AI-ДАЙДЖЕСТ

OpenAI показала перплексію своїх моделей

Dragos знайшов злами провайдерів інференсу, 87% енергії Anthropic даватиме газ, Replit порівняла команди агентів.

понеділок, 23 листопада 2026

У ЦЬОМУ ВИПУСКУ

1. OpenAI опублікувала перплексію своїх моделей і пропонує спільні ліміти для галузі
2. 2 млн промптів-пасток: Dragos знайшов злами трьох провайдерів, OpenRouter відключив трьох постачальників
3. Better Data Center Project: 87% енергії Anthropic у 2028 році даватиме газ, 11% вугілля
4. Сюжет Nvidia: сенатор вимагає допиту Бессента і Лутніка, Бессент не виключає повернення Хуанга
5. Replit: 32 агенти Opus прискорили NumPy у 18,9 раза, команди GPT-6.2 вийшли гіршими за одного агента
6. Збій Delta в Атланті і пояснення хвилі збоїв у США
7. GL.iNet: через діру в хмарі зловмисник поставив шкідливе ПЗ на роутери і KVM [одне джерело]
8. Південна Корея: міністр освіти пішов у відставку через скандал з AI-списуванням [одне джерело]

ТЕМА 1

OpenAI опублікувала перплексію своїх моделей і пропонує спільні ліміти для галузі

ДЖЕРЕЛА OpenAI, 22.11 OpenAI · @merettm, 22.11 @merettm · @gdb, 22.11 @gdb · HN Hacker News

Якуб Пахоцкі, головний науковець OpenAI, опублікував есей «Perplexity and Intelligence». Перплексія показує, наскільки добре модель передбачає наступний токен. Що вона нижча, то краще модель стискає дані. OpenAI вперше показала графік цієї метрики для своїх базових моделей, від GPT-3.5 до внутрішньої моделі, яка ще не вийшла. Пахоцкі визнає, що ці цифри дають конкурентам ясну картину. Але, за його словами, спільні правила безпеки зараз важливіші.

Цифри з есею такі. GPT-3.5 мала log-перплексію близько 0,99 і вміла вести розмову. GPT-4 мала близько 0,91 і вже корисно міркувала ланцюжком думок. Нижче 0,8 з'явились «трансформаційні» здібності в кібербезпеці й математиці. На моделі з показником близько 0,76 OpenAI почала отримувати проривні результати в відкритих

математичних задачах. Наступне вікно нових здібностей Пахоцкі очікує на рівні 0,68-0,7. Що саме з'явиться там, він не знає. Вимір іде на закритому наборі з внутрішнього коду OpenAI.

Найгостріша частина есею стосується темпу. На графіку видно різке прискорення в 2026 році. OpenAI пояснює його переважно довгими дослідженнями і обчисленнями, і лише меншою мірою автоматизацією. Але Пахоцкі пише, що вже в 2027 році AI може дати «лавину» фундаментальних покращень у самому пре-тренуванні. Прогрес у вирівнюванні й моніторингу, на його думку, за цим не встигне. Тому він пропонує спільну межу: лаби не переходять певний рівень перплексії, поки не досягнуть конкретної мети в розумінні моделей. Приклад такої мети - наступник моніторів ланцюжка думок.

Есей сам називає слабкі місця метрики. Лаби міряють на різних наборах даних. Перплексія не враховує міркування під час відповіді. І ніхто не знає, де її нижня межа. Грег Брокман підтримав текст і назвав його основою для спільних планок безпеки. Тред на HN зібрав 519 балів.

Чому це важливо: досі лаби домовлялись про безпеку через тести на конкретні здібності. Такі тести швидко застарівають, і модель вчиться їх розпізнавати. Пахоцкі пропонує міряти саму «сировину» інтелекту до будь-якого донавчання. Пропозиція запрацює, лише якщо інші лаби опублікують свої цифри на порівнянних наборах. Поки цього не зробив ніхто.

ТЕМА 2

2 млн промптів-пасток: Dragos знайшов злами трьох провайдерів, OpenRouter відключив трьох постачальників

ДЖЕРЕЛА Dragos, 19.11 Dragos · OpenRouter, 22.11 Openrouter · HN Hacker News · HN Hacker News

HD Moore, автор Metasploit і зараз керівник досліджень у Dragos, з липня по листопад відправив близько 2 млн промптів до AI-сервісів. Кожен промпт мав унікальні пастки: URL, DNS-ім'я і справжній ключ AWS від порожнього акаунта. Будь-яке звернення до пастки означає, що хтось прочитав промпт. Звіт вийшов 19.11. 22.11 OpenRouter відповів своїм оголошенням.

