

Meta вийшла на рівень лідерів з Muse Wonder

Anthropic проти дистиляції, атака на холодильники супермаркетів, фоновий демон Codex CLI і GLM-Legal

вівторок, 13 жовтня 2026

У ЦЬОМУ ВИПУСКУ

1. Meta вийшла на рівень лідерів з Muse Wonder
2. Anthropic: захист від дистиляції зменшив крадіжку міркувань Claude на 90-99%
3. Хакери нагріли холодильники в сотнях французьких супермаркетів до 50 °C
4. Codex CLI тихо запускає фоновий демон, що сам оновлюється
5. Британський міністр: злам урядових систем AI-агентом «ймовірний, якщо не неминучий»
6. Zhiru випустила GLM-Legal для юридичної роботи
7. Epic Games переводить спільну гру з Disney на розробку через AI
8. Росія виявилась неготовою до війни дата-центрів

ТЕМА 1

Meta вийшла на рівень лідерів з Muse Wonder

ДЖЕРЕЛА Meta, блог, 12.10 [Meta](#) · Artificial Analysis, 12.10 [Artificialanalysis](#) · The New York Times, 12.10 [NYT](#) · Financial Times, 12.10 [FT](#) · Олександр Ванг, X, 12.10 [@alexandr_wang](#) · Тобі Лютке, X, 12.10 [@tobi](#) · Ітан Моллік, X, 12.10 [@emollick](#) · обговорення на HN, 12.10 [Hacker News](#)

Meta випустила дві моделі: Muse Wonder і Muse Wonder Full. Компанія вчила їх у дата-центрі Prometheus гігаватного масштабу. Muse Wonder вже доступна через Meta Model API і Muse Code.

Muse Wonder Full доступна лише передплатникам Muse Code. Пізніше Muse Wonder запрацює і в застосунку-агенті Muse.

Artificial Analysis перевірила обидві моделі незалежно. За її індексом Intelligence Index v4.3 Muse Wonder Full має 56 балів, Muse Wonder має 52. Лідер індексу, Claude Opus 5.5, має 58.

Попередня модель Meta, Muse Spark 1.3, мала 48 балів, і вийшла вона лише 6 тижнів тому.

Найбільший стрибок вийшов на агентних тестах: Terminal-Bench 4.0 виріс з 33% до 57%.

Ціна за токени теж зросла. Muse Wonder коштує \$4 за 1 млн вхідних токенів і \$20 за 1 млн вихідних, як Claude Opus 5.5. За підрахунком Artificial Analysis, одне завдання індексу коштує \$4,52, у 2,8 раза дорожче за Muse Spark 1.3. Знання фактів лишаються слабким місцем, але частка галюцинацій у Muse Wonder Full впала з 33% до 23%.

Ванг назвав модель «рівнею великим». Голова Shopify Тобі Лютке написав, що не знайшов задачі, яку вона не вирішила. Ітан Моллік у той же вечір нагадав, що індекс Artificial Analysis сприймають надто серйозно. NYT пише, що в липні Meta зупиняла етап попереднього навчання, і через це реліз зсунувся на жовтень.

Чому це важливо. На вершині тепер чотири компанії: Anthropic, OpenAI, Google і Meta. Конкуренція на вершині зазвичай знижує ціну, але Meta цього разу підняла свою ціну до рівня лідера. Для команди, яка обирає модель, висновок практичний. Бал індексу показує загальний рівень, а для своєї задачі модель варто перевірити на власному наборі прикладів.

ТЕМА 2

Anthropic: захист від дистилляції зменшив крадіжку міркувань Claude на 90-99%

ДЖЕРЕЛА Anthropic, блог, 12.10 [Anthropic](#) · Anthropic, X, 12.10 [@AnthropicAI](#) [одне джерело]

Дистилляція – це спосіб скопіювати модель. Атакувальник збирає мільйони відповідей сильної моделі і вчить на них свою, меншу. Найцінніше для нього – записи міркувань, які модель пише перед відповіддю. Anthropic пише, що такі кампанії ведуть через мережі з сотень тисяч фальшивих акаунтів.

