AI-ДАЙДЖЕСТ

# Qwen3.8-Max вийшов з відкритими вагами

*Перша Max-модель з відкритими вагами, $2/$6 за мільйон токенів.*

понеділок, 3 серпня 2026

---

У ЦЬОМУ ВИПУСКУ



**1. Qwen3.8-Max: перша Max-модель з відкритими вагами, $2/$6 - і пелікан Саймона Вілісона за 17 центів** 181 бал HN. Вийшов сьогодні вночі за Києвом, найсвіжіше в випуску.

Дослівно з блогу (цитата по треду HN, бо сама сторінка рендериться клієнтом): «Сьогодні ми офіційно випускаємо Qwen 3.8-Max, найздібнішу модель у сімействі Qwen на сьогодні. **Це також перший раз, коли ми відкриємо ваги моделі класу Qwen-Max** - відкриті ваги вийдуть наступного тижня.»

Звірені цифри:

• **Ціни: $2.0 за млн вхідних / $6.0 за млн вихідних, імпліцитне кешування $0.25** - з офіційного поста @Alibaba_Qwen • Підтримка `reasoning_effort` з чотирма рівнями, дефолт - **xhigh** • Окремо анонсовано **Qwen3.8-27B з відкритими вагами наступного тижня**

**Найцінніше в треді - замір від @simonw.** Він одразу прогнав свій канонічний тест (пелікан на велосипеді): «Це зайняло **11 хвилин** і воно забуло колеса!» І порахував ціну:

**17 центів** за 29 734 вихідних токени. Він же спіймав плутанину: два тижні тому виходив `qwen3.8-max-preview`, сьогодні повна модель, і з формулювань блогу це неочевидно.

**Чому це важливо.** Дві речі. Перша - **$2/$6 проти того, що вимірювалось тиждень тому**: це приблизно рівень, на якому «дешева модель на просту операцію» перестає бути теорією. Теза Малєєва про токенмаксинг («інженери беруть найдорожчу модель на найпростішу операцію, бо поміняти модель важче, ніж витратити $30») отримує альтернативу, яку тепер незручно ігнорувати. Друга - **11 хвилин і забуті колеса**: низька ціна за токен і низька ціна за результат це різні речі. На підписці перша колонка нуль, а 11 хвилин рахуються завжди. [proven - щодо цін і факту релізу; **fuzzy** - щодо якості, бо це один тест однієї людини]

блог Qwen · ціни @Alibaba_Qwen · @simonw про китайські моделі · підрахунок 17 центів · HN-тред

**2. «Kimi K3 на MI355X вигідніший за B300» - а в таблиці всередині B300 виграє в кожному рядку** 204 бали HN. Найкорисніший пункт випуску, хоча тема - залізо поза щоденним фокусом.

Заголовок статті Wafer: «Running Kimi K3 on MI355X at Better Performance per Dollar Than B300». Звучить як перемога AMD. **Сама стаття дала таку таблицю:**

| | 8× MI355X | B300 |
|---|---|---|
| Decode tok/s на потік | 118 | **172** |
| Пікова агрегована | 952 | **1 568** |
| Пікова на GPU | 119 | **196** |
| **На долар/GPU-год** | **48** | 33 |

**B300 швидший у трьох рядках з чотирьох.** Єдиний рядок, де виграє MI355X, останній, і то лише тому, що поділено на ціну **$2.50 проти $6.00 за GPU-годину**, яку Wafer призначив сам. Коментатор на HN сформулював рівно це: «На кожному рядку B300 обійшов MI355X». І другий, жорсткіше: «Wafer робить себе синонімом слопу в inference-просторі. Перебільшені несправедливі порівняння в усіх їхніх результатах... **$2.50/GPU-год для MI355X, $6.00 для B300 - це неточне порівняння в реальних умовах.**»

Технічна частина при цьому не порожня: Kimi K3 - **2.8 трлн параметрів**, «понад 1.5 ТБ VRAM до виділення KV-кешу на 1M токенів контексту», і **навіть нода B200 з 8 GPU його не вміщає**. Це справжня причина, чому MI355X з 288 ГБ на GPU тут узагалі в розмові.