Знахідок три. Перша: атакувальники зламали трьох провайдерів інференсу і копіювали запити на вході. В одному випадку вкрали кілька сотень ключів AWS, і ключі перевіряли вже за кілька годин. Далі атакувальник шукав дані в S3 і Secrets Manager і пробував увімкнути моделі в Bedrock. Жоден з трьох провайдерів не знав про злам, поки Dragos не повідомив.

Друга: продавці на маркетплейсі інференсу виставляли модель за найнижчою ціною, вмикали логування запитів і читали промпти. Одна людина з домашнього інтернету у Франції пробувала паролі з промптів для входу root по SSH.

Третя знахідка стосується приватності. Китайська лаба забирала секретні URL з більш ніж половини промптів до її моделей. Серед них були близько двох третин запитів з прапором Zero Data Retention (ZDR). Те саме відбувалось, коли модель цієї лаби продавав американський провайдер. Через 24 дні після запиту одне посилання з'явилось у робочому чаті Feishu, і його відкрили з шести пристроїв у Пекіні і Шанхаї.

OpenRouter пише, що злам самого OpenRouter не знайшли. Компанія відключила трьох постачальників і прибрала 125 кінцевих точок зі статусу ZDR. Публічні звіти вже опублікували Atlas Cloud, Novita, Phala і Together. Together пише про витік промптів за червень і липень.

Чому це важливо: прапор приватності діє лише на першому кроці маршруту, а запит часто проходить через кілька роутерів. Dragos радить три кроки: обирати провайдера явно, не брати маркетплейси з найдешевшою ціною і класти в контекст агентів власні пастки. Агент відправить у промпт усе, що бачить, включно з ключами.

ТЕМА 3

Better Data Center Project: 87% енергії Anthropic у 2028 році даватиме газ, 11% вугілля

ДЖЕРЕЛА Better Data Center Project, 16.11 Betterdatacenterproject · Blood in the Machine, 22.11 Bloodinthemachine · HN Hacker News

Звіт вийшов 16.11, але у вікні він став другою сторі HN доби з 914 балами і 915 коментарями. Better Data Center Project - некомерційна група, що вивчає енергію дата-центрів. Вона розібрала контракти однієї компанії, Anthropic. Станом на жовтень Anthropic законтрактувала 15 ГВт обчислень. З них 5,9-9,6 ГВт прив'язані до конкретних майданчиків.

Головна оцінка: у 2028 році близько 87% енергії Anthropic даватиме газ і 11% вугілля. Автори рахують, які електростанції покриють нове навантаження на мережі. Без контрактів на нові відновлювані джерела це газові і вугільні станції.

Публічні контракти Anthropic на відновлювану енергію є лише в Норвегії і Канаді. Водночас компанія витратила $3,4 млрд на газові генератори потужністю 1,7 ГВт.

Викиди CO2 у 2028 році група оцінює в 17-38 млн тонн, середній сценарій дає 32 млн тонн. Це більше за всю електрогенерацію Каліфорнії у 2024 році (31,1 млн тонн). 80% власної газової генерації на майданчиках становлять прості турбіни і двигуни, а вони брудніші за станції комбінованого циклу. Окремо звіт згадує Colossus 2 від SpaceX. Там Anthropic бере обчислення з травня, а більшість турбін працює без дозволів.

Автори самі називають межі оцінки. Частина викидів формально піде в звіти орендодавців, наприклад Amazon або SpaceX.

Викиди Google, Microsoft і Meta теж ростуть, тому порівняння не означає, що Anthropic стане найбільшим забруднювачем.

Чому це важливо: звіт показує механізм. Нове навантаження на мережу закриває найдешевша доступна генерація, а це газ і вугілля. Обіцянка «розширити енергетичний мікс» нічого не змінює без конкретних контрактів. Ця сама логіка стосується будь-якої компанії, що будує дата-центр біля газової мережі.

ТЕМА 4

Сюжет Nvidia: сенатор вимагає допиту Бессента і Лутніка, Бессент не виключає повернення Хуанга

ДЖЕРЕЛА The Guardian, 22.11 Guardian · Semafor, 22.11 Semafor · Financial Times, 22.11 FT · @GavinSBaker, 22.11 @GavinSBaker

Вчора дайджест писав про мережу Еліс Хуанг і $3,6 млрд грошей Китаю. У неділю сюжет перейшов на телевізійні шоу США.

