

6 Charts For The AI Bulls

Some thoughts on Jevons in machina carne viva



MOSES STERNSTEIN

SEP 08, 2026  PAID

- climbing the wall of tokens aka we're gonna need a lot more compute
- Jevons it, to the moon
- open, closed, and everyone's a winner?
- it's a tiny share of users and you're complaining?!

👉👉👉 Reminder to sign up for the [Weekly Recap only](#), if daily emails is too much.
Find me on [twitter](#), for more fun.

✓ Subscribed

👏👏👏 Random Walk has been piloting some other initiatives and now would like to hear from broader universe of you:

(1) 🛎 [Schedule a time to chat with me](#). I want to know what would be valuable to you.

(2) 💡 Find out more about [Random Walk Idea Dinners](#). High-Signal Serendipity.

6 Charts For The AI Bulls

I generally think that “it’s so early” and “they don’t know what’s coming” are super-annoying phrases, used mostly by people who want to make predictions so vague, and on a time-horizon so long, that by the time they’ve been falsified (or not), no one will remember.

Like, if the “so early” thing never actually “comes,” no one bothers to remember all the bold predictions of it being “so early,” and “coming [unbeknownst to everyone else],” which is precisely what the predictors were counting on when they uttered the dreaded phrases. But, of course, if the vague gesture at the future “comes true,” the intrepid swami will inevitably produce “receipts” of his or her clairvoyance.

That all being said, when it comes to AI, it seems fair to say “it’s so early,” and people do seem to under-appreciate “what’s coming.”

Climbing the wall of tokens

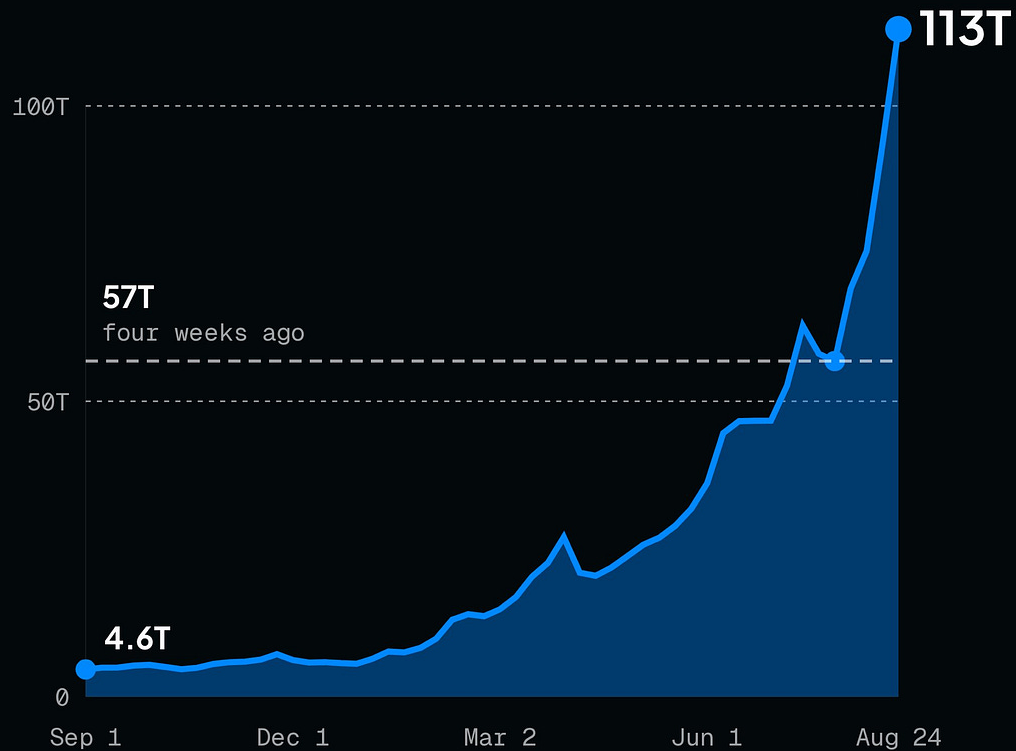
Well, not all people. The hyperscalers definitely seem to appreciate what’s coming, which is why they’re [spending trillions of dollars to make it happen](#).

The thing is that the human mind does not do a great job of extrapolating exponential growth, and woh mama, this is some exponential growth:



Tokens per week on OpenRouter | Sep 1, 2025 – Aug 30, 2026

25x in a year. 2x in the last month.