**Чому це важливо.** Заголовок технічно не брехня: «performance per dollar» справді на боці AMD. Але **вибір знаменника зробив автор**, і саме цей вибір несе всю вагу висновку. Правдоподібне пояснення є завжди, і на ньому легко зупинитись, поки не подивитись, з чого зроблене число. Правило з цього: коли трапляється метрика-дріб (щось «на долар», «на ват», «на токен»), дивитись треба обидва доданки, з яких він складений. [proven - таблиця і ціни з самої статті Wafer, коментарі з HN-треду]

стаття Wafer · HN-тред з розбором

**3. MIT Sloan: 1000 людей, три моделі - ШІ-фінпорадник дає непогані поради, але відповідь залежить від того, ХТО питає** 338 балів HN, **379 коментарів** - найгарячіша дискусія доби. **Стаття від 21.07**, вона просто випливла вчора на HN.

Як робили: побудували модель життєвого циклу (доходи, робота, інвестиції, податки) як еталон хороших рішень. Дали **1000 дорослих** написати **власні** промпти до **GPT-5.2, GPT-5.6 і Gemini 3 Flash**. Просимулювали, що буде, якщо людина 22 - 89 років житиме за цими порадами. Потім повторили з «академічними» промптами, з повними вхідними даними і явними припущеннями.

**Що вийшло добре:** поради виявились кращими, ніж очікували самі дослідники: більше заощаджень, участь у ринку акцій, диверсифікація, зниження частки акцій після 45. Дослівно Taha Choukhmane: «Нас дещо здивувало, наскільки хорошими були поради».

**Що вийшло погано:** • ШІ **погано адаптується до шоків** - людині після втрати роботи радив різати витрати надто різко, **навіть коли в неї були заощадження** • дозволяв портфелю **дрейфувати** замість активного ребалансування • **поради залежать від того, хто пише промпт**:

> *Промпти, написані **чоловіками**, більш фінансово грамотними або тими, хто вже мав досвід з ШІ, давали **~5% більше капіталу** до пенсії. Різниця в рекомендованій частці акцій компаундилась у **~$50 000 (4%) менше капіталу у віці 60** для жінок і менш фінансово грамотних. А тим, хто раніше не користувався ШІ для фінпорад, модель радила **нижчу норму заощаджень** - **майже $100 000 (6%) менше** до 60 років.*

Механізм автор пояснює прямо: люди просто пишуть різні промпти. Жінки частіше вживали слова «сім'я», «продукти», «платити»; чоловіки - «стратегія», «крипта», «зростання».

**Чому це важливо.** Головний висновок: **якість відповіді дорівнює якості вхідних даних**. Модель без явних цифр добудує середню людину і відповість середній людині. Практично це означає, що в питання про портфель мають входити реальні цифри і озвучені припущення: горизонт, дохід, подушка. І окремий рядок: **дрейф замість ребалансування** легко не помітити, бо це не виглядає як рішення. [proven - рецензована робота з описаною методологією; дата 21.07, симуляція, не реальні портфелі]

MIT Sloan · HN-тред

**4. Леві: чому найважча робота автоматизується першою** 76.6K переглядів. Вчора була його теза про розрив між побутовим і глибоким ШІ, сьогодні вона добудована до механізму.

Дослівно: «Входимо в дивну динаміку, коли деяка з "найважчої" роботи у світі насправді **першою піддається автоматизації - саме через свою верифіковність**. Математика, кібербезпека і код - попри те, що це шалено складні й високоцінні галузі - мають

перевагу: **їх можна об'єктивно перевірити на правильність**. Це дає дві негайні вигоди: тренування моделей отримує **чіткіший сигнал винагороди**, а запуск моделей дозволяє знати, що воно працює правильно, бо результат можна тестувати **масштабовано**.»

