Why Token Optimization Is a Gift to the Hyperscalers

UNCOVERALPHA
JUN 29, 2026 · PAID

Hey everyone,

A few weeks ago I wrote a piece called Most of the Economy Won't Run on the B Model, where I argued that the AI market will bifurcate: the frontier model goes small slice of work where intelligence is unbounded in economic value (drug disc novel math, the hardest agentic reasoning), and the middle of the economy — classification, extraction, summarization, routine code, support — runs on the cheapest model that clears the quality bar.

This article is the sequel to that one. Who captures the value when the world swi from token maxing to token optimization?

The shift away from always-buy-the-best-model is, on its surface, bearish for the labs and looks like it should compress the whole stack. But it is quietly one of the bullish structural setups for the three hyperscalers — Microsoft, Amazon, and G

Let me explain why.

Think about a highway toll road. There are two businesses operating on it. The f the company that manufactures the cars — they make the actual machine that do work of getting you somewhere, and there's a real margin in a car. The second is company that owns the tollbooth. The tollbooth owner doesn't care what car you Ferrari, Toyota, or a 12-year-old used Honda. Every one of them pays the same to cross the bridge. Now imagine a world where, suddenly, everyone realizes they w commuting to work in a Ferrari for no reason, and they all downgrade to the che Honda to save money. The carmaker's revenue per vehicle collapses, but people d drive less because they switched to a cheaper car. They drive more, because now

cheap enough to justify trips they'd never have taken before. And every one of th extra trips still crosses the bridge. The tollbooth's revenue goes up.

In the AI stack, the AI labs are the carmakers. The hyperscalers are the tollbooth And in the next months, we are entering a period where most of the economy is trading down from the Ferrari to the Honda, while simultaneously driving 10x m miles.

From token maxing to token optimization

For the last 18 months, the dominant behavior in enterprise AI was token maxin found the single best model on the leaderboard, you pointed every workload at it you didn't think too hard about cost, because the whole thing was a pilot and the was small relative to the perceived upside of "does this even work?"

That era is ending fast. Companies are essentially blowing through their AI budg a quarter. Altman said recently that the "my company spent my entire 2026 budg Q1, can you make this more efficient?" complaint went from something that "ne came up" to "all of a sudden a huge issue." And it's not just him; you can see it ac the industry, from companies like Salesforce and Meta to many other smaller one blowing their planned yearly budget in a matter of days.

The natural behavior change from this is that instead of having one model for everything, companies start routing: a small, cheap, often open-weight model ha the 80% of requests that are less complex, and only the genuinely hard requests escalate to the frontier. This is token optimization, and it doesn't reduce total tol consumption — it accelerates it. The moment inference gets cheap enough, you rationing it. You run the agent in a loop. You let it read the whole codebase. You run it five times and vote on the answer. A single coding-agent session now chew through millions of tokens of context where a chatbot query used a few thousand

You can see this in the hard numbers the hyperscalers themselves disclose. Micr said it processed over 100 trillion tokens in a single quarter in 2025, up 5x year-o year, with a record 50 trillion in one month alone. By its fiscal Q3 2026 call, Micr said over 300 customers were on track to process more than a trillion tokens eac

Foundry this year — and that this was accelerating 30% quarter-over-quarter. Go went from 480 trillion tokens per month at I/O in May 2025, to 980 trillion by Jul 1.3 quadrillion by October — and in its Q1 2026 filing disclosed that its first-part models alone were processing more than 16 billion tokens per minute via direct A up 60% in a single quarter.

At the same time, the price per unit of capability is falling substantially— the Sta HAI AI Index found inference cost for GPT-3.5-level performance fell more than fold in two years, and a16z pegs the decline at roughly 10x per year for any fixed capability level. And yet, total tokens processed are growing several-fold per yea

Where the margin lives in a routed token economy

When a company calls the SOTA model directly through, say, an AI lab's first-pa API, the lab captures the full economic rent of the token. There are many industr reports surfacing around the margins that providers like Anthropic have right no and many of them claim that the margins went substantially up this year to as hi 70% gross margins. That price embeds the lab's R&D, its brand, its leaderboard position — call it the "model-provider margin."

And to be fair, when token optimization happens, there might not be a direct hit margins in the short-term at these frontier labs (as there will always be use cases the best AI model), but it does mean that growth acceleration becomes smaller th would be if everyone stayed for every use case on the frontier path.

The company routes that same workload to an open-weight model — GLM 5.2, DeepSeek V3.2, Qwen3 Coder, Kimi K2.5, Llama, MiniMax M2.5 — running on a hyperscaler's managed inference. The "model-provider margin" essentially goes t zero, because the model is open-weight and nobody is charging a brand premiun it. But the token still has to run on somebody's GPUs, inside somebody's data cer behind somebody's managed API with its security, compliance, logging, and SLA And that somebody are the hyperscalers. They still charge their full infrastructur margins on the token. AWS has historically run a roughly 35–38% operating marg Google Cloud, which lost money for years, posted operating margins above 33% i

2026 and is still climbing. That margin doesn't care whether the token came fron $50/million frontier model or a $1/million open-weight model. To return to the a that from before, the tollbooth charges the same toll regardless of the car.