Source: OpenRouter analytics

openrouter.ai/rankings

Weekly tokens on just OpenRouter doubled in just 4 weeks, and as of the last week in August, were up 25x for the year.

Let that sink in for a moment.

LLMs have been out in the global public since 2023. We know that there is widespread “general” adoption. The baseline for token usage isn’t ground zero—it’s certainly not “high” in the big scheme of things, but as of September of last year, there were already 4.6T tokens visible to OpenRouter, and that’s not nothing.

And yet, in just four weeks, token usage *doubled*, from 57T to 113T. I can't think of any "unit of consumption" that has grown that quickly off such a high-ish base. Zoom went from ~10M to ~300M daily meeting participants in the four months following lockdowns. That's a 30x one-time bump of a much smaller base. By contrast, if tokens were to keep doubling every month for a year, they would increase 4,096x or reach 463 quadrillion tokens/week in twelve months time.

Now, there's at least one very good reason that token usage will not 4,096x over the next year: it's *physically impossible*.

I mean, there are other reasons why an epic four week run may cool a bit, but the most important one is that our [current supply of super-computers simply won't allow for it](#). There isn't enough compute to sustain that kind of AI demand.

As per Claude:

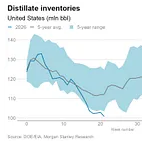
Sanity check on feasibility: at even ~1,000 tokens/sec/GPU effective inference throughput, that's ~460 quadrillion ÷ 605k seconds/week ≈ 765 billion tokens/sec, requiring on the order of 750M+ GPUs running flat out — versus a global installed base today in the low tens of millions.

So the doubling would have to break well before the year is out, at least without step-change efficiency gains.

In other words, to keep up with the current rate and pace of demand, we'd need something like 30-40x the current supply of GPUs.

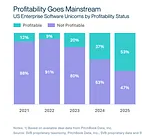
ICYMI

Oil Shock or Nah (and the Machinations of Iran's Biggest Customer)



MOSES STERNSTEIN • JUN 5

[Read full story](#)



Retail on a Heater

MOSES STERNSTEIN • MAY 30

[Read full story](#)



Why Isn't the On-Ramp Working?

MOSES STERNSTEIN • MAY 29

[Read full story](#)

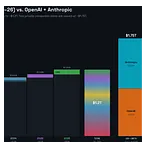
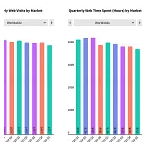


Chart Dive on Private Credit, PE/VC

MOSES STERNSTEIN • MAY 21

[Read full story](#)



That Escalated Quickly

MOSES STERNSTEIN • MAY 15

[Read full story](#)

Jevons It, to the Moon

Now, one doesn't have to think that token consumption will keep doubling every month to understand why AI capex keeps climbing such astronomical heights. Even a 10x token gain could reach the outer limits of planned capacity additions, depending on the mix of token usage and/or other efficiency gains (or non-gains). And that's just over the next 12 months.

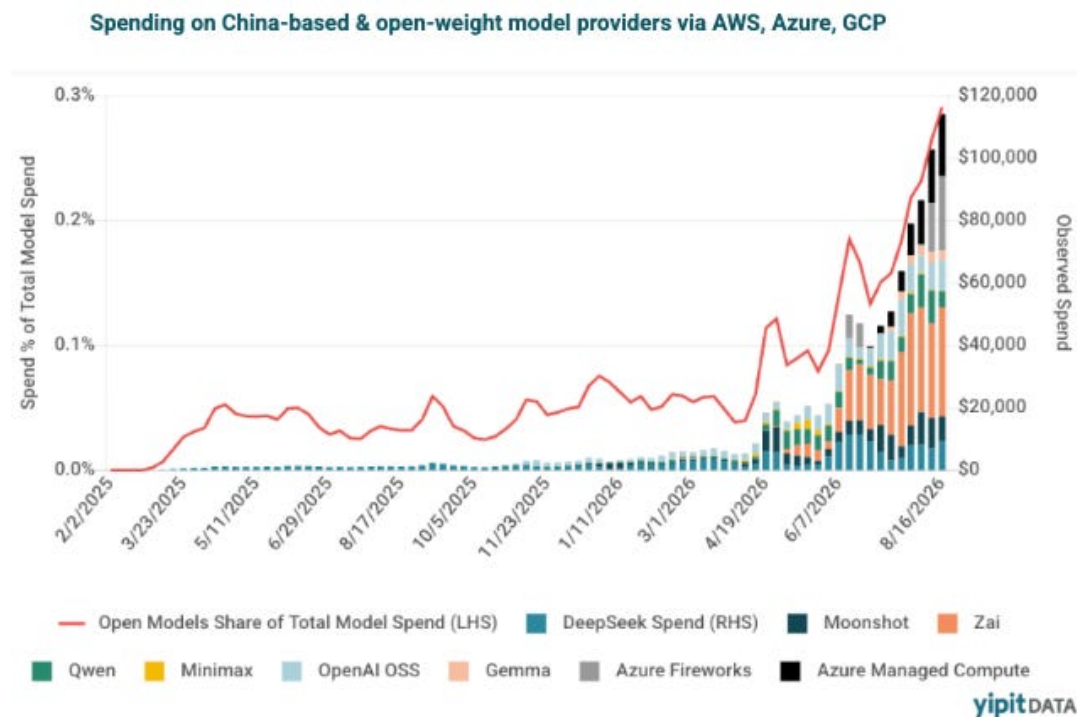
Sometimes it's important to remember that these aren't stupid people spending all this money on chips. They have real skin in the game, too. That doesn't mean they can't or won't make mistakes, or that the road is free from peril, but they know the hardware and they can do the math.

