AUG 12, 2026

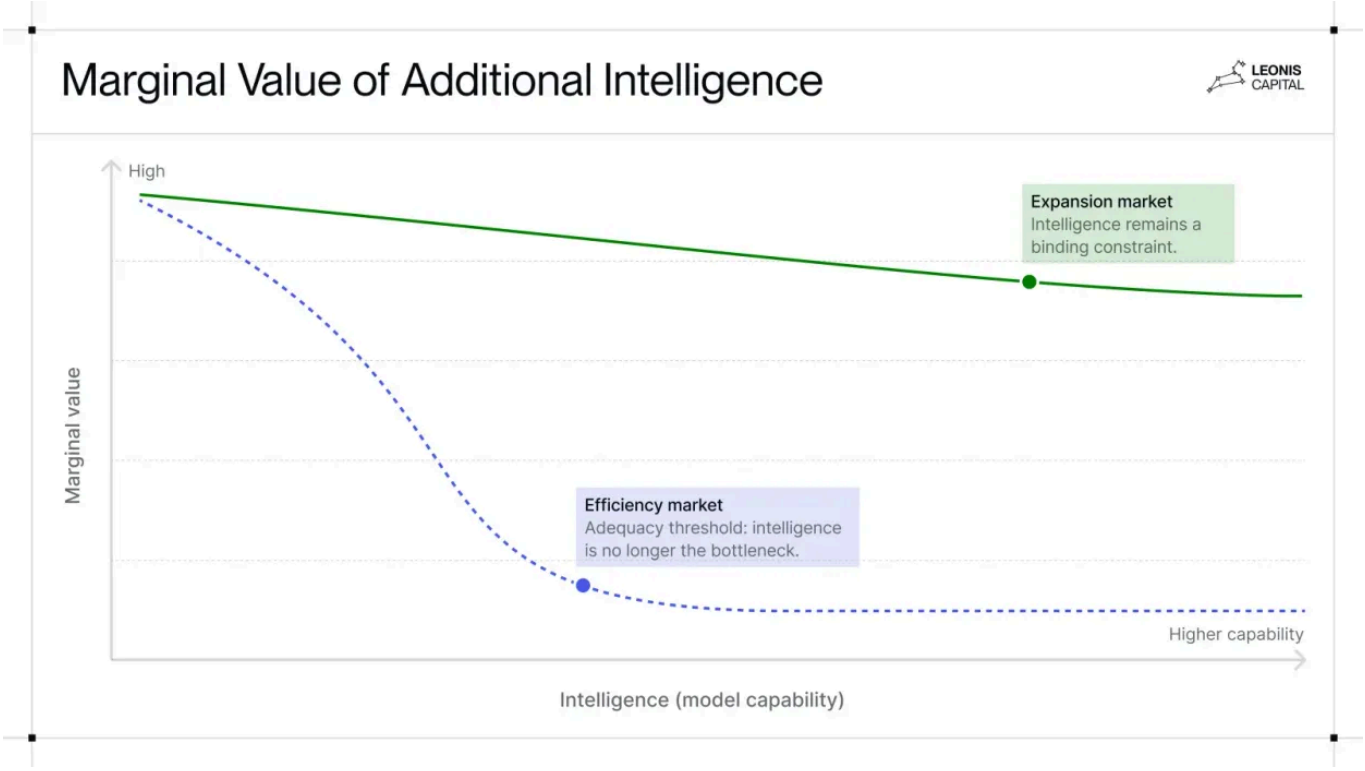
The Two Token Economics of AI

Why the same token is a bargain for one buyer and an overpay for another



By Liang Wu, Jenny Xiao, and Jay Zhao

Leonis Capital · leoniscap.com/research/the-two-token-economics-of-ai/



It wasn't too long ago that the most prominent voices in AI told companies to spend more. Jensen Huang said a $500,000 engineer should be consuming at least $250,000 of AI tokens a year, and that Nvidia was working toward a $2 billion annual token budget for its own engineers. Workers took the hint, and "tokenmaxxing" became a named behavior, mostly employees burning tokens to look productive.