So here is the structural beauty of it for the hyperscalers, and the structural dang the labs:

- Per-token economics compress, but the compression lands almost entirely o model layer, not the infrastructure layer. The AI lab's margins on simple workloads get squeezed. The hyperscaler's infrastructure margin is sticky.
- Total token volume explodes, and almost every single token crosses the hyperscaler's tollbooth. More usage, on a partly-depreciated, increasingly ef installed base, means absolute infrastructure revenue and gross profit dollar up even as the price of any individual token falls.

This is Jevons' paradox pointed directly at the cloud P&L. The hyperscalers sque more revenue out of the infrastructure they already own, and they don't need the model-provider margin to do it. Outside of Google, they were never in that busin the first place.

The orchestration layer brings real value

If the future is multi-model — and it clearly is — then someone has to own the orchestration: the layer that decides which model handles which request, holds t fine-tunes, runs the agent loop, manages memory and tool-calling, handles fallba when a model is down. This layer can bring immense value to those who own it, all three hyperscalers are naturally positioned to become just that.

Amazon's Bedrock is the clearest example. It's no longer "a place to call Claude." 2026, the catalog spans 18 providers and 110+ individually addressable model var and AWS bolted on Intelligent Prompt Routing, which automatically routes each request to the cheapest model in a family that can handle it — AWS claims up to cost savings with no accuracy loss. On top of that sits Bedrock AgentCore, a production agent harness with built-in Runtime, Memory, Gateway, Browser, Ide and Observability. Bedrock reached a multi-billion-dollar annualized run rate wi customer spend growing 60% quarter-over-quarter across 100,000+ customers.

Microsoft's Azure AI Foundry and Google Vertex are the same play, although Go preferred scenario is where the Gemini model family dominates the workloads.

The natural position for hyperscalers to be the orchestration layer is great becau companies already have their data, security perimeter, billing, and compliance liv the cloud environments. They can pick from a menu of models." The lab moves f the front end — the thing the customer chooses and bonds with — to the back e interchangeable component sitting behind the hyperscaler's harness. Customers start building the fine-tunes, the RAG pipeline, the agent scaffolding, the eval su the cloud orchestration layer, not at the model layer; switching models becomes

something normal and not a migration task. One of my readers, who runs his ow harness, put it perfectly in the comments of my last piece: the harness is worth a much as the model itself, and you're paying for it either way.

The moat for the labs becomes shallow with most use cases but still remains imp at the most complex ones (the ones where economic upside is not capped). But fo hyperscalers, they benefit in all use cases that run the economy.

One might argue why companies couldn't move even outside of the cloud, but th problem is here that besides the normal reasons that people migrated a big part workloads from on-prem to cloud (easier to scale, use, collaborate, manage), now top of those reasons, you also have the problem compute shortages where infrastructure outside of cloud environments is even harder to get and managing infrastructure becomes a much more difficult job. But an important aspect, I thi that is forming is also the cybersecurity one. Given the recent developments here clear that only a handful of companies will have first-row access to the most capa cyber models, to first shield their systems before the model is then available for everyone else. If you are an enterprise, you want safety, and it seems the only way you to get it is if you have your infrastructure at one of those preferred partners t get first-row access to cyber-capable models before everyone else does. The hyperscalers are those partners.

The hyperscalers' job, then, is almost simple to state: enable as many models as possible, make routing and fine-tuning and observability frictionless, and let customers optimize tokens to their heart's content. Every model they add makes tollbooth more valuable.

The government approval window just made this worse for the labs

Now layer on a development from June that I don't think the market has connect this thesis at all: the June 2, 2026 executive order: Promoting Advanced Artificial Intelligence Innovation and Security.

The headline version is that it sets up a voluntary framework where developers o "covered frontier models" — designation determined by the NSA via a classified

benchmarking process focused on cyber capabilities — give the federal governm access to the model for up to 30 days before release to "other trusted partners." (T earlier draft had a 90-day window; it was cut to 30 to avoid blunting U.S. competitiveness.) The order pointedly does not create mandatory licensing — bu establishes a structured pre-release evaluation pathway, an AI cybersecurity clearinghouse run out of Treasury, and a government role in selecting which "tru partners" get early access. The catalyst, per CFR's reporting, was rising concern models like Anthropic's Claude Mythos being able to autonomously find and exp software vulnerabilities. Commerce ordered Anthropic to cut off non-U.S. access Mythos 5 and Fable 5 models on export-control grounds, and those models were from Bedrock days after launch. Now there is growing concern that, even when r are released to the public, U.S. citizens may get access before people from other countries. Hyperscalers have clients worldwide, so it becomes even more import them to offer multiple models to companies and for companies to have an orchestration layer that isn't just one AI lab, since they will have to juggle geopo compliance as well.