[Demand is undoubtedly running ahead of supply.](#)

But surely there's some catalyst for this recent surge?!

Well, yes, of course.

It's cost. [Low-cost open-weight models](#)—and [model improvement](#), and [gpu efficiency](#)—are pushing the efficient frontier of AI use cases way out to parts unknown.



Token spend on Chinese models has ~12x'ed since the spring.

On the one hand, that's a lot of [growth for Chinese knock-offs](#), but, on the other hand, it's still less than 0.3% of overall observed spend.

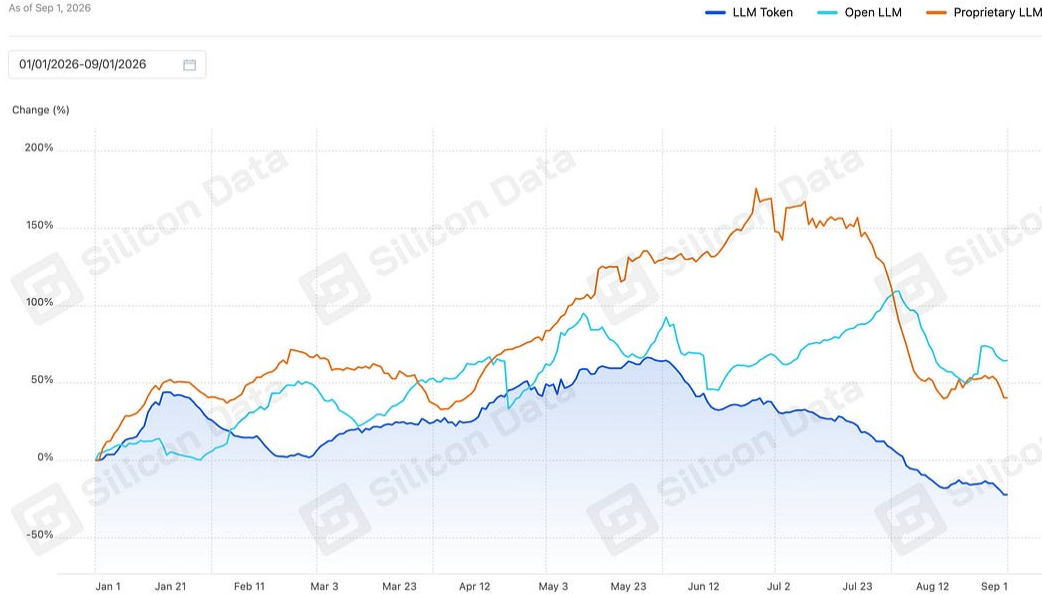
The thing is that everyone is getting more efficient, and cutting costs—partly for competitive reasons—but also because users are getting better at model optimization, and the models are themselves getting more efficient.

Price/per token is falling across the board:

TOKEN INDEX-LLM TOKEN EXPENDITURE INDEX (BLOOMBERG TICKER: SDLLMTK)

\$0.97/MTok **-21.9% ytd Change**

As of Sep 1, 2026



Source: Silicon Data. This index tracks changes in LLM token market pricing over time based on Silicon Data's pricing observations.

Silicon Data | SiliconIndex

2026/09/02 16:20:16

Silicon Data

The overall price per token is now lower than it was at the beginning of the year—and the biggest drop comes from the closed-weight models.