Сенатор-республіканець Джон Кеннеді в Meet the Press назвав роман Хуанга «найдорожчим сексом з часів Олени Троянської».

Але сам Хуанг його цікавить менше. Кеннеді хоче допитати під присягою міністра торгівлі Говарда Лутніка і міністра фінансів Скотта Бессента. Питання: як Китай обійшов експортний контроль.

Guardian нагадує контекст. Megaspeed, де Еліс Хуанг була виконавчою директоркою, провела продажі чипів приблизно на $12 млрд в обхід експортного контролю. Лутніка критикують за те, що він ігнорував попередження про контрабанду. Бессент та інші хотіли внести Megaspeed у чорний список, але міністерство торгівлі затримало це рішення. Кеннеді каже, що керівники міністерства або некомпетентні, або «свідомо це дозволили».

Директор національної розвідки Джей Клейтон сказав, що розвідка перевіряє Еліс Хуанг. Чипи Nvidia він назвав «коштовністю, яку всі хочуть». Бессент у Face the Nation сказав, що уряд говорив з тимчасовим керівництвом Nvidia і юристами про те, хто відповідав за комплаєнс. FT вийшов з заголовком, що Бессент не виключає повернення Хуанга. Інвестор Гевін Бейкер процитував слова Бессента: «Я казав, що треба шукати людину ззовні. Дженсен тепер ззовні».

Чому це важливо: питання вийшло за межі ради директорів Nvidia і тепер стосується уряду США. Якщо Сенат викличе Лутніка, то предметом слухань стане робота експортного контролю. Повернення Хуанга тоді залежить не тільки від акціонерів.

ТЕМА 5

Replit: 32 агенти Opus прискорили NumPy у 18,9 раза, команди GPT-6.2 вийшли гіршими за одного агента

ДЖЕРЕЛА Replit, 19.11 Replit · @vsachar, 22.11 @vsachar

Дослідники Replit випустили Optimization-Bench. Агент отримує код популярної бібліотеки і дві години, щоб прискорити реальне навантаження без зміни результату. Перевірку робить окрема машина, до якої агенти не мають доступу. Вона також ловить підгонку під тести. Replit порівняла одного агента і команди з 8, 16 і 32 агентів для шести моделей.

Розбір вийшов 19.11, у листах X він розійшовся 22.11.

На NumPy один агент Claude Opus 5.5 дав прискорення в 3,66 раза, команда з 32 агентів дала 18,92 раза. Команди Opus і Sonnet 5.5 завжди перемагали своїх соло-агентів, медіана 1,30 раза на 16 і 32 агентах. GPT-6.2 Astra була найсильнішим соло-агентом: на NumPy більш ніж утричі швидша за соло-Opus. Але її команди ставали гіршими з ростом. На 32 агентах вони програли її ж соло-результату на трьох з чотирьох задач.

Різницю дала організація. Opus відправляв 82% повідомлень на спільну дошку, по треду на кожне навантаження, і оголошував «торкаюсь btree.c» перед правкою. Координатор Opus майже не писав працівникам. У GPT-6 Astra і GPT-6.1 Sol працівники писали координатору напряму, 20-22 повідомлення кожен, і координатор став вузьким місцем. На Git усі навантаження ділять ті самі 15 файлів, і там команди Opus не перегнали соло.

Чому це важливо: сильний соло-агент ще не означає сильну команду. Результат дають чітке володіння шматками роботи і спільний журнал, який може прочитати будь-хто. Ці самі правила працюють у людських командах.

ТЕМА 6

Збій Delta в Атланті і пояснення хвилі збоїв у США

ДЖЕРЕЛА The New York Times, 22.11 NYT · Financial Times, 22.11 FT · The Guardian, 22.11 Guardian

18.11 дайджест писав, що поспішні патчі після AI-знахідок валять системи в США. У неділю, на старті тижня Дня подяки, затримались рейси Delta через Атланту. Збій дала система, що рахує вагу і центрування літака. За даними Delta, з 1 654 затриманих рейсів близько 1 158 пов'язані саме з нею. Пополудні систему відновили. FAA наголосило, що її системи керування рухом тут ні до чого.

