AI DIGEST

OpenAI cut the price of GPT-5.6 Sol by more than 20%. The live price list says $4/$20 per million, and the promo holds through 21 November

AI digest, Saturday 22.08 - the day the frontier got cheaper and trust in text got more expensive: OpenAI cut the price of Sol by 20%, and three separate pieces in one day say people physically cannot read machine text any more.

Saturday, 22 August 2026

IN THIS ISSUE

TOPIC 1

OpenAI cut the price of GPT-5.6 Sol by more than 20%. The live price list says $4/$20 per million, and the promo holds through 21 November

OpenAI's official announcement - 1.6M views, the loudest post of the day in both lists. The second post in the same thread clarifies the channels. Reaction from @gdb: "towards giving our customers the lowest price on the market for any task, as well as the highest ceiling on capability".

Figures from the live price list, not from the tweet. curl against openai.com is blind (403 on everything, including a made-up address), so the pricing page was opened in a logged-in browser:

MODEL	INPUT	CACHED INPUT	OUTPUT
gpt-5.6-sol	**$4.00**	$0.40	**$20.00**
gpt-5.6-terra	$2.00	$0.20	$12.00
gpt-5.6-luna	$0.20	$0.02	$1.20

Long context costs exactly double ($8/$30 on Sol).

Two things the tweet leaves out and the price list has: • The promo deadline is named exactly: "GPT-5.6 Sol's promotional pricing is available at least through **November 21, 2026**". The tweet says a vague "for the next 3 months", the docs give a date and the words "at least". That is a minimum guarantee with a number on it.

• **Data residency costs +10%** for models released after 05.03.2026. The fresh Claude Code (v2.1.239, last night) got a mirror line: `/cost` estimates now account for a **1.1x premium on US-only inference**. Both vendors started showing the residency surcharge as its own line within a day of each other.

Dedup against yesterday, and a correction. On 19.08 item 7 carried "GPT-5.6 Sol got twice as cheap" with prices of **$2.50/$15** and a note saying "but only on OpenRouter". Today the cut landed **in OpenAI's official price list**, and it is **smaller** than what OpenRouter gave a month ago: $4/$20 against $2.50/$15. That was a router with its own margin, this is a vendor price list. But anyone holding only the 19.08 figure in their head will read today's news as a price rise.

Why it matters. For anyone on a subscription, another vendor's price list changes nothing today. The interesting part is about **measurement**: within a day both vendors learned to show the residency surcharge as a separate line in the cost estimate. Most pipelines do not measure the cost of their own sessions at all. On a subscription that does not hurt; the question goes live the day something starts running through a per-token API.

TOPIC 2

Kagi added a switch to strip paywalled links from results - the most popular page of the day on HN, 1045 points. The feature itself is one line in the changelog, the audience wrote the rest

Kagi changelog, HN thread - **first place for the day by a wide margin** (the next item has 703).

The feature runs to **one sentence**: "we've added a setting for removing paywalled links from search results automatically". The rest of the release is a stocks widget, Markdown rendering in messages, search across threads, and two dozen bugfixes with issue numbers.

Why it is interesting as a signal. 1045 points went to **the mere option of not seeing paywalls**. The thread is a referendum on the state of the web, with Kagi as the occasion. A second detail from the same changelog: the fix list contains the issue **"'CHATGPT' Wikipedia article is flagged as slop"**. So a mechanism for flagging junk content exists, and it **fired by mistake on a Wikipedia article**. In the thread that is almost more interesting than the paywall filter: paid search already filters for "quality" too, and that mechanism has false positives.

Why it matters. In any pipeline that verifies sources, paywalls cut the check short: the body of the article is out of reach and all that remains is a "behind a paywall" note. Kagi **hides** the problem: the link disappears from results, the article stays unavailable. The useful version is the opposite - know **in advance** that a source is paid, and do not carry it as verified. An explicit list of paid domains next to the list of domains that serve robots a 403 does that job. It is the neighbouring category, and Kagi is not needed for it.

TOPIC 3

Dan Luu: "There is no reason for software to be slow". His agent spent a month building JIT compilation of regexes into ripgrep: 2-4x on long queries and 7% on the holdout set

168 points, the original, thread. Read in full.

This is a direct continuation of item 4 from 19.08, the "benchmark apocalypse": an agent spent a month beating Rust regex by 1.4x and turned out ten times slower on the control set. Today is the next step of the same experiment, and Luu holds the same bar of honesty.

What was done: • the same **FRE** engine, retrained so it would not be overfitted to `rebar` • then an idea nobody had tried because it was expensive: compile native regex code **on a separate thread** while ripgrep is already searching with the ordinary matcher, and switch over once compilation is done • result: **2-4x** on long simple queries, **~7%** on representative holdout queries from his real history • the key line: "we can just try this experiment with a few minutes of human time" - code surgery that for a human would have been visible work

The most valuable point is about economics. He quotes Michael Malice: there was a meme that "AI does not help, because code was never the hard part". In many areas that is true, but

JIT compilers are exactly the case where writing the code was the hard part. JITs are rare because historically they were too expensive to pay off. That barrier is down. Marc Brooker adds: the likely outcome is dynamic software tuned **to one specific workload**. Not to a class of workloads.