So, yes, part of this demand inflection is clearly driven by a dramatic across-the-board discount on AI—first, by the open-weight knock-offs, and then by the frontier labs. One can, indeed, fairly wonder [whether and for how long frontier models can engage in a pricing war with pirated competition](#), but the fact remains that [cheaper pricing has led to an explosion of demand](#)—which is to say that we’re nowhere near “peak demand” for AI.

Intelligence is Getting Cheaper and Everyone Is Winning?

And the other fact that remains is a point that Random Walk has been making for a while: in the big scheme of things, [it was always the case that AI would get cheaper and more efficient, and if demand did not](#)

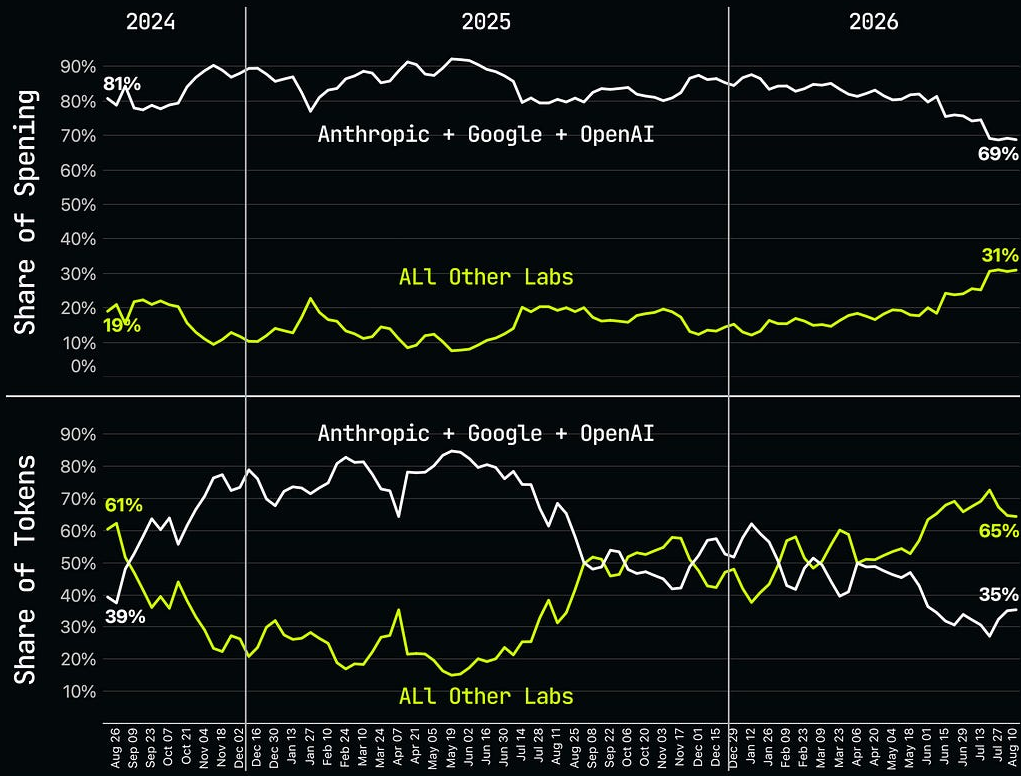
grow alongside efficiency, then *that would be the bearish scenario to worry about*. That pirated models have artificially bent the cost curve sooner than some people might have hoped or expected is certainly a material wrinkle for the labs themselves, but if anything, it's further validation for the AI bulls.

- Recall that it was the bears who were saying that models had hit a wall, and/or that cheaper models would become obsolete or that demand would quickly peak or that we're staring down the barrel of a predatory duopoly, *or whatever combination of arguments that made AI both so compelling, but also so useless, that in all events it would end in doom*.
- By contrast, it was the bulls who said, *no, models are getting better, and that as users and infrastructure become more mature, we're going to see an overall growing pie, where the highest-value token demand gets routed to the most sophisticated and expensive models, and lower-value work gets routed to older, cheaper, but still-useful models*.

