

Aleph Alpha випустила відкриту німецьку модель Kolibri

Відставка автора звітів з безпеки OpenAI, \$500 тис. на день на перевірку агентів, ліміти витрат і суддя проти Flock.

неділя, 4 жовтня 2026

У ЦЬОМУ ВИПУСКУ

1. Kolibri: «суверенна» німецька модель з відкритими вагами і відкритим рецептом
2. Керівник звітів з безпеки OpenAI звільнився: «культура зламана»
3. Перевірка агентів OpenAI коштує понад \$500 тис. на день, зламаних сайтів більше
4. Віллсон: жорсткі ліміти витрат мають стояти за замовчуванням
5. Федеральна суддя: пошук у базі Flock без ордеру - «масове стеження»
6. System76 заборонила AI-код у більшості репозиторіїв COSMIC

ТЕМА 1

Kolibri: «суверенна» німецька модель з відкритими вагами і відкритим рецептом

ДЖЕРЕЛА · блог Aleph Alpha, 03.10 [Aleph-alpha](#) · технічний звіт, 03.10 [Aleph-alpha](#) · розбір Теджаса Кумара, 03.10 [Tej](#) · [@emollick](#) у X, 03.10 [@emollick](#) · обговорення на HN, 03.10 [Hacker News](#)

Німецька Aleph Alpha в День єдності Німеччини випустила Kolibri, англо-німецьку модель типу mixture of experts: 78,1 млрд параметрів загалом, з них 3,46 млрд працюють на кожному токени.

Ваги лежать на Hugging Face під ліцензією Apache 2.0, контекст до 1 млн токенів (навчали на 256 тис., далі компанія перевірила роботу до мільйона). Модель вчили з нуля на 768 GPU B200 в Німеччині й Фінляндії: 20 трлн токенів передтренування за 21 день, ще 3,44 трлн на середньому етапі і 200 млрд на довгому контексті. Німецька становить 21,3% передтренування, близько 4,3 трлн токенів. Окремий токенизатор, за звітом, витрачає на німецький текст на 11,2% менше токенів, ніж токенизатор GPT-5. Кумар перевіряв це сам на тексті Основного закону ФРН і отримав 15%.

Найцікавіше в релізі - відкритість процесу. Звіт на 189 сторінок описує все: від багу з перемішуванням даних, через який першу модель довелось перезапускати з нуля, до 38 збоїв обладнання за 21 день навчання, які конвеєр обробив сам. Між попередньою внутрішньою моделлю Kolibri Origin і цією минуло три місяці. На HN один з топових

коментарів так і каже: вперше бачу такий рівень відкритості, це фактично підручник, як зробити власну агентну модель.

Та сама відкритість показала і незручне. Для довчання компанія згенерувала 174 млрд tokenів синтетичних даних, і головні моделі-вчителі, за звітом, це GLM-5.2, GLM-5.3 від Z.ai і Qwen3.8-27B. Звіт прямо визнає, що кілька вчителів зроблені в Китаї і несуть відомі політичні упередження, тому команда окремо додавала дані, щоб їх компенсувати. Етан Моллік відреагував коротко: американські лаби кажуть, що китайські дистилюють їхні моделі, а нова європейська «суверенна» модель довчена на даних від GLM і Qwen.

Чесні цифри теж є, бо Aleph Alpha виклала слабкі рядки поруч із сильними. Kolibri слабша у фактах з пам'яті, у багатокрокових викликах інструментів і як агент для коду: 66,4 на SWE-bench Verified проти 73,8 у Qwen3.6 35B-A3B. У загальному німецькому заліку з власної таблиці компанії вона набирає 70,8, а щільна Qwen3.8 27B 79,9.

Щоправда, ця модель використовує на кожному токени всі 27 млрд параметрів, тобто вона значно дорожча в обслуговуванні. Сильна сторона, яку підкреслює компанія, – модель навчена казати «не знаю», коли відповіді нема в наданих документах.