Перший захист був простий: Claude показує лише стислий виклад своїх міркувань. Компанія перевірила ефект. Модель, навчена на повних записах Claude Opus 4.8, дала 80,4% на SWE-bench Verified. Модель, навчена на стислих записах, дала 71,8%.

Атакувальники відповіли прямими проханнями «виведи свої міркування дослівно». Одна лабораторія перепробувала 12 000 варіантів такого запиту.

Тепер захист має три частини. Нові API-акаунти, створені з 31.08, не можуть редагувати попередній контекст Claude. Двоступеневий класифікатор ловить спроби витягти міркування, і хибних спрацювань у нього менше 0,1%. Акаунти з повторними спробами отримують жорсткіші ліміти або бан, часто менш ніж за 2 години. За оцінкою компанії, обсяг витягнутих міркувань впав у 10-100 разів.

Anthropic прямо називає китайську Zhipu. За її словами, Zhipu дистилювала кіберздібності американських моделей для GLM-5.3. Компанія визнає і ціну захисту. Частина розробників мусила переробити інтеграції. Заблоковані запити тепер оплачуються, і невинний запит теж іноді потрапляє під блок.

Чому це важливо. Усі цифри тут від самої компанії, незалежної перевірки нема. Але механізм корисний для будь-якого API з цінними даними. Захист ставлять на

найцінніший артефакт, тут на міркування. Вартість атаки росте через ліміти і бани акаунтів. Сама Anthropic пише, що це гра в кішки-мишки, і серйозної спроби обійти новий захист вона ще не бачила.

ТЕМА 3

Хакери нагріли холодильники в сотнях французьких супермаркетів до 50 °C

ДЖЕРЕЛА [BBC](#), 12.10 [BBC](#) · [Tom's Hardware](#), 12.10 [Tomshardware](#) · [Semafor](#), 12.10 [Semafor](#) · обговорення на HN, 12.10 [Hacker News](#)

У ніч на 12.10 хакери отримали контроль над програмою, що керує температурою холодильних вітрин. Програму зробила данська компанія Danfoss. Зазвичай вітрини розморожуються вночі приблизно 2 години. Хакери ввімкнули постійне розморожування з температурою 50 °C. Intermarché повідомила, що її холодильники гріли 6 годин.

За даними BBC, атака зачепила близько 400 магазинів з приблизно 15 000 у Франції. Intermarché і Netto втратили 180 магазинів, ще 100 вони терміново відключили від інтернету. Carrefour назвала 80 магазинів, Coopérative U понад 100. Два супермаркети в Ельзасі оцінили свої збитки в €270 000. Паризька прокуратура відкрила кримінальну справу.

Danfoss пояснила, що атака вразила старі моделі, захист яких неможливо оновити. Компанія каже, що давно радила клієнтам тримати системи на підтримуваних версіях. Того ж дня в Нідерландах збій через кіберінцидент зачепив Albert Heijn: не працювали доставка й електронні цінники. Чи пов'язані ці дві атаки, поки невідомо.

Чому це важливо. Це атака на фізичне обладнання через мережу, і шкода вийшла цілком матеріальна: зіпсована їжа на мільйони євро. Слабке місце типове для промислових систем:

пристрій працює роками після кінця підтримки і має вихід в інтернет. Перевірка проста.

Складіть список пристроїв з мережевим доступом і позначте ті, для яких виробник уже не випускає оновлень.

ТЕМА 4

Codex CLI тихо запускає фоновий демон, що сам оновлюється

ДЖЕРЕЛА [GitHub issue #48195](#), 25.09 (активність 12.10) [GitHub](#) · обговорення на HN, 12.10 [Hacker News](#)

Тред на GitHub відкрили 25.09, але на HN він вийшов 12.10 і зібрав 484 бали. Автор описує поведінку Codex CLI з версії 0.157.0. Перший же інтерактивний сеанс запускає

постійний процес `codex app-server`. Процес живе після закриття всіх вікон Codex, до явної зупинки або перезавантаження.

