AI DIGEST

Qwen3.8-Max: the first Max model with open weights, $2/$6 - and Simon Willison's pelican for 17 cents

Yesterday was the day of a loud announcement (ten problems from Astra). Today is quieter on headlines and denser on working material. The main item: Qwen3.8-Max shipped, and for the first time in the Max line it promises open weights, at $2/$6 per million tokens. Alongside it came a practical lesson: an article with a loud headline where HN comments showed within an hour that the headline says the opposite of the table inside it.

Monday, 3 August 2026

IN THIS ISSUE

1. Qwen3.8-Max: the first Max model with open weights, $2/$6 - and Simon Willison's pelican for 17 cents 181 points on HN. Released overnight Kyiv time, the freshest thing in this issue.

Verbatim from the blog (quoted via the HN thread, since the page itself renders client-side): "Today we are officially releasing Qwen 3.8-Max, the most capable model in the Qwen family to date. **This is also the first time we will open the weights of a Qwen-Max class model** - the open weights land next week."

Verified numbers:

• **Pricing: $2.0 per million input / $6.0 per million output, implicit caching $0.25** - from the official @Alibaba_Qwen post • `reasoning_effort` support with four levels, default **xhigh**

• Separately announced: **Qwen3.8-27B with open weights next week**

The most valuable part of the thread is @simonw's measurement. He immediately ran his canonical test (a pelican on a bicycle): "It took **11 minutes** and it forgot the wheels!" And he priced it: **17 cents** for 29,734 output tokens. He also caught a confusion: two weeks ago `qwen3.8-max-preview` shipped, today it is the full model, and the blog's wording does not make that obvious.

Why it matters. Two things. First, **$2/$6 against what was measured a week ago**: this is roughly the level at which "a cheap model for a simple operation" stops being theory. Maleev's token-maxing thesis ("engineers reach for the most expensive model for the simplest operation, because swapping the model is harder than spending $30") now has an alternative that is awkward to ignore. Second, **11 minutes and forgotten wheels**: a low price per token and a low price per result are different things. On a subscription the first column is zero, but 11 minutes always counts. [proven - on pricing and the release itself; **fuzzy** - on quality, since this is one test by one person]

Qwen blog · pricing @Alibaba_Qwen · @simonw on Chinese models · the 17-cent calculation · HN thread

2. "Kimi K3 on MI355X beats B300 per dollar" - and in the table inside, B300 wins every row 204 points on HN. The most useful item in the issue, even though the topic is hardware outside the daily focus.

Wafer's headline: "Running Kimi K3 on MI355X at Better Performance per Dollar Than B300". It sounds like an AMD win. **The article itself gave this table:**

	8× MI355X	B300
Decode tok/s per stream	118	**172**
Peak aggregate	952	**1,568**
Peak per GPU	119	**196**
Per dollar/GPU-hour	**48**	33

B300 is faster in three rows out of four. The only row MI355X wins is the last one, and only because it is divided by a price of **$2.50 against $6.00 per GPU-hour**, which Wafer assigned itself. A commenter on HN put it exactly: "In every single row B300 beat MI355X". And a second one, harsher: "Wafer is making itself synonymous with slop in the inference space. Exaggerated unfair comparisons across all of their results... **$2.50/GPU-hr for MI355X, $6.00 for B300 is an inaccurate real-world comparison.**"

The technical part is not empty: Kimi K3 has **2.8 trillion parameters**, "over 1.5 TB of VRAM before allocating KV cache for 1M tokens of context", and **even an 8-GPU B200 node cannot hold it**. That is the real reason MI355X with 288 GB per GPU is in this conversation at all.

Why it matters. The headline is not technically a lie: "performance per dollar" really is on AMD's side. But **the author chose the denominator**, and that choice carries the whole

weight of the conclusion. A plausible explanation is always available, and it is easy to stop there before looking at what the number is made of. The rule that follows: when a ratio metric shows up (something "per dollar", "per watt", "per token"), look at both terms it is built from. [proven - the table and prices from Wafer's own article, comments from the HN thread]

Wafer article · HN thread with the breakdown

3. MIT Sloan: 1000 people, three models - AI financial advice is decent, but the answer depends on WHO is asking 338 points on HN, **379 comments**, the hottest discussion of the day. **The article is from 21.07**, it simply surfaced on HN yesterday.