І зворотний бік: «В інших доменах **миттєвої верифіковності значно менше**. Які юридичні пункти клієнт погодить, яку маркетингову кампанію запускати за мінливого настрою, яке повідомлення захоче почути потенційний покупець, які фінансові цілі й бюджет ставити... **немає "однієї правильної відповіді"**. Вони залежать від думок і рівня ризику операторів, дуже чутливі до вхідного контексту, і **в багатьох випадках правильну відповідь неможливо дізнатись ще довго** після того, як модель видала результат.»

Висновок Леві: навіть при експоненційному зростанні моделей «**на прикладному шарі буде зроблено значно більше, ніж самою моделлю**», і самі процеси доведеться змінювати.

**Чому це важливо.** Це пояснює вчорашню математику і сьогоднішній пункт 3 одним механізмом: Astra валить теорію чисел (Lean компілюється або ні), а фінпорадник дрейфує (правильність видно через 30 років). Там, де перевірку можна зробити дешевою і швидкою, автоматизація зайде глибоко; де не можна, ШІ лишається прискорювачем. Питання «як здешевити верифікацію» важливіше за питання «яку модель узяти». [promising - аргументована рамка практика, збігається з незалежним постом @max_spero_ про рівні верифіковності]

Леві · @max_spero_ (цитований пост про рівні верифіковності - точного лінка збір не дав, тому даний профіль)

**5. Моллік: «нехай лабораторії відпрацьовують» - список нерозв'язаних питань поза математикою** 20.9K переглядів, свіжий (03:23 за Києвом). Пряме продовження вчорашнього пункту 1.

Його теза: «Математика отримує багато уваги за свої нерозв'язані проблеми, але **є нерозв'язані й важливі проблеми в багатьох галузях**, які потенційно можна вирішити емпірично, якщо ШІ справді стане достатньо добрим. Проблеми, які, якщо їх розв'язати, **принесли б велику цінність суспільству**.»

І далі він **виписує свій список** по підприємництву. Що саме **спричиняє** підприємницький успіх, поза кореляціями; чи винятковий ріст **передбачуваний** взагалі, чи це емерджентний шлях, помітний лише постфактум; які навички можна **навчити** так, щоб вони реально підвищили успіх; яка **найменша можлива інтервенція**, що переводить місце з низько-підприємницької рівноваги в стійку екосистему; чому одні фірми з ростом стають менш адаптивними, а інші лишаються гнучкими.

Окремим постом він додав: «Думаю, галузям було б корисно перелічити деякі з таких проблем. **Хай ШІ-лабораторії це відпрацьовують.**»

**Чому це важливо.** Найдешевша ідея тижня, і вона працює без жодної нової технології: скласти власний список нерозв'язаних питань у своїй галузі. Формула проста: «що досі

невідомо і що змінилось би, якби було відомо». Список справ і список питань це різні документи, і другий майже ніхто не веде. [promising - це фреймінг, не результат]

Моллік - список · «нехай лабораторії відпрацьовують»

**6. Qwen3.8-Max на шейдер-тесті Молліка: «solid, але не рівень Kimi K3»** Свіжак, 42 хвилини до збору. Незалежний практик за годину після релізу, тому це окремий пункт.

Моллік прогнав свій власний тест, складний шейдер: «нескінченне місто неоготичних веж, наполовину затоплене штормовим океаном». Оцінка: «Базове враження після купи експериментів - це **солідна модель, але не рівня Kimi K3**, за наявним досвідом наразі.»

**Чому це важливо.** Ілюстрація до пункту 1: офіційний блог каже «новий рівень для кодингу», а незалежний практик з власним тестом за годину каже «солідно, але не топ». **Жодне з двох тверджень не бреше**, вони міряють різне. Тому релізи лабораторій не варто переказувати без другого голосу. [proven - власний замір Молліка, але n=1 і суб'єктивна оцінка]

Моллік про Qwen3.8-Max (той самий шейдер-тест на Kimi K3 робився 16.07, цитується всередині цього ж поста)

**7. QM: 5 234 до 7 616 зірок за добу. Третій день, темп не падає** Продовження вчорашнього пункту 4: цифра знову виросла, і це вже тренд.