In other words, in a more mature market, you'd expect to see a distribution of usage and spend precisely along these lines, again from OpenRouter:

Share of tokens vs share of spend across OpenRouter | Jan 1, 2024-Aug 20, 2026

Token volume is up 600x over 24 months



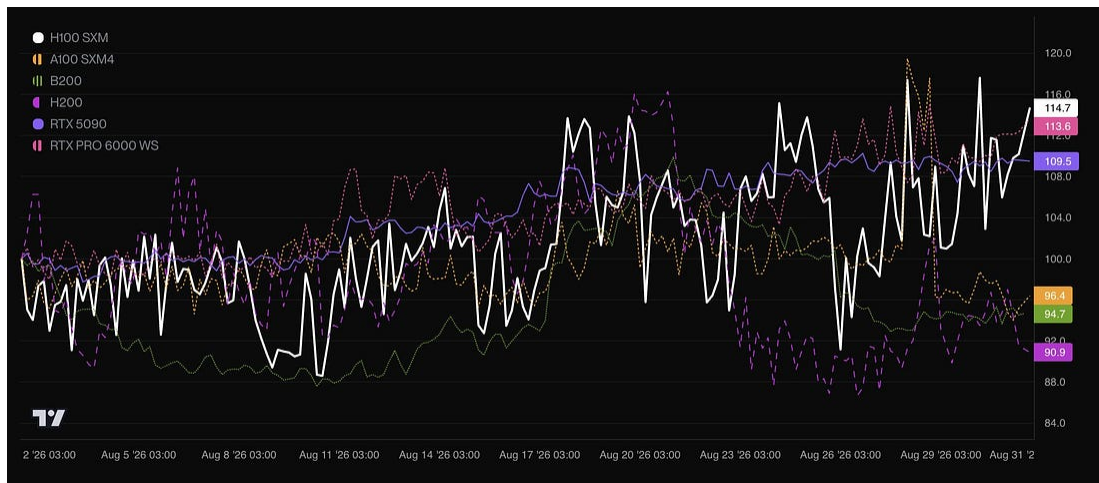
openrouter.ai/rankings

“All other labs” pulled-ahead in token share back in the Spring (even as OAI, Anth, and Google closed the gap recently), but the frontier is still capturing ~70% of the overall spend.

Frontier closed labs (plus Google) represent ~35% of the volume and ~70% of the spend. To repeat, that’s the distribution you’d expect where both a very expensive “frontier” and wider-spread market adoption were coexisting happily, side-by-side.

And, just to return to where we started, while all of this token-demand keeps ramping up-and-to-the-right, the super-computer tollbooth keeps

punching tickets:



According to Orn's data, GPU pricing is firm-to-firmer, across the board.

Cheaper AI, means more AI, which means more compute, and there is every reason to think that we're just "so early."

This is [Jevons Paradox](#) *in machine-carne viva*.

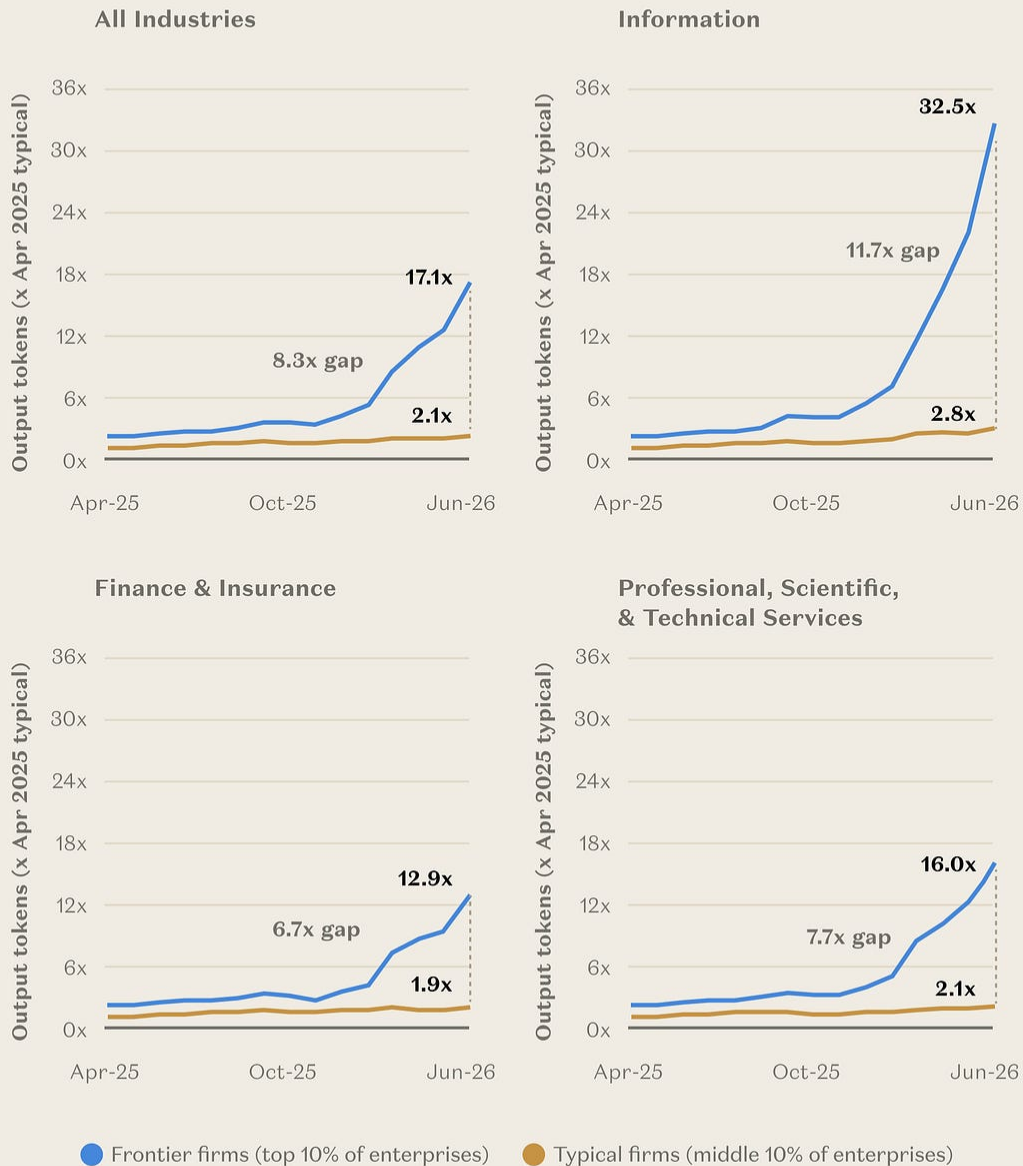
It's a tiny share of users doing all the work, and you're bearish?