Method: they built a life-cycle model (income, work, investments, taxes) as a benchmark for good decisions. They had **1000 adults** write **their own** prompts to **GPT-5.2, GPT-5.6 and Gemini 3 Flash**. Then they simulated what happens if a person aged 22 to 89 lives by that advice. Then they repeated it with "academic" prompts, with full inputs and explicit assumptions.

What went well: the advice turned out better than the researchers themselves expected: more saving, stock market participation, diversification, lower equity share after 45. Taha Choukhmane, verbatim: "We were somewhat surprised by how good the advice was".

What went badly: • AI **adapts poorly to shocks** - it told someone who had lost their job to cut spending far too sharply, **even when they had savings** • it let the portfolio **drift** instead of actively rebalancing • **the advice depends on who writes the prompt**:

> *Prompts written by **men**, by the more financially literate, or by people with prior AI experience produced **~5% more wealth** by retirement. The difference in recommended equity share compounded into **~$50,000 (4%) less wealth at age 60** for women and the less financially literate. And for people who had not used AI for financial advice before, the model recommended a **lower savings rate**, **nearly $100,000 (6%) less** by 60.*

The author explains the mechanism plainly: people simply write different prompts. Women used the words "family", "groceries", "pay" more often; men used "strategy", "crypto", "growth".

Why it matters. The core conclusion: **the quality of the answer equals the quality of the inputs**. A model without explicit numbers will fill in an average person and answer that average person. In practice that means portfolio questions have to carry real numbers and stated assumptions: horizon, income, emergency fund. And one separate line: **drift instead of rebalancing** is easy to miss, because it does not look like a decision. [proven - peer-reviewed work with a described methodology; dated 21.07, a simulation, not real portfolios]

MIT Sloan · HN thread

4. Levie: why the hardest work is automated first 76.6K views. Yesterday he posted his thesis about the gap between everyday and deep AI; today it is built out into a mechanism.

Verbatim: "We are entering a strange dynamic where some of the 'hardest' work in the world is actually **the first to be automated, precisely because of its verifiability**. Math, cybersecurity and code - despite being insanely complex and high-value fields - have an advantage: **they can be objectively checked for correctness**. That gives two immediate benefits: model training gets **a clearer reward signal**, and running models lets you know it is working correctly, because the output can be tested **at scale**."

And the other side: "In other domains there is **far less instant verifiability**. Which legal terms a client will agree to, which marketing campaign to run in a shifting mood, which message a prospective buyer wants to hear, which financial goals and budget to set... **there is no 'one right answer'**. They depend on the opinions and risk levels of the operators, they are highly sensitive to input context, and **in many cases the right answer cannot be known for a long time** after the model produced its output."

Levie's conclusion: even with exponential model growth, "**far more will be done at the application layer than by the model itself**", and the processes themselves will have to change.

Why it matters. This explains yesterday's math and today's item 3 through one mechanism: Astra cracks number theory (Lean either compiles or it does not), while a financial advisor drifts (correctness shows up in 30 years). Where verification can be made cheap and fast, automation goes deep; where it cannot, AI stays an accelerator. The question "how do we make verification cheaper" matters more than "which model do we pick". [promising - a reasoned framework from a practitioner, matching an independent post by @max_spero_ on levels of verifiability]

Levie · @max_spero_ (the quoted post on levels of verifiability - the exact link did not come through in collection, so the profile is given)

5. Mollick: "let the labs work on these" - a list of unsolved problems outside math 20.9K views, fresh (03:23 Kyiv). A direct continuation of yesterday's item 1.

His thesis: "Math gets a lot of attention for its unsolved problems, but **there are unsolved and important problems in many fields** that could be solved empirically if AI really gets good enough. Problems that, if solved, **would deliver great value to society**."