Guardian розібрав усю хвилю збоїв з 12.11. Висновок прозаїчніший за чутки про кібератаку: це переважно збіг. Наземну зупинку польотів по всій країні дав новий маршрутизатор. Його поставили в межах заміни мідних ліній FAA на оптику, а цей проєкт триває з 2002 року. Збій Microsoft дав баг з липневого оновлення. Але Агентство національної безпеки (АНБ) сказало Конгресу, що його збій спричинили швидкі виправлення вразливостей, які знайшов AI.

Чому це важливо: дві причини діють разом. Стара інфраструктура не отримує інвестицій. Водночас AI-пошук вразливостей змушує компанії оновлюватись удвічі швидше. Кожне поспішне оновлення старої системи несе ризик збою.

ТЕМА 7

GL.iNet: через діру в хмарі зловмисник поставив шкідливе ПЗ на роутери і KVM [одне джерело]

ДЖЕРЕЛА GL.iNet, 20.11 Gl-inet · HN Hacker News

GL.iNet робить популярні туристичні роутери і KVM-пристрої. 20.11 компанія знайшла вразливість у своєму сервісі віддаленого керування. Сторонній зловмисник використав її і поставив шкідливе ПЗ на частину роутерів і KVM. Компанія закрила діру в хмарі, тому нових заражень цим шляхом нема. Але виправлення в хмарі не прибирає шкідливе ПЗ з уже заражених пристроїв.

GL.iNet радить поставити свіжу безпекову прошивку для своєї моделі навіть без ознак зараження. Ознака зараження - процес `crond -i /tmp/.log`. Прошивка прибирає шкідливе ПЗ без скидання налаштувань. Скільки пристроїв постраждало, компанія не каже.

Чому це важливо: хмарне керування - це вхід на пристрій, який стоїть усередині мережі. KVM дає повний доступ до клавіатури і екрана підключеного комп'ютера. Якщо віддалене керування не потрібне, його краще вимкнути.

ТЕМА 8

Південна Корея: міністр освіти пішов у відставку через скандал з AI-списуванням [одне джерело]

ДЖЕРЕЛА The Guardian, 22.11 Guardian

Все почалось з іспиту на податкового консультанта. Найкращий учасник зізнався, що сфотографував питання прихованою мікрокамерою, відправив їх в AI-систему і читав відповіді зі смартфона. Це сталось за чотири дні до сунинга, головного вступного іспиту країни. Уряд відключив публічний wifi біля центрів іспиту, подвоїв кількість наглядачів і дозволив лише олівці. Ґудзики на одязі учасників заклеювали наліпками. Усім телефонам країни уряд надіслав тривожне сповіщення, щоб знайти приховані пристрої. За інформацію про списування уряд обіцяв щонайменше $75 000.

Іспит пройшов без великих інцидентів. Але міністр освіти Чхве Кьо Джін, 72 роки, пішов у відставку. Раніше на слуханнях він сказав, що через вік не впевнений, що зможе впіймати порушників.

Чому це важливо: іспити, що перевіряють знання без доступу до інформації, тепер вимагають майже військового режиму.

Мікрокамера і смартфон з AI дешевші за будь-який нагляд. Тому формат таких іспитів, найімовірніше, доведеться змінювати.

misc

- **Доказ Масулло в Lean:** спільнота веде публічну мапу формалізації рукопису Алессандро Масулло. Один з учасників на HN пише, що сотні знайдених помилок ще не означають діру в основній ідеї, і частину вже виправили. Розділ 13 рукопису, за коментарями, майже порожній. Hacker News
- **P vs NP наживо:** pequalnp.com показує флот AI-агентів, що атакує задачу P vs NP. На ранок 23.11 агенти перевірили 371 з 553 лем. Це експеримент, доказу нема. Pequalnp
- **Leonida:** репозиторій статичної рекомпіляції Xbox-версії GTA VI для PC. Автори перевели всі 26 503 шейдери у SPIR-V, але гра ще не запускається. Вчора @amasad писав про цей проєкт як про координацію рою агентів через GitHub issues. GitHub
- **Backblaze:** з 01.12 нова політика приватності дозволяє продавати або передавати персональні дані для таргетованої реклами. За FAQ компанії, файли резервних копій сюди не входять. Neowin
- **Автономна зброя:** Червоний Хрест попереджає, що скоро буде пізно регулювати автономну зброю. Конференція CCW в Женеві погодилась лише обговорювати інструмент регулювання наступні два роки. Guardian
- **США і Китай:** Трамп сказав, що США працюватимуть з Китаєм над «спільними запобіжниками» для AI. Semafor пов'язує зміну тону з протестами проти дата-центрів і результатами проміжних виборів.

  Semafor