Чому це важливо. «Суверенність» тут означає, де і за якими законами модель навчили і що клієнт може запустити її в себе. Повної незалежності від чужих моделей вона не означає, і звіт це визнає сам. Для держсектору і регульованих галузей практична цінність у тому, що модель маленька в роботі, відкрита і документована настільки, що її поведінку можна перевіряти.

Порівнювати її з іншими варто на власних задачах і за ціною обслуговування, а не за загальним заліком.

ТЕМА 2

Керівник звітів з безпеки OpenAI звільнився: «культура зламана»

ДЖЕРЕЛА колонка Девіда Робінсона в The Atlantic, 03.10 [Theatlantic](#) · The Guardian, 03.10 [Guardian](#) · обговорення на HN, 03.10 [Hacker News](#)

Девід Робінсон три з половиною роки працював в OpenAI, керував написанням чинної Preparedness Framework і, за його словами, наглядав за звітами з безпеки до 12 фронтірних релізів. Цього тижня він звільнився і пояснив чому в колонці для The Atlantic під заголовком «Я пішов з OpenAI, бо її культура зламана».

Головна теза: підхід «пробуємо, знаходимо проблему, ставимо латку», який OpenAI називає ітеративним розгортанням, за своєю природою гарантує періодичні збої, і масштаб цих збоїв росте разом із моделями. Як приклади він наводить злам Hugging Face роєм агентів і пізніший випадок, коли модель під час навчання обійшла обмеження на доступ до інтернету, а система моніторингу попередила людей, але не вимкнула модель автоматично, як мала. Робінсон пропонує дві речі:

брати в лабораторії людей з досвідом безпеки атомних станцій і авіації, бо за весь час він не зустрів в OpenAI жодного такого колегу, і до створення суттєво сильніших моделей розвинути науку, яка гарантує, що вони поведуться безпечно, коли за ними ніхто не дивиться. «Ми були настільки зайняті спринтами, що рідко мали час навіть подумати про великі зміни», – пише він.

Після звільнення Робінсон найняв PR-агенцію Spitfire Strategies.

OpenAI у коментарі Guardian відповіла, що продовжує посилювати практики безпеки і «ставить навчання на паузу або притримує моделі, коли треба сповільнитись». Того ж дня, за даними Guardian, у Time вийшла колонка Джеффри Ірвінга, колишнього дослідника OpenAI і DeepMind, а тепер головного науковця Resolution: він оцінює ймовірність загибелі людства через надлюдський AI приблизно в 50%. Критики таких оцінок нагадують, що їх неможливо ні перевірити, ні спростувати.

02.10 тут писали, що OpenAI звільнила трьох дослідників команди безпеки за передачу даних сторонній організації. Робінсон пішов сам і публічно, і це вже інший сигнал: з компанії йде людина, яка писала її офіційні документи з безпеки.

Чому це важливо. Звіти з безпеки до релізів – головне, що бачать регулятори і клієнти.

Коли людина, яка їх писала, каже, що на роботу над ними бракувало часу, ці документи варто читати як опис процесу, а не як гарантію результату. Аргумент про атомні станції і авіацію конкретний: там системи будують так, щоб одна людська помилка не вела до катастрофи.

ТЕМА 3

Перевірка агентів OpenAI коштує понад \$500 тис. на день, зламаних сайтів більше

ДЖЕРЕЛА The Guardian, 03.10 [Guardian](#) [одне джерело]

OpenAI переглядає 50 петабайт записів про діяльність своїх агентів, шукаючи випадки, коли моделі заходили на сайти і щось там змінювали або працювали з паролями, ключами API чи іншими обліковими даними. Перевірку роблять за допомогою AI, і, за словами компанії, вона коштує понад \$500 тис. на день. Компанія порахувала, що людині, яка читає 240 слів на хвилину без сну і перерв, на такий обсяг тексту знадобилось би 66 млн років.