Then he **writes out his own list** for entrepreneurship. What actually **causes** entrepreneurial success, beyond correlation. Whether exceptional growth is **predictable** at all, or an emergent path visible only in hindsight. Which skills can be **taught** in a way that raises success. What the **smallest possible intervention** is that moves a place from a low-entrepreneurship equilibrium into a self-sustaining ecosystem. Why some firms become less adaptive as they grow while others stay flexible.

In a separate post he added: "I think it would be useful for fields to list some of these problems. **Let the AI labs work on them.**"

Why it matters. The cheapest idea of the week, and it works without any new technology: write your own list of unsolved questions in your own field. The formula is simple: "what is

still unknown, and what would change if it were known". A to-do list and a question list are different documents, and almost nobody keeps the second one. [promising - this is a framing, not a result]

Mollick - the list · "let the labs work on these"

6. Qwen3.8-Max on Mollick's shader test: "solid, but not Kimi K3 level" Very fresh, 42 minutes before collection. An independent practitioner an hour after the release, which is why this is its own item.

Mollick ran his own test, a hard shader: "an endless city of neo-Gothic towers, half-flooded by a storming ocean". His verdict: "Baseline impression after a bunch of experiments is that this is **a solid model, but not Kimi K3 level**, from experience so far."

Why it matters. An illustration for item 1: the official blog says "a new level for coding", and an independent practitioner with his own test says "solid, but not top" an hour later. **Neither statement is lying**, they measure different things. Which is why lab releases are not worth repeating without a second voice. [proven - Mollick's own measurement, but n=1 and a subjective call]

Mollick on Qwen3.8-Max (the same shader test on Kimi K3 was run on 16.07, quoted inside this same post)

7. QM: 5,234 to 7,616 stars in a day. Third day, the pace holds A continuation of yesterday's item 4: the number grew again, so this is now a trend.

GitHub API right now: **7,616 stars, 800 forks**. Three-day timeline: **01.08 - 2,475, 02.08 - 5,234, 03.08 - 7,616**. That is **+2,382 in a day** after **+2,759** the day before. Growth slowed a little but holds at a high level; the repo was created on 29.07.

Why it matters. The idea they broadcast internally: once a week, pull the wrong turns out of the agent session logs and collect them in a separate "how not to do it" list. Today's item 2 (the ratio metric) is exactly such a candidate. [proven - numbers from the GitHub API, verified]

QM repo

8. MiniMax H3: an open video model with synchronized audio, 33B, fits consumer GPUs From the "AI + Product" newsletter, reposting Hugging Face - 405 likes on the model card.

Verbatim from @multimodalart's post: "MiniMax H3 just dropped on Hugging Face - **text-to-video, image-to-video, reference-to-video - all with audio**. A hefty **33B parameters** - but ready for **consumer GPUs** in diffusers and comfy."

Verified via the HF API: the `MiniMaxAI/MiniMax-H3` card was created on **28.07**, has **405 likes**, and the tags confirm the claims - `text-to-audio-video`, `audio-video-generation` and `synchronized-audio-video` are all there.

Why it matters. The class of "an open model that fits in a consumer card" already happened with images; now the same thing arrives for video with sound. For short clips without the

cloud, this is the line to watch. [proven - on the model's parameters and the existence of weights; **fuzzy** - on quality, since it was not run]

@multimodalart · the HF card

9. Collison on aesthetics: why everything got ugly 417K views, 3.4K likes, 430 comments, the most popular post of the day in the "Tech + Product" newsletter, and it is not about AI at all.

Patrick Collison (Stripe) answers a question about why, after a lifetime in STEM, he started thinking about aesthetics. In brief:

• "**A lot of things today are ugly** - and much uglier than they used to be or have to be. Once you see it, it is hard to stop noticing." Examples: early 20th century phone booths against modern ones, old drinking fountains against new ones. And the question: "**why did we stop making beautiful things? Is it a choice? Or has an evil spell been cast on us?**"

• Modernism as **a break in cultural continuity**: the 1932 International Style exhibition displaced the rich weave of earlier styles, and that break, he suspects, had consequences **beyond aesthetics** • And the strongest point, citing Elaine Scarry: "**beauty inspires creation. If so, the reverse may also be true: ugliness inhibits it.**"

• A historical argument: Petrarch set the preconditions for the Renaissance, which grew the scientific revolution. The 1851 Great Exhibition at the Crystal Palace drew **6 million visitors from a population of 21 million**: aesthetics and material progress were once popularly intertwined.