GitHub API просто зараз: **7 616 зірок, 800 форків**. Хронологія за три дні: **01.08 - 2 475, 02.08 - 5 234, 03.08 - 7 616**. Тобто **+2 382 за добу** після **+2 759** позавчора. Приріст трохи сповільнився, але тримається на високому рівні; репо створено 29.07.

**Чому це важливо.** Ідея, яку вони транслюють у себе: раз на тиждень витягати з журналу агентних сесій хибні ходи і зводити їх в окремий список «як не треба». Сьогоднішній пункт 2 (метрика-дріб) - рівно такий кандидат. [proven - цифри з GitHub API, звірено]

репо QM

**8. MiniMax H3: відкрита відео-модель з синхронним звуком, 33B, під споживчі GPU** З листа «AI + Product», репост Hugging Face - 405 лайків на картці моделі.

Дослівно з поста @multimodalart: «MiniMax H3 щойно вийшла на Hugging Face - **text-to-video, image-to-video, reference-to-video - усе зі звуком**. Потужні **33B параметрів** - але готова для **споживчих GPU** в diffusers і comfy.»

Звірено по HF API: картка `MiniMaxAI/MiniMax-H3` створена **28.07**, **405 лайків**, теги підтверджують заявлене - там і `text-to-audio-video`, і `audio-video-generation`, і `synchronized-audio-video`.

**Чому це важливо.** Клас «відкрита модель, що влазить у споживчу карту» вже пройдений з картинками, тепер те саме приходить у відео зі звуком. Для коротких роликів без хмари це та лінійка, за якою варто стежити. [proven - щодо параметрів моделі й наявності ваг; **fuzzy** - щодо якості, вона не запускалась]

@multimodalart · картка на HF

**9. Колісон про естетику: чому все стало потворним** 417K переглядів, 3.4K лайків, 430 коментарів - найпопулярніший пост доби в листі «Tech + Product», і він взагалі не про ШІ.

Патрік Колісон (Stripe) відповідає на питання, чому він, усе життя провівши в STEM, почав думати про естетику. Стисло:

• «**Багато речей сьогодні потворні** - і значно потворніші, ніж були раніше або мусять бути. Щойно це побачено, перестати помічати вже важко.» Приклади: телефонні будки початку XX століття проти сучасних, старі питні фонтанчики проти нових. І питання: «**чому перестали робити красиво? Це вибір? Чи на нас наклали злі чари?**»

• Модернізм як **розрив культурної безперервності**: виставка International Style 1932 року витіснила багатий гобелен попередніх стилів, і цей розрив, за його підозрою, мав наслідки **за межами естетики** • І найсильніше, з посиланням на Елейн Скеррі: «**краса надихає творення. Якщо так, то може бути правдою й зворотне: потворність його гальмує.**»

• Історичний аргумент: Петрарка створив передумови Ренесансу, який виростив наукову революцію. Всесвітня виставка 1851 у Кришталевому палаці зібрала **6 млн відвідувачів при населенні 21 млн**: естетика і матеріальний розвиток колись були взаємозалежні популярно.

**Чому це важливо.** Це не новина, і взята вона свідомо. Для будь-якої команди, що робить продукт, теза «потворність гальмує творення» це робоча гіпотеза про якість роботи. [fuzzy - це есей-роздум, не дослідження; береться як рамка, не як факт]

Колісон

**10. Naval: «Agent Programming Interface» - три слова, 85K переглядів** Найлаконічніший пост доби. Навал написав рівно три слова: **«Agent Programming Interface»** - 85.2K переглядів, 1.1K лайків за годину.

Розшифровки він не дав. Але напрошується очевидне: **API писались для програм, а користувачами стають агенти**. Те, що зручно програмі (стабільна схема, суворі типи), не збігається з тим, що потрібно агенту: опис намірів, приклади, зворотний зв'язок про помилку природною мовою.

