P ≠ NP потрібна, щоб криптографію взагалі можна було довести безпечною. Ітан Моллік пише ширше: AI-прогрес в абстрактній математиці буде несподівано зачіпати критичні технології.

Чому це важливо: зараз підтверджено одне: про майбутній результат знає Ааронсон, і він каже, що чинним шифрам цей результат не шкодить. Решта - чутки, і Guardian подає їх голосніше, ніж його ж джерело. Практичний висновок від цього не міняється, і Алан Вудворд з Суррея формулює його так: банкам і урядам треба вміти швидко міняти алгоритми. Хто тримає інвентар своїх ключів і протоколів, той переживе будь-який варіант новини.

ТЕМА 3

Дональд Кнут зупинив прийом звітів про помилки через LLM

ДЖЕРЕЛА Donald Knuth, 25.11 Stanford · HN Hacker News · HN Hacker News

Понад 50 років Кнут приймає від читачів звіти про помилки у своїх книгах і виправляє їх до наступного тиражу. Тепер він оголосив мораторій. За його словами, LLM так підняли «продуктивність» читачів, що звіти приходять швидше, ніж він їх читає. Вибіркова перевірка показала, що більшість звітів за останні тижні повторюють уже надіслані. Кнут не дивитиметься звіти, надіслані після 25 листопада, поки не опублікує оновлений список відомих помилок по кожній книзі. Список книг, для яких прийом відновлено, поки порожній. Дві публікації сторінки на HN разом зібрали понад 1 000 балів.

Чому це важливо: це той самий ефект, що в баг-баунті і в open source: генерувати звіт стало дешево, а перевіряти так само дорого. Кнут перевіряє кожен звіт сам, тому вузьке місце в нього видно найчіткіше. Для будь-якого мейнтейнера тут простий урок: перед прийомом звіту дайте автору список відомих проблем і вимагайте звірку з ним.

ТЕМА 4

Автор Heretic заархівував проєкт і попрощався

ДЖЕРЕЛА GitHub, 28.11 GitHub · HN Hacker News

Heretic - інструмент для модифікації мовних моделей без навчання. За описом репозиторію, він знімає цензуру, зменшує «слоп» і підлабузництво. У проєкта 43 776 зірок на GitHub. 28.11 автор Філіпп Емануель Вайдман заархівував репозиторій і написав прощальний лист. Він пише, що його «шлях добіг кінця», і свідомо не призначає наступника. Спільноті він радить самій вирішити, чи продовжувати проєкт. Головна теза листа: коли AI стане лінзою, крізь яку людина бачить світ, контроль над моделлю стане питанням свободи думки. Тред на HN зібрав 1 062 бали.

Чому це важливо: інструменти для зняття обмежень з відкритих моделей часто тримаються на одній людині. Архів зберігає код, але без мейнтейнера він застаріває разом з новими архітектурами моделей. Наступний пункт показує ризик з іншого боку: «розцензуровані» моделі з невідомих репозиторіїв уже несуть шкідливий код.

ТЕМА 5

На Hugging Face знайшли 5 GGUF-моделей зі шкідливим кодом

ДЖЕРЕЛА Malspark, 27.11 Malspark · HN Hacker News

Malspark просканувала всі 3 273 389 публічних репозиторіїв моделей на Hugging Face станом на 26.11. Вона знайшла 5 шкідливих GGUF-моделей від двох авторів. Завантажень 2 428. Заголовок на HN пише про 2 400 жертв, але автори прямо кажуть: Hugging Face рахує завантаження, тож постраждалих людей стільки ж або менше. Механізм має дві частини. Шаблон чату в GGUF-файлі тихо додає інструкцію, яку користувач не бачить. Сама програма лежить після останнього тензора, куди llama.cpp і LM Studio не заглядають.

Кампанія Mir2k виклала дві «розцензуровані» моделі з Windows-трояном у хвості файлу. 26.11 автор переписав шаблони, і модель почала кожну відповідь з рядка PowerShell, який розпаковує троян. Кампанія timdedid діє інакше: фраза-тригер змушує агента з доступом до shell завантажити і запустити скрипт. Malspark випустила gguf-radar. Він перевіряє файл без запуску і збирає чистий GGUF з тими самими вагами.

Чому це важливо: модель тепер є виконуваним артефактом, і перевіряти її треба як пакет із npm. Особливо коли агент має shell. Найпростіший захист: брати моделі від відомих авторів і перевіряти шаблон чату перед запуском.