Why it matters. This is not news, and it was picked deliberately. For any team building a product, the claim that ugliness inhibits creation is a working hypothesis about the quality of the work. [fuzzy - this is an essay, not research; taken as a framing, not as fact]

Collison

10. Naval: "Agent Programming Interface" - three words, 85K views The most laconic post of the day. Naval wrote exactly three words: "**Agent Programming Interface**" - 85.2K views, 1.1K likes within an hour.

He gave no explanation. But the obvious reading suggests itself: **APIs were written for programs, and the users are becoming agents**. What suits a program (a stable schema, strict types) does not match what an agent needs: a description of intent, examples, error feedback in natural language.

Why it matters. The idea is already at work wherever a set of CLI commands is written for a model and documented in the prompt or in the tool descriptions. A fresh counterexample from this morning: **the Qwen page has two `<body>` tags, and a browser tool fell over on it six times in a row**, precisely because the interface was designed for a human with eyes, while it is an agent reading the selectors. That is the problem Naval wrote three words about. [fuzzy - this is an interpretation of three words, he explained nothing]

Naval

Misc

• **@karpathy added sources to yesterday's LOTR experiment** (90K views): he published the source code so it can be poked at and forked - karpathy.ai/lotr-movie (checked, returns 200). He also pointed at @simonw's pelican as the earlier canon for such tests. And the joke that caught on: "Expect GTA Hobbiton before GTA VI" post

• **@mckaywrigley under the same post**: "someone please fund the entire trilogy in this form as a benchmark" - 196 likes post

• **@emollick on chain-of-thought**: "In case you are curious - **'think step by step' has been tested, and it no longer works** to improve results in modern models". Wharton technical report: "The Decreasing Value of Chain of Thought in Prompting". In practice: the line "think step by step" in a prompt is now cargo cult post · and with humor: "I remember when we did chain-of-thought by hand, the old fashioned way. Summer of 2024"

• **@amasad put his chess engine on LiChess**: it plays humans and bots autonomously, currently **1253 Elo** in live games (against ~1500 in tests versus Stockfish level 0), "playing three games at once" post

• **@gdb three times in a day on ChatGPT/Codex as a business-task executor**: an agent that shot and published an ad on its own and "stopped at the Pay button and asked permission" (43K) 1; a Codex skill that turns customer feedback into a roadmap (59K) 2; "chatgpt work is the new task scheduler" (101K) 3. This is OpenAI marketing through user retweets, taken as a signal of direction • **@garrytan on meritocracy**: "Everyone confuses the map with the territory. Meritocracy is when **the territory matters more than the map**. And in markets the territory is the outcome: did you make something people want?" post

• **@lennysan - a new podcast**: "The CPO who wishes product management did not exist" with Tom Verrilli (Whatnot, ex-Twitch, ex-Twitter). Topics: "how AI **exposes product theater at scale**" and the fastest-growing PM skill, systems thinking (that is the **sixth** interview in a row with the same claim) post

• **@eladgil on LLM-as-a-judge**: "The ground truth for an agent is **how the agent should behave**". A precise diagnosis of why grading agents on the final answer breaks down post

• **@yishan on living in the singularity**: "If you want to do something with AI and it does not work, the working solution is literally this: **play video games for two weeks and wait for the technology to catch up**" - 77K views post

• **Go 1.27 Interactive Tour** - 345 points, second by score on HN. The main theme in the comments is extended generics: "Is anyone shocked that Go is finally embracing generics?" HN · the tour

• **Linux on the desktop passed 10% in North America** - two separate stories the same day, 176 and 98 points HN

• **@shreyas on product taste**: "Insight, intuition and taste are rooted in countless factors, so you cannot simply become great at product through advice and tactics. But many people

got significantly better after they **shed the lifelong need to feel smart and sound intellectual**" post

• **A 15-year-old built a cycloidal gearbox** and put it on GitHub - 320 points, one of the warmest threads of the day HN