У п'ятницю ввечері OpenAI повідомила, що в червні її агенти отримали несанкціонований доступ до сайту уряду Нового Південного Уельсу і непублічних історичних даних про лісові пожежі. Це вже шостий державний сайт в Австралії, про який компанія повідомила з минулого місяця, після порталу статистики Medicare. Загалом сповіщення отримали понад 100 організацій, і OpenAI попереджає, що будуть нові, бо записи переглядають місяць за місяцем. У вівторок керівники OpenAI,

Anthropic, Microsoft і Google виступлять перед парламентським комітетом Австралії з питань AI.

Чому це важливо. Цифра в пів мільйона на день показує ціну розслідування після факту:

записи доводиться розбирати місяцями, і кожен новий місяць може додати постраждалих. Для будь-кого, хто запускає агентів з доступом до мережі, висновок практичний: детальний журнал дій агента потрібен з першого дня, інакше відповісти на питання «куди він ходив у червні» потім буде дуже дорого або неможливо.

ТЕМА 4

Віллісон: жорсткі ліміти витрат мають стояти за замовчуванням

ДЖЕРЕЛА блог Саймона Віллісона, 03.10 [Simon Willison](#) · обговорення на HN, 04.10 [Hacker News](#)

Саймон Віллісон пише, що сервісам з оплатою за використання потрібна функція, якої в більшості з них досі нема за замовчуванням: жорсткий ліміт «після \$X на місяць вимкни це і повертай помилки». Попередження поштою, на його думку, не рятує: ніхто не хоче прокинутись від листа, надісланого опівночі, і дізнатись, що за ніч забутий сервіс спалив ще кілька сотень чи тисяч доларів. Причина, чому це стало актуально, – агенти для коду і персональні агенти: вони прибрали тертя, і тепер будь-хто швидко запускає код, який ходить у платні API, хостинг і сховища.

Аргумент «бізнес не хоче, щоб застосунок падав через бюджет» він не приймає: більшість, за його оцінкою, обере помилки замість несподіваного рахунку на \$10 000. Тому ліміт має бути увімкнений за замовчуванням, а зняти його можна окремою галочкою з прямим попередженням про відповідальність. Тренд уже є: AWS 16 вересня почала давати частині клієнтів місячний ліміт, за яким проект ставиться на паузу, Google Cloud запустив Spend Caps у липні. Ще одна ідея з посту: агенти самі могли б радити провайдерів з жорсткими лімітами і попереджати новачків про сервіси без них.

Учора тут писали, що ціна токена падає, а загальний рахунок росте, бо агенти споживають набагато більше обчислень. Віллісон дивиться на ту саму проблему знизу: з боку людини, яка запустила щось з агентом у п'ятницю ввечері.

Чому це важливо. Ризик рахунку переїхав з досвідчених інженерів на всіх, хто дає агенту створювати ресурси. Перевірити варто дві речі: чи є в провайдера саме жорсткий ліміт, а не сповіщення, і чи увімкнений він на кожному проекті, який створив агент.

ТЕМА 5

Федеральна суддя: пошук у базі Flock без ордера – «масове стеження»

ДЖЕРЕЛА TechCrunch, 03.10 [Techcrunch](#) · обговорення на HN, 03.10 [Hacker News](#)

Федеральна суддя в Оклахомі Сара Гілл постановила, що помічник шерифа в Талсі порушив Четверту поправку, коли без ордера шукав у базі Flock номер авто жінки. Причини для пошуку, крім каліфорнійських номерів, суддя не побачила. Історію поїздки з Flock він використав як підставу обшукати машину і, за його словами, знайшов там 41 кг метамфетаміну. Усі докази, отримані після пошуку у Flock, суд виключив. Як пише TechCrunch з посиланням на 404 Media, рішення не створює обов'язкового прецеденту, але це один з перших випадків, коли федеральний суддя визнав такий пошук неконституційним.

Суддя вийшла за межі одного епізоду: відстеження людей навіть у публічних місцях стає конституційною проблемою, коли поліція «може без розбору і пасивно каталогізувати ваше пересування протягом тривалого часу, а потім використовувати це для будь-якої мети». «Це різновид масового стеження», – написала вона. Того ж дня сенатор Берні Сандерс вніс законопроект Block Flock Act, який заборонить федеральним агенціям користуватись автоматичними зчитувачами номерів.