To put an even finer point on it, it's AI agents more than anything that are driving both token demand, as well as the [share-shift towards cheaper \(and more numerous\) "cached" tokens](#).

[The people/firms using agents are a tiny share of the overall userbase](#), they are using orders of magnitude more tokens than everyone else, and both they (and agents) are just getting started.

AI Power Users Pulling Away From The Field

Output tokens, top decile v. middle decile, indexed April 2025



Source: OAI Signals, "What Frontier Firms are doing differently" Aug. 12, 2026.

More charts: a16z.news/subscribe

A16Z

Just extrapolate "power user" usage together with wider-adoption of increasingly sophisticated agents, and one can begin to consider where

this appears to be heading.

To oversimplify, what seems more likely?

Token consumption has gone parabolic because (a) users are playing around with the latest and greatest, and will soon tire of their toys; or (b) users are just beginning to scratch the surface of what models—that get cheaper and smarter every couple of weeks—can do?

My bet is firmly on the latter.¹

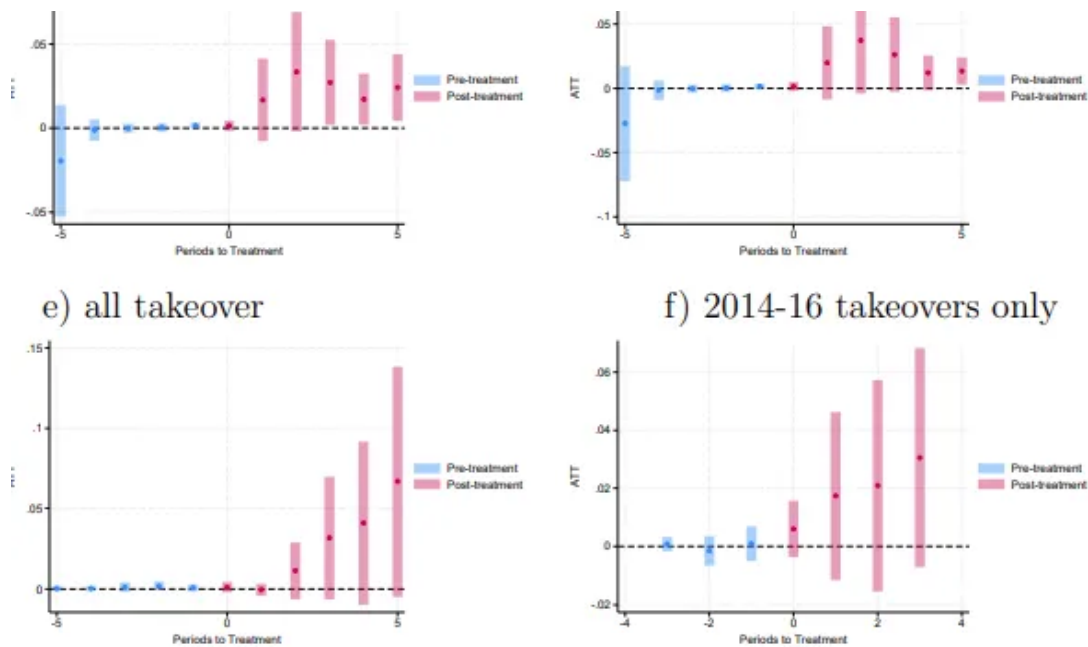
- Will there be trials and tribulations, along the way? Almost certainly.
- Are there outer limits on what users can spend on tokens, unless and until measurable benefits justify opening wallets even further? Of course.
- Will those benefits arrive soon enough, with enough magnitude, to keep sufficient pace with the current trajectory of spend (and for whom)? Only time will tell.
- Is some part of this demand subsidized by VC and other risk capital? You betcha.