On the other side, the bills arrived. Uber blew through its entire 2026 AI coding budget by April and capped engineers at $1,500 a month, with its COO admitting the link between token spend and output "is not there yet". Microsoft revoked Claude Code licenses months after issuing them. Meta discovered its employees had consumed 73.7 trillion tokens in a month, partly driven by an

internal leaderboard, and dismantled it. Forbes ran the math under the headline "AI Costs More Than The People It Replaced," and some executives are quietly concluding that the engineers they laid off were cheaper than the model that replaced them.

Two paradoxes hang over this. First, why are some companies desperately trying to cut AI costs while others spend freely on frontier models? Second, if open-source models keep improving toward frontier capability, why do closed frontier models still command premium pricing?

The market has been trying to answer these one at a time and getting neither right. Both resolve together once you notice a single question hiding underneath them.

Is intelligence the bottleneck for the business?

If the answer is yes, more intelligence directly produces more revenue, and the buyer will pay for the best model they can get. If the answer is no, intelligence is already sufficient and something else is holding output back, and the buyer will pay only what the labor being replaced was worth. Two different answers, two different economics.

Is Intelligence The Bottleneck In Your Business?

The economic value of AI depends on whether intelligence is the bottleneck on what the business is trying to produce. If it is, the buyer sits in an *expansion market*, where additional intelligence directly creates more economic value. If it isn't, the buyer sits in an *efficiency market*, where additional intelligence only drives cost down against a fixed output.

Expansion buyers use AI to create revenue. Something else is capped, but intelligence is the input that lifts the cap. Better trades produce better returns. Better drug candidates produce more shots at a blockbuster. More shipped products produce more customers. Because intelligence is the constraint, more of it produces more output, and the buyer's ceiling on spend is set by the returns those outputs generate, not by any wage bill. There is no "good enough," because a better model produces more of the thing that makes money. Buying frontier intelligence for a billion-dollar trade is like buying the fastest car on a track where lap time is the prize. The performance is the point, and you would pay more for more of it.

Efficiency buyers use AI to cut costs or drive efficiency to get more out of their business. Intelligence is not the constraint on output, something else is. For Uber, this is riders and drivers, for Costco, this is the number of physical stores and their supply chain, for an insurer, it is underwriting capacity and regulation. And for most enterprises it is headcount and process. AI helps at the margin by replacing labor, resolving the ticket, extracting the field, summarizing the document, and processing the claim. The ceiling on spend is the wage bill it replaces, because no rational buyer pays more to automate a task than the labor cost. Once a model is good enough for the job, extra intelligence adds nothing. Resolving a support ticket with a frontier model is like

commuting in a Ferrari. You arrive at the same place having paid for performance you never used.

The difference is each segment's production function.

In expansion markets, output rises with intelligence:

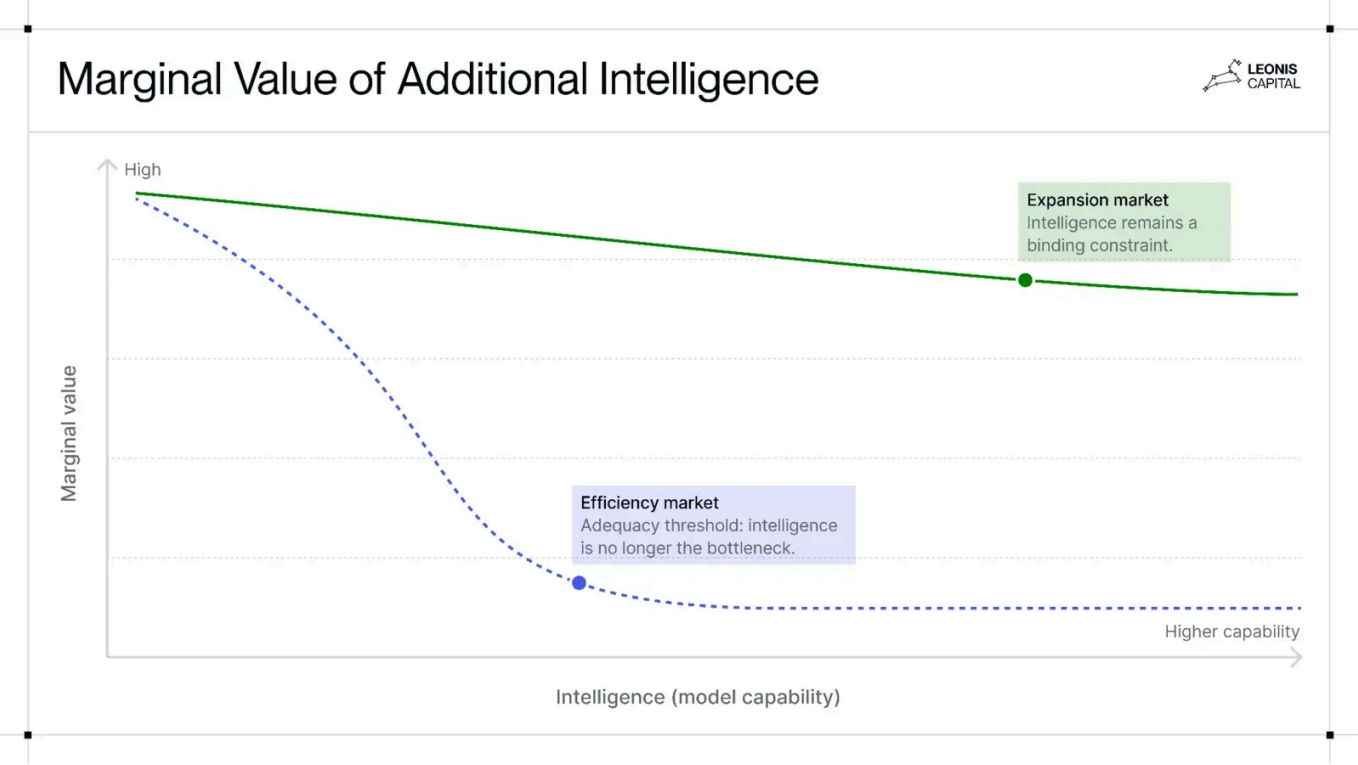
```
Output = f(Intelligence)
```

In efficiency markets, output is capped by the non-intelligence constraints:

```
Output = min(Intelligence, other constraints)
```

Once intelligence exceeds the threshold set by those other constraints, additional intelligence produces no additional output. Which production function you are running determines your optimization problem, and that is where the two camps diverge. Expansion buyers maximize output, because intelligence is the constraint they are trying to relax. Efficiency buyers minimize cost, because intelligence is a substitute for something already priced.

The cleanest way to see this is to plot the marginal value of intelligence for each buyer type.



Layering on the above chart, we can see Efficiency versus Expansion buyer's production function on a chart as it relates to intelligence and output.

It's the Objective, Not the Task

A common shortcut is to sort task categories into the two segments, for example customer support is efficiency, coding is expansion. That shortcut fails. The same task belongs to different markets depending on what the buyer is trying to do with it.

Software engineering is the clearest case. For a mature business with limited room for top-line growth, code is a cost center, and AI coding spend largely follows efficiency logic. That is why Uber, whose engineers now generate roughly 70% of committed code with AI, still capped their spending. A ridesharing platform's rider demand is not intelligence-constrained, so more sophisticated code does not translate into more rides. But for an early-stage startup, ten times more code can genuinely mean ten times more product and ten times more customers. Intelligence is the binding constraint, and the same coding task lives in a different market. Ironically, fast-growing startups pay for frontier coding intelligence more readily than enterprises with a thousand times their budget.

The same enterprise can even run both plays at once. In June, Coinbase CEO Brian Armstrong published the company's AI cost strategy. Coinbase had cut total AI spend nearly in half even as token usage kept climbing. The levers were all efficiency moves, defaulting engineers to cheaper open-weight models (GLM 5.2, Kimi 2.7) through an internal gateway, routing each task to the cheapest model that clears the bar, and pushing its cache hit rate from 5% to 60%. What holds for Coinbase's engineering org would flip for a new business line. For example, its push into prediction markets is exactly the kind of bet where a frontier model could expand the market itself rather than just execute existing work more cheaply. Same company, opposite objective, opposite spend logic.