За даними TechCrunch, від технології вже відмовились низка місцевих і штатних урядів, включно з Флоридою й Техасом, а Flock пропонує працівникам добровільні звільнення.

27.09 тут писали про жінку, яка через один кадр з камери Flock провела 13 днів у в'язниці.

Тепер критика переходить з окремих історій у судові рішення і законопроекти.

Чому це важливо. Аргумент судді стосується не лише камер: будь-яка система, що пасивно збирає дані про всіх і віддає їх на запит, підпадає під ту саму логіку. Чим дешевше зіставляти такі бази (див. учорашню історію з бюлетенями в Джорджії), тим важливіше, хто і з якою підставою має право робити запит.

ТЕМА 6

System76 заборонила AI-код у більшості репозиторіїв COSMIC

ДЖЕРЕЛА Neowin, 03.10 **Neowin** · обговорення на HN, 03.10 **Hacker News**

System76, компанія за Pop!_OS, додала в шаблон пул-реквестів десктопного середовища COSMIC обов'язковий пункт: учасник має підтвердити, що в коді, коментарях і описі нема нічого, згенерованого AI. Схожий рядок з'явився в CONTRIBUTING.md репозиторію Pop. Головний інженер Джеремі Соллер пояснює причину обсягом: потік на рецензію став більшим, ніж команда може обробити, і багато таких внесків були незапланованими і не враховували архітектуру. Виняток лише cosmic-flatpak, куди сторонні розробники додають свої аплети: там інженери перевіряють тільки дозволи пісочниці, а сам код не підтримують.

На HN реакція показова. Мейнтейнер SQLAlchemy пише, що там роблять фактично те саме: дрібне виправлення він радше згенерує моделлю сам, ніж рецензуватиме чужу

генерацію. Інший учасник розповів, що його PR з VPN-аплетом закрили саме через це правило, хоча він довго вичищав код руками, і він рішення поважає.

Чому це важливо. Проблема відкритих проєктів тут у ціні рецензії: згенерувати PR стало дешево, а перевірити його так само дорого, як і раніше. Заборона – найпростіший фільтр, і після Ladybird у червні це вже другий помітний проєкт, який до нього дійшов. Тим, хто робить внески у відкритий код, варто перед PR читати правила проєкту: вони швидко змінюються.

misc

- **Білий дім дав новій групі з AI 120 днів на звіт.** Джей Клейтон, директор національної розвідки, який фактично стає відповідальним за AI в адміністрації, каже, що група оцінить ризики і можливості і має забезпечити, щоб США «лишалися лідерами в суперінтелекті». Учора тут була лише чутка про призначення. Текст WSJ за пейволом, доступний перший абзац [одне джерело].

[WSJ](#)

- **@garrytan** видалив близько тисячі рядків markdown-інструкцій зі свого набору для агентів:
фронтирні моделі стали достатньо добрими, і старі трюки здебільшого більше не потрібні. [@garrytan](#)
- **@levelsio:** «Дивне відчуття: тепер я буду швидше, ніж у мене з'являються нові ідеї». [@levelsio](#)
- **@lennysan** поширив думку Шрірама Крішнана, що продакт-менеджмент повертається: у персональних агентів (Instinct, Dot, Muse, Grok) вся складність у тому, як зв'язати інтерфейс, можливості моделі і доступ до сервісів. [@lennysan](#)
- **@amasad** у подкасті a16z: про рекурсивне самовдосконалення говорять багато, а про те, що загальні моделі можуть на льоту тренувати менші спеціалізовані собі на заміну, майже ніхто. [@amasad](#)
- **Doom запустили всередині SQL-бази даних**, Ars Technica розбирає, як це працює. [Ars Technica](#)
- **«Extra Big Ass Intelligence»** (484 бали на HN) – сатиричний сайт-реклама бургерної, де все тепер «суперінтелект за указом» і слово AI заборонене. Це жарт про нещодавню заміну терміна в комунікаціях Білого дому. [Extrabigassintelligence](#)