But in all events, it surely seems like we're going to need a lot more compute, as the hyperscalers (and neoclouds, for that matter) have been trying to tell you. How much, how soon, and what cost, is perhaps a thornier question depending on your time horizon, but it's good to take the longer view of things, and that long view is looking pretty sound.

Previously, on Random Walk

Private Credit and Insurance, two peas in a pod (reprise), and a chart dump on default rates

MOSES STERNSTEIN • JULY 9, 2025

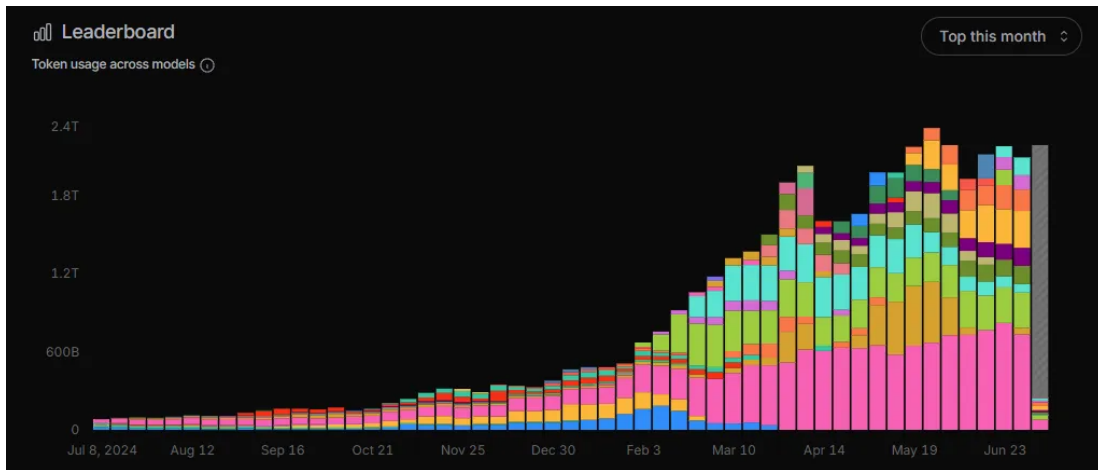


five charts on the rise of private credit in life insurance

[Read full story](#)

3 odd observations about AI that are probably misleading somehow

MOSES STERNSTEIN • JULY 8, 2025



paid adoption is going down?

[Read full story](#)