Why it matters. This is the same frame as yesterday's Karpathy on abstractions wearing thin, from the other end and with numbers. Anywhere "too expensive to optimise" was once the right call, that call may have gone stale. The argument "this is too expensive for a one-off task" now has to be checked every time.

TOPIC 4

DHH raised $8M for Linux on the desktop: eight patrons at a million each, among them the CEOs of Stripe, Shopify, Cloudflare, Dell and Jack Dorsey

163 points, the announcement - read in full. On X it is the loudest post of the day after OpenAI: @dhh, **1.2M views**, 9.9k likes, reposted by @patrickc, one of the patrons.

The patron list verbatim, $1M each: Tobi Lütke (Shopify), Patrick Collison (Stripe), Michael Dell (Dell), Jack Dorsey (Block), Matthew Prince (Cloudflare), Brendan Iribe (Sesame, Oculus co-founder), Jason Fried (37signals) and DHH himself.

The foundation is non-profit: it holds the trademarks, funds infrastructure and supports the open-source projects Omarchy depends on. DHH's wording: "it is a silly amount of money, so the intention is to make it last a long time".

Why it matters. Technically nothing changes. But **$8M of private money into desktop Linux in 2026** is worth attention for who is paying it: people whose companies hold half the infrastructure everyone uses (Stripe, Cloudflare, Shopify). Nearby in the feed sit @levelsio on Mac Search being broken and Theo on moving to Linux for agents ("the file system is faster"). That is mood. There is no measurement of agent performance on macOS against Linux, and inventing one is not worth it.

TOPIC 5

Three different authors wrote the same thing in one day: people have stopped reading machine text. One of them on the physical inability to focus

The strongest piece is "I'm becoming AI blind", 306 points, thread. Read in full.

Rafał Cymerys describes a specific symptom: he **physically cannot** analyse a document someone sent him. "It feels like it is being read, but it is impossible to focus on the content." It ends in endless ping-pong with the sender about things the document already says.

His three examples are painfully recognisable: a design doc that reads like a paste from Claude with the telltale jargon ("This cuts just through it", "The first gate is real"); a 20-page marketing deck where smart strategy is mixed with technical nonsense; requirements that read like **the model's internal deliberation**, unsure of its own decisions.

The explanation of the mechanism is the most valuable part: he considers himself "**pre-trained**" on AI posts in LinkedIn, newsletters and websites. His brain learned to catch the markers of machine text and **skip it without thinking**, the way banner blindness works.

Dedup, and why this is a separate item. Yesterday item 5 was dontpastetheai.com, 997 points: a **request** not to send machine text. The day before, a linter against "load-bearing" in the build, a fight with **machine language**. Today the same thing from the side of **a reader who can no longer do it**. Three days, three different places, rising sharpness: request, tool, symptom. It is called a trend only because it is the third independent piece in a row.

Why it matters. Cymerys has a marker that lands precisely: "pitching every little thing as a breakthrough". That is the main disease of a daily format, which pushes you to present a changelog line as an event. The best concrete advice in the text: if a document describes checkboxes in RBAC, do not sell it as though fire had just been invented. The only safeguard is a measured voice of your own, written down somewhere outside the author's head.

TOPIC 6

Codex on AWS Bedrock: cache writes ate 85% of the bill. $1,386 over four days where it should have been around $200

146 points, issue openai/codex#37674 - read in full, **closed**, 12 comments.

The substance: the native `amazon-bedrock` provider in Codex CLI **cannot turn on explicit prompt caching** for GPT-5.6 Sol. The request carries neither `prompt_cache_options` nor `prompt_cache_breakpoint`, and the provider config only covers transport and authorisation, with no transformation of the request body.

The author's figures from Cost Explorer (4 days, 05-08.08):

REQUESTS	CACHE WRITES	COST OF CACHE WRITES	TOTAL
3,656	171.94M tokens	$1,182	**$1,386**

85% of model spend is cache writes that were never read back: in a local session 76 requests produced 6.7M `cache_write_input_tokens` and **zero** `cached_input_tokens`. That is about 88k cache writes per request. The author marks it honestly: this is an estimate from usage data, not the final AWS invoice.

Why it matters. The most practical item in the issue, despite modest points. The failure mode generalises well beyond Codex and Bedrock: **an intermediate layer (gateway, provider, router) silently drops a parameter, and you find out from the bill.** Nothing crashes, everything works, the metrics are green. The closest analogue is yesterday's fix in

Claude Code v2.1.237: "Fixed prompt caching for sessions using an LLM gateway or custom base URL", the same class of bug in the same tool, patched two days ago. The rule is worth remembering: **if there is an intermediate layer between the user and the model, check that the cache is actually being read. Not only written**.