It strengthens the case

for entering AI through the hyperscaler. If you're a regulated enterprise — a ban hospital, a utility, exactly the critical-infrastructure operators the EO names — y safest path to frontier capability is increasingly through a cloud provider that ha the compliance, security review, data residency, and access-control machinery fo and gives you a menu of pre-vetted models. The EO effectively raises the compli overhead of touching frontier AI, and compliance overhead is precisely the thing hyperscalers are built to absorb and resell.

There is another aftereffect of this government policy. It narrows the economic window the AI labs have to monetize a frontier model. A lab builds a genuinely s of-the-art model. Historically, it gets to sell direct access to that frontier capabili premium for as long as it takes competitors and open-weight models to catch up that catch-up window is the lab's profit window. But now insert a mandatory-in-practice review period and a "trusted partners first" distribution gate at the fron

that window for the most capable (most cyber-capable, most monetizable) model frontier model's clock starts later and runs against open-weight competitors whc capabilities are compounding fast.

Also imagine a closed competitor releases a model that is less cyber-capable thai designated frontier model and therefore sails through review and reaches the op market faster, it might be days after the SOTA model, as the SOTA model was lo in the review process longer, meaning the duration of the economic value of the for the AI lab shrinks even more. The most capable model can paradoxically reac broad commercial availability slower than a slightly weaker one. The premium w — the period where a lab has the best model and can freely sell it to everyone. Th slice of time where "best model" equals "uniquely monetizable model" is shrinki long as we have at least two companies competing in the SOTA race.

And who is indifferent to all of this? The tollbooth. The hyperscalers host the frc models and the open-weight models and the not-quite-frontier closed models. Whichever one wins the customer's routing decision in any given quarter, the tol still crosses their bridge. They are structurally hedged across the entire model la which is exactly where you want to be standing when the model layer is getting commoditized and regulated at the same time.

Summary

The intuitive read on "token optimization" is bearish: prices are collapsing, the A are getting commoditized, surely this compresses the whole AI trade. And it is b — for the labs' pricing power on commodity workloads (but not for the most con tasks), and for anyone whose thesis depends on a single model maintaining a dur premium.

But the same forces that compress the model layer expand the infrastructure laye take. Here's the asymmetry as I see it:

The hyperscalers win on volume, mix, and stickiness simultaneously. Volume, be cheaper tokens get consumed in vastly greater quantity and everyone pays the infrastructure toll — you can watch it in the token-processing disclosures growi

several-fold a year against per-token prices falling 10x a year. Mix, because as workloads shift to cheap and open-weight models, the lab margin in each token evaporates but the hyperscaler margin is untouched, so the hyperscalers capture larger share of a token's total economics than before. And stickiness, because the orchestration harness — Bedrock AgentCore, Foundry, Vertex — is where the customer's real switching costs accrue, not the model.

On top of that, this lands on a cost base that is increasingly paid-for. The 2026 ca enormous — AWS guiding to ~$200B, Microsoft over $100B, Google $175–185B - the bear case is that this capex never earns a return. But if you can squeeze an or magnitude more useful tokens out of installed hardware and keep filling new har with exploding demand, you get operating leverage going vertical: incremental t revenue landing on partly-depreciated, dramatically-more-efficient-per-token infrastructure. That's the scenario where cloud converts from a capex incinerato a cash machine.

The labs, by contrast, are being squeezed into a narrower and narrower strip of t map: the genuinely frontier, genuinely unbounded-value workloads where somec will still pay the Ferrari price — and even that strip now has a regulatory speed b bolted to the front of it. It's a real business. Anthropic's trajectory shows there's enormous value there. But it's a narrower business than the "every token runs on model" dream, and it's one where the customer relationship increasingly belongs the cloud provider sitting in front of them.

The market right now is mostly priced for a world where capex keeps compound and the hyperscalers are making a losing bet on it. I think that has it close to backwards. The token-optimization shift everyone reads as a deflationary threat the owners of the tollbooth, a Jevons-paradox tailwind with a regulatory moat th in for free.

As always, I hope you found this article valuable. I would appreciate it if you cou share it with people you know who might find it interesting.

Thank you!

Disclaimer:

I own hyperscaler Meta (META), Amazon (AMZN), Microsoft (MSFT), Google (GOOGL) stock.

Nothing contained in this website and newsletter should be understood as invest or financial advice. All investment strategies and investments involve the risk of Past performance does not guarantee future results. Everything written and expr in this newsletter is only the writer's opinion and should not be considered inves advice. Before investing in anything, know your risk profile and if needed, consu professional. Nothing on this site should ever be considered advice, research, or invitation to buy or sell any securities.

99 Likes · 14 Restacks

Discussion about this post

Comments Restacks

Write a comment...

William Dwyer William's Substack 29 de jun.

Liked by UncoverAlpha

I enjoy your articles and your views.

Visit my substack: gentleinspiration.substack.com

Friendly Christian Advice for a Happy Life

LIKE (1) REPLY

Murali Murali 8d *Edited*

"On top of that, this lands on a cost base that is increasingly paid-for. The 2026 capex is enor AWS guiding to ~$200B, Microsoft over $100B, Google $175–185B — and the bear case is tha capex never earns a return. But if you can squeeze an order of magnitude more useful tokens