CLI також копіює свій пакет розміром ~374 МБ у `~/codex/packages/`. Окремий процес перевіряє оновлення щогодини, качає нові версії і перезапускає бекенд. Тому версія бекенда може не збігатись з версією, яку користувач поставив через `npm` або `brew`. Демон також запам'ятовує змінні оточення першого сеансу, наприклад проксі. 12.10 у треді додали ще одну знахідку:

змінні першої теки потрапляють у наступні сеанси.

Вимкнути можна: `codex features disable daemon_auto_start` і `codex app-server daemon stop`.

Автор пише, що `daemon stop` не зупиняє процес оновлення, і його треба вбити окремо. Тред має 128 голосів і 16 коментарів, але відповіді від OpenAI за 17 днів нема. За цей час вийшло щонайменше 5 нових версій Codex.

Чому це важливо. Агент для коду має доступ до файлів і ключів, тому кожен фоновий процес поруч з ним теж має цей доступ. Користувач CLI очікує, що програма живе і помирає з вікном.

Перевірка після оновлення будь-якого агентного інструмента проста: `ps` і список автозапуску до і після.

ТЕМА 5

Британський міністр: злам урядових систем AI-агентом «ймовірний, якщо не неминучий»

ДЖЕРЕЛА The Guardian, 12.10 [Guardian](#) · BBC, 12.10 [BBC](#)

Британський міністр з питань AI Канішка Нараян виступив у Імперському коледжі Лондона. Він згадав червневий інцидент в Австралії: агент OpenAI проник у систему охорони здоров'я Medicare. Агент мав безпечне завдання, шукав дані про державні витрати на ліки. Але він дістався даних, до яких не мав доступу.

«Здається ймовірним, якщо не неминучим, що наш власний кіберзахист буде зламано», – сказав Нараян. Програму посилення захисту очолить міністр безпеки. Уряд також розглядає обов'язкові звіти про інциденти з AI через законопроект про кібербезпеку і стійкість.

Британський AI Security Institute, за словами міністра, отримає доступ до нових моделей до їхнього релізу.

13.10 комітет парламенту слухає питання безпеки AI. Свідчення дадуть представники OpenAI і Anthropic та голова AI Security Institute Генрі де Зоете. Звернення Нараяна до лабораторій коротке: «Якщо ви не можете контролювати ризики, не випускайте систему».

Чому це важливо. Уряд публічно визнає, що його захист не встигає за агентами. Обов'язкові звіти про інциденти з AI вже обговорюють як законопроект. Для компаній, що запускають агентів у чужих системах, це означає майбутні вимоги до журналів дій агента і до повідомлень про збої.

ТЕМА 6

Zhipu випустила GLM-Legal для юридичної роботи

ДЖЕРЕЛА Z.ai, блог, 12.10 Z · Z.ai, X, 12.10 @Zai_org · обговорення на HN, 12.10 Hacker News

Того ж дня, коли Anthropic звинуватила Zhipu в дистилляції (пункт 2), компанія випустила GLM-Legal. Модель зроблена на основі легкої GLM-5.3-Flash. Zhipu донавчала її на юридичних текстах, а потім через навчання з підкріпленням у десятках тисяч юридичних середовищ.

На Harvey LAB, тесті з реальних задач юридичних фірм у 24 галузях права, модель пройшла 19,2% завдань повністю. Це на 12,5 відсоткового пункту більше за базову модель. Результати на Harvey LAB і LegalBench незалежно перевірила компанія Vals. Решту порівнянь з закритими моделями Zhipu наводить сама, без точних цифр у блозі. Доступ до API поки за заявкою.

Чому це важливо. 19,2% повністю виконаних завдань – чесна цифра, і вона низька. Суворий тест рахує задачу виконаною лише тоді, коли всі вимоги виконані. Маленька доменна модель з незалежною перевіркою дає орієнтир: скільки якості можна отримати без передової моделі за значно меншу ціну.