**Чому це важливо.** Ця ідея вже працює там, де набір CLI-команд пишеться під модель і документується в промпті або в описі інструментів. Свіжий контрприклад з сьогоднішнього ранку: **сторінка Qwen має два теги `<body>`, і на цьому браузерний інструмент упав шість разів поспіль**, рівно тому, що інтерфейс проєктувався під людину з очима, а селектори читає агент. Оце і є та проблема, про яку Навал написав три слова. [fuzzy - це інтерпретація трьох слів, він нічого не пояснював]

Naval

**Misc**

• **@karpathy доклав джерела до вчорашнього LOTR-експерименту** (90K переглядів): виклав вихідний код, щоб було грабельно і форкабельно - karpathy.ai/lotr-movie (перевірено, віддає 200). Плюс послався на пелікана @simonw як на попередній канон тесту. І жарт, який пішов у народ: «Чекайте на GTA Hobbiton раніше за GTA VI» пост

• **@mckaywrigley під тим самим постом**: «хтось, будь ласка, профінансуйте всю трилогію в такому вигляді як бенчмарк» - 196 лайків пост

• **@emollick про chain-of-thought**: «На випадок, якщо цікаво - **протестовано "think step by step", і воно більше не працює** для покращення результатів у сучасних моделях». Технічний звіт Wharton: «Спадна цінність ланцюжка міркувань у промптингу». Практично: рядок «думай покроково» в промпті - це вже карго-культ пост · він же з гумором: «Пам'ятається час, коли chain-of-thought робився вручну, по-старому. Влітку 2024-го»

• **@amasad вивів свій шаховий рушій на LiChess**: грає з людьми і ботами автономно, зараз **1253 Elo** на живих партіях (проти ~1500 у тестах проти Stockfish level 0), «грає три партії одночасно» пост

• **@gdb тричі за добу про ChatGPT/Codex як виконавця бізнес-задач**: агент, що сам зняв і опублікував рекламу, «зупинився на кнопці Оплатити і спитав дозволу» (43K) 1; Codex-навичка, що перетворює фідбек клієнтів на роадмап (59K) 2; «chatgpt work - це новий планувальник задач» (101K) 3. Це маркетинг OpenAI через ретвіти користувачів, береться як сигнал напряму • **@garrytan про меритократію**: «Усі плутають карту з територією. Меритократія - це коли **територія важливіша за карту**. А на ринках територія - це результат: чи зроблено те, чого люди хочуть?» пост

• **@lennysan - новий подкаст**: «CPO, який шкодує, що продакт-менеджмент існує» з Томом Верріллі (Whatnot, ex-Twitch, ex-Twitter). Теми: «як ШІ **оголює масштабний продуктовий театр**» і навичка PM, що росте найшвидше - системне мислення (це вже **шосте** інтерв'ю поспіль з тією ж тезою) пост

• **@eladgil про LLM-as-a-judge**: «Ground truth для агента - **те, як агент має поводитись**». Точний діагноз, чому оцінювати агентів по фінальній відповіді ламається пост

• **@yishan про життя в сингулярності**: «Якщо хочеться зробити щось з ШІ і не виходить, робоче рішення буквально таке: **пограти у відеоігри два тижні і почекати, поки технологія підтягнеться**» - 77K переглядів пост

• **Go 1.27 Interactive Tour** - 345 балів, другий за балами на HN. У коментарях головна тема - розширені дженерики: «Хтось один у шоці, що Go нарешті приймає дженерики?» HN · тур

• **Linux на десктопі перевалив за 10% у Північній Америці** - дві окремі історії того самого дня, 176 і 98 балів HN

• **@shreyas про смак у продукті**: «Інсайт, інтуїція, смак укорінені в незліченних факторах, тому неможливо просто стати чудовим у продукті через поради й тактики.

Але багато людей ставали значно кращими після того, як **позбувались довічної потреби почуватись розумними і звучати інтелектуально**» пост

• **15-річний зібрав циклоїдний редуктор** і виклав на GitHub - 320 балів, один з найтепліших тредів доби HN

---