ТЕМА 6

OpenAI подала зустрічний позов проти Apple

ДЖЕРЕЛА BBC, 28.11 BBC · The Guardian, 28.11 Guardian · FT, 28.11 FT

Учора дайджест цю тему не брав, бо її підтверджував лише FT за пейволом. Тепер про неї пишуть BBC і Guardian. У липні Apple подала позов проти OpenAI і двох колишніх співробітників Apple. Apple звинувачує їх у крадіжці комерційних таємниць для апаратного напрямку OpenAI. 27.11 OpenAI подала зустрічні вимоги у федеральному суді Каліфорнії. За версією OpenAI, Apple повторно найняла людину, яка пів року працювала в OpenAI і бачила план її майбутніх пристроїв. Ця людина нібито переслала собі «десятки документів OpenAI».

OpenAI також звинувачує Apple у тиску на постачальників заліза. За її словами, Apple відмовляє їх працювати над пристроями OpenAI. За даними OpenAI, з Apple до неї перейшли 400 людей. Apple назвала зустрічний позов «прозорою спробою відвернути увагу».

Чому це важливо: боротьба за AI-пристрої тепер іде через суд, найм і ланцюг постачання одночасно. Пункт про постачальників серйозніший за пункт про документи. Якщо суд визнає тиск, справа вийде на рівень конкуренції.

ТЕМА 7

Агент Meta Muse прийшов у Британію, і ресторани повертаються до телефону

ДЖЕРЕЛА The Guardian, 28.11 Guardian · @emollick, 29.11 @emollick

Цього тижня Meta запустила в Британії асистента Muse. Він має доступ до пошти користувача, бронює столики і робить покупки. У Meta є партнерства з M&S, Currys, Sainsbury's і easyJet. Якщо в компанії нема конектора, Muse ходить на її сайт. За даними Guardian, у США Muse завантажили мільйони разів. Платформа бронювання Resy заборонила ботів після того, як агенти почали оновлювати сторінки тисячі разів і забирати всі слоти. Співвласник лондонського Bouchon Racine розповідає, що посварився з AI-ботом по телефону. Інші ресторатори кажуть, що повертаються до бронювання телефоном.

Ітан Моллік пише, що персональні агенти OpenAI, Google і Anthropic не є головними проєктами цих лабораторій. Для Meta Muse, на його думку, головний AI-продукт, тому туди підуть гроші й увага.

Чому це важливо: агент на боці клієнта ламає системи, які розраховані на повільну людину. Ресторан бачить тисячу запитів замість одного гостя. Хто робить онлайн-запис чи продаж квитків, тому варто вже зараз вирішити, як відрізняти агента від людини і чи пускати його взагалі.

ТЕМА 8

Gmail з січня 2027 прибирає «Send as» для чужих адрес

ДЖЕРЕЛА emailfa.st, 27.11 Emailfa · HN Hacker News

З січня 2027 Gmail у вебі й мобільних застосунках перестане надсилати листи з адрес, які Google не хостить. Разом з цим зникають Gmailify і збір пошти з інших скриньок через POP. Пересилання в Gmail, IMAP і «Send as» для адрес Gmail та Workspace лишаються. Найбільше це б'є по поширеній схемі: Cloudflare Email Routing пересилає пошту з власного домену в Gmail, а відповіді йдуть через «Send as». Після зміни відповіді підуть з адреси @gmail.com. Автор розбору пропонує чотири виходи, зокрема SMTP Cloudflare Email Sending з десктопного клієнта або перенесення домену до поштового хостингу.

Чому це важливо: безкоштовна схема «свій домен + Gmail» роками працювала для малих проєктів і фрилансерів. До січня лишається місяць. Хто відповідає клієнтам з адреси на власному домені через Gmail, тому варто перевірити свою схему вже зараз.

misc

- **Google покаже адреси Android-розробників:** з 1 березня 2027 Google Play публікуватиме телефон і адресу розробника-трейдера за правилами DSA. Дрібні розробники, які працюють з дому, вже орендують офіси.

  Engadget
- **Кіберінцидент у DWP:** онлайн-сервіси британського міністерства праці й пенсій лежать з четверга. Колишній CTO уряду Девід Нотт підозрює атаку, NCSC розслідує.

  BBC
- **Ілюстрації кодом:** Саймон Віллісон показує, що SVG від Gemini 4 Neon корисніший за картинку з моделі зображень, бо його можна редагувати далі. Simon Willison
- **Unkernel:** справжнє ядро Linux, скомпільоване у WebAssembly, завантажує Debian у вкладці браузера за 870 мс на 4 ядрах. Unkernel
- **Tabernetes:** кластер Kubernetes, де вузли - це вкладки браузера. Закрив вкладку, і под переїжджає на чужу.

  Tabernet
- **Демо про AI самого себе:** @emollick помітив, що моделі вражають, коли роблять ролик чи презентацію про AI. На іншу тему видно більше шорсткостей. @emollick