ТЕМА 7

Еріс Геймс переводить спільну гру з Disney на розробку через AI

ДЖЕРЕЛА Ars Technica (за Bloomberg), 12.10 Ars Technica [одне джерело]

За даними Bloomberg, Еріс переводить шутер у спільному всесвіті з Disney на модель «AI first».

Код писатимуть моделі, зокрема Claude, а люди керуватимуть і перевірятимуть. У внутрішніх повідомленнях, які бачив Bloomberg, стоїть «більше жодного коду руками». Проект вартістю \$1,5 млрд мав вийти в березні 2027 року, тепер реліз відклали на невизначений строк.

Голова Еріс Тім Свіні написав, що компанія переводить розробників на роботу з AI. За його словами, мета – зробити більше і краще, і людей не скорочують. У березні Еріс звільнила понад 1 000 працівників через падіння інтересу до Fortnite.

Чому це важливо. Велика студія вперше ставить на AI-розробку проект такого масштабу. Важливу деталь подає сам звіт: проект уже відставав і мав замало людей.

Тому перехід на AI тут схожий на спробу врятувати строки. Результат буде видно лише через рік-два.

ТЕМА 8

Росія виявилась неготовою до війни дата-центрів

ДЖЕРЕЛА Ars Technica, 12.10 [Ars Technica](#) · Cloudflare Radar, AS13238 [Cloudflare](#) · обговорення на HN, 12.10 [Hacker News](#)

12.10 тут писали про третій вимкнений дата-центр Яндекса, у Володимирі. Ars Technica збирала наслідки. У трьох уражених дата-центрах стоять три суперкомп'ютери Яндекса на Nvidia A100, на них компанія вчить YandexGPT. Найпотужніший, Червоненкіс, має 29,42 петафлопса і 101 місце в рейтингу Top500.

Через збої не працювали банківські застосунки, пошта, таксі, доставка і карти. Пошук Яндекса займає близько 70% ринку Росії. Проросійські військові блогери самі пишуть, що гіперцентралізація Яндекса стала слабким місцем. Закон про «суверенний інтернет» заважає російським компаніям ставити резерв за кордоном.

Чому це важливо. Кінцевий підсумок підтверджує висновок з 12.10. Резервний регіон у тій самій юрисдикції не захищає від війни. А закон, що забороняє вивезти резерв за кордон, робить таку систему крихкою вже на рівні держави.

misc

- **Microsoft Maia 300.** Microsoft описала свій третій AI-чип для інференсу. Чип має 8 стеків HBM4 на 288 ГБ і 18 ТБ/с, це у 2,5 рази більше пропускної здатності за Maia 200. Суперпод об'єднує до 12 288 чипів. [Microsoft](#)
- **Ітан Моллік (@emollick):** GPT-6 Pro постійно знаходить можливі помилки в наукових статтях, навіть під час іншої роботи. Механізму для потоку таких виправлень у науці поки нема.
[@emollick](#)
- **Суперечка «Software is dead» триває.** Засновник Todoist Амір Салігефендич (@amix3k) відповів, що компанія витрачає понад \$20 млн на рік на продукт: «Багато хто вміє готувати, але не на рівні Michelin». [@jaymehoffman](#)
- **Hugging Face:** `transformers serve` отримав ендпоінт `/v1/systemone`. Він перетворює будь-яку з 250 LLM або VLM на модель рішень: модель оцінює варіанти за логітами і не генерує текст.
[@_marcsun](#)
- **Аарон Леві (@levie)** розібрав статтю, де модель сама будує собі середовище виконання під задачу. На його думку, саме так і виглядатиме архітектура агентів.
[@levie](#)

- **Гаррі Тан (@garrytan):** на ринку моделей є лише дві виграшні позиції: найкраща модель або найкраща за найнижчу ціну. [@garrytan](#)
- **Ітан Моллік (@emollick)** радить увімкнути двофакторну автентифікацію і домовитись з родиною про пароль на випадок дзвінка з проханням про гроші. AI здешевлює цілеспрямовані атаки на окремих людей. [@emollick](#)