TOPIC 7

Thinking Machines is giving Inkling away free on OpenRouter, but only to agent harnesses and in exchange for telemetry

Announcement from @thinkymachines via Collison's repost (54k views), documentation - checked, 200.

The condition verbatim: the model is free on OpenRouter **only with agent harnesses**, for the next few weeks, and the data **will be used** (detached from accounts) to improve the model's behaviour. This is collection of real agent traces, named honestly in the announcement itself.

Alongside it in the same feed: Ox Alpha, a stealth model on OpenRouter with 1M context, text+image+video input, 230 points on HN. Collison on it: "It's very impressive". Whose it is has not been announced, and guessing is not worth it.

Why it matters. The mechanic is new and worth recording: **free access in exchange for agent traces**. This is the third story this week about agent telemetry becoming currency (18.08 - an autonomous agent in Jira, 20.08 - Linear with figures on agent tasks). Anyone about to run something through OpenRouter reads the data terms. Not only the price.

TOPIC 8

Anna's Archive: AI companies are buying up paper books and destroying them. The topic is booming on HN across two threads, but the text itself is from 5 August

703 points on the .pk mirror and another 542 on .gl - two separate threads about one piece, together more than anything else in the day.

The claim: companies buy millions of used books through intermediaries, scan them and **destroy** them to get data "untouched by machines" from before 2022. The motives the author names: deny competitors the scan, remove legal risk, and plainly, destruction is cheaper than scanning without damage. The call is for volunteers to scan rare books while they are still around.

Three caveats, without which this item would be misinformation:

1. **The piece is not from today.** The header says **2026-08-05**; it took off on HN yesterday, two weeks after publication.
2. It is a **guest post by a volunteer**, translated from Chinese. It has no editorial standing.

3. The key specific, Anthropic's "Project Panama", is presented as **revealed as part of the $1.5B settlement**. The settlement over books does exist and is publicly known, but **the link between that particular code name and it is not verified in primary sources**, and the author does not vouch for the detail. The rest of the post is a call to action and argument, without documents.

Why it matters. The item is here because these are the top two places on HN for the day. But the label is honest: **[fuzzy]** on the specifics, [proven] only on the fact that the text exists and collected those points.

TOPIC 9

Naval: "Frontier labs are incentivised to be afraid - it is a moral justification for staying closed"

@naval, 210k views, 3.6k likes - the loudest opinion of the day in the AI + Product list.

> *"Frontier labs have an incentive to be afraid, because it gives them a moral justification for staying closed. Chinese labs consistently take the position of doing what is good for everyone and being open. So the little guys end up in the awkward position of rooting for the Chinese..."*

Next to it, @emollick from the opposite side: lab CEOs were talking about existential risk **long before their companies were worth anything**, so they believed it; the PR move is more that they **stopped** talking about risk.

Why both versions are given without a pick. These are two incompatible explanations of one behaviour, and neither has data behind it, only an interpretation of motive. Showing the seam is more honest than slipping one version through as a conclusion. The only factual comment: in the past week OpenAI **halted training** over cyber capabilities (19.08, item 1). That is an action. Not a statement, and it leans towards Mollick.

TOPIC 10

DeepSeek shipped vision with no noise: `deepseek-v4-flash-vision-exp`, 472 points and no announcement

Documentation, thread. Their classic route: a page in the docs.

What it does: images alongside text (JPEG, PNG, GIF, WebP), with the format determined **by file content**, not by extension and not by MIME. Three ways to pass an image, all in the OpenAI-compatible Chat Completions format, and it works through the Responses API too. The request body limit is **48 MiB** on base64.

Why it matters. For image tasks (photos, screenshots, PDF analysis) nothing needs to change right now: the model already in use handles them. This is an option for bulk batch processing of images, where price starts to matter. An OpenAI-compatible endpoint means switching costs one config line.

misc: Kobo now runs apps - 466 points, Cobalt opens the e-reader to third-party code · Claudette: "make Claude stop sounding like a BuzzFeed article" - 225 points, the fourth piece of the day about machine style · accidentally logged hundreds of thousands of calls to US military bases - 478 points, e164.arpa interception · EU: copyright does **not** protect AI content 186 points · a week on Codex instead of Claude, a practitioner's notes 83 points · $250M at a $2.3B valuation for data centres **in orbit** - Nvidia and Cisco among the investors, 288k views · @lennysan: the habit of asking "could AI do this?" **before** doing it · @paulg: in capitalism the winners are those who make customers' lives better. Not those who squeeze out pennies - "2x versus 10x" · Exa hired the CTOs of Robinhood and Lyft · 100k hours of video of human hands opened for robotics, 15k tasks

---