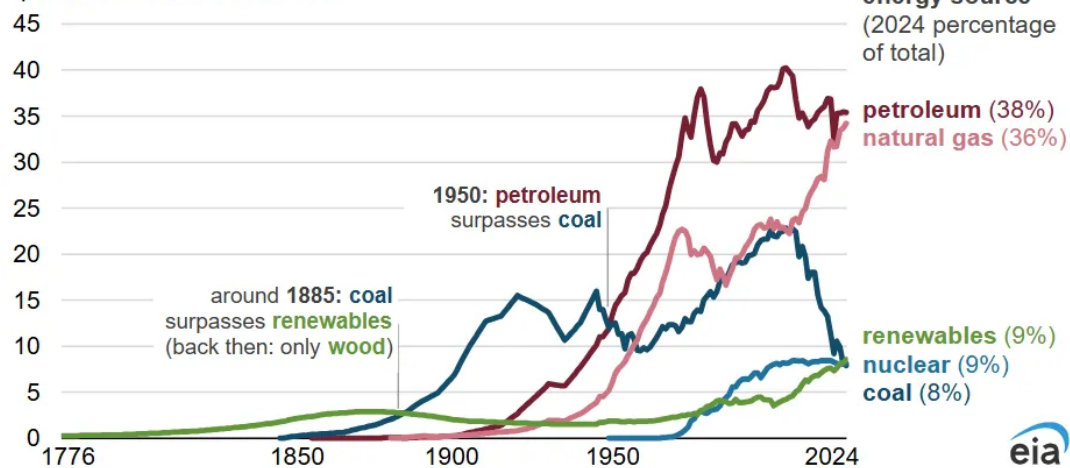
Energy in 1776

MOSES STERNSTEIN • JULY 4, 2025

How has U.S. energy use changed since 1776?

U.S. energy consumption (1776–2024)

quadrillion British thermal units



Data source: U.S. Energy Information Administration, *Monthly Energy Review*

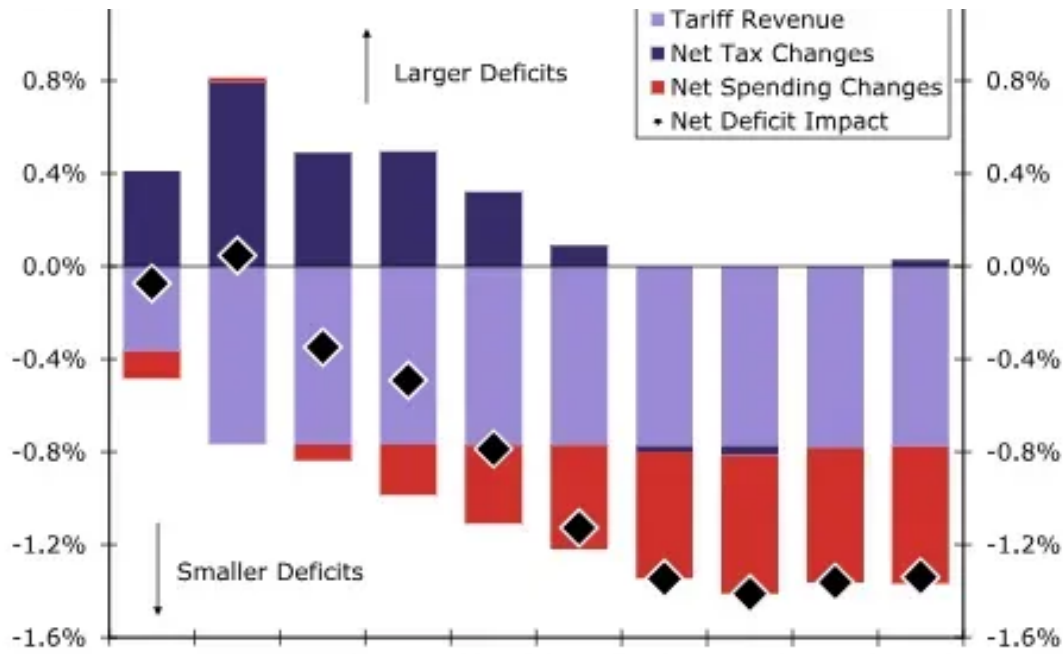
Data values: Primary Energy Consumption by Source and Estimated primary energy consumption in the United States, selected years, 1635–1945

It's July 4th, so Happy Birthday America, and we're going to keep it light and only semi-topical.

[Read full story](#)

Maybe the OBBB is actually fiscally responsible?

MOSES STERNSTEIN • JULY 3, 2025



frontloaded stimulus < backloaded cuts?

[Read full story](#)

LPs want their liquidity and they want it now

MOSES STERNSTEIN • JULY 1, 2025



an anecdote from the world of cap-intro

[Read full story](#)

*Random Walk is an idea company dedicated to the [discovery of idea alpha](#). **It's not investment advice, conflicts of interest abound, it's totally unreliable, and it reflects the thoughts and ideas of Random Walk, only.** Find differentiated data, perspectives and people, and keep your information mix lively. A foolish consistency is the hobgoblin of small minds. Fight the Great Idea Stagnation. Join Random Walk. Follow [me on twitter](#). Follow me on substack:*



Get more from Moses Sternstein in the Substack app

Available for iOS and Android

[Get the app](#)

-
- 1 I suppose it's also possible that only the power users figure this thing out, and no one else uses agents at all, but that seems extremely unlikely for all the obvious reasons.