The same sorting happens across functions, e.g., customer support is efficiency while customer upsell is expansion, even when the two sit in the same part of the org chart. Complex planning may need the frontier model, execution tasks may not.

Buyer	Function	Segment	Model Choice	Spend Logic
Marketplaces (e.g., Uber)	Engineering, Ops	Efficiency	Cheapest adequate	Cap at payroll offset
AI support vendors (e.g., Decagon)	Customer Support	Efficiency	Cheapest adequate	Cap at payroll offset
Early-stage startup	Coding	Expansion	Best available	Spend while output grows
Enterprises (new product lines)	New product Line	Expansion	Best available	Spend while output grows
Hedge funds	Trading	Expansion	Best available	Spend to generate returns
Biotech	Drug discovery	Expansion	Best available	Spend against blockbuster upside
Consulting Firms (e.g., McKinsey)	Consulting Delivery	Expansion	Best available	Spend to match rivals

Why Most Enterprises Will Be Efficiency Buyers

Under this framework, almost all Fortune 500 enterprises are efficiency buyers. Their businesses are not intelligence-constrained. They are constrained by demand, distribution, regulation, physical assets, network effects, headcount, and the other things large enterprises actually run on. AI helps at the margin, but the margin is capped by whichever of those constraints is binding, not by how good the model is.

The implication for enterprise AI spend is that after the tokenmaxxing frenzy is over, most companies will settle into efficiency-buyer behavior. The FinOps tooling, the routing startups, the CFO think-pieces reading like a structural crisis in AI economics are not describing a crisis at all. In fact, 98% of practitioners now manage AI spend, up from 63% in 2025, and 73% of enterprises exceeded their AI cost projections in the past year. Those numbers describe the market repricing its dominant segment. The 2026 backlash is not a verdict that AI spending is unsustainable, instead it showed **that efficiency buyers had briefly paid expansion prices for efficiency work and have started correcting**. A few observations follow from this.

The next phase is therefore not less enterprise AI adoption, **but a shift from access to allocation.** In the first phase, companies bought seats, encouraged experimentation, and treated token consumption as a proxy for adoption. In the next, they will manage AI as a portfolio: centrally governed multi-model gateways, budgets assigned by workflow and risk level, cheaper defaults for efficiency work, and flexible frontier spending for expansion work or high-stakes judgment. We expect model access will become an escalation policy rather than a company-wide entitlement, and procurement will increasingly measure cost per accepted task and value per dollar instead of seats, benchmark scores, or raw token volume.

That distinction will matter in public markets as well. Enterprise AI usage can keep rising even as spending per task falls and workloads migrate toward cheaper inference, meaning token growth alone will reveal little about revenue quality or pricing power. Investors will need to distinguish businesses capturing premium expansion demand from those processing commoditizing efficiency volume, and from the control-plane platforms that own routing, governance, workflow context, and outcome measurement.

In light of this, cheap open-source models have an opportunity to capture the market. Once intelligence exceeds the task threshold, Fable 5, GPT-5.6, Kimi 3, and GLM 5.2 all produce nearly identical business value. The question for the open-source labs was never whether they could beat closed-source models head to head. It was whether they could **get good enough for the efficiency market**, which is where most of the enterprise workflow volume actually sits. Kimi 3 and GLM 5.2 have plausibly crossed that chasm this quarter. For efficiency buyers, routing to whichever open-weight model clears the bar is not a temporary optimization, but a new equilibrium. Companies are starting to fit this pattern and it will generalize over time.

Open source is a bigger near-term problem for the closed frontier labs than the discourse admits. A meaningful share of frontier revenue today comes from coding and other tasks that are efficiency work for the buyer, even if they used to require frontier capability to complete. As the open-source floor rises to meet those tasks, that revenue erodes. The strategic answer is not to defend the efficiency turf, which the labs will lose to open source and cheaper alternatives. Instead, it is to move faster into expansion markets where intelligence is genuinely the bottleneck and buyers will keep paying for the best. Anthropic's timeline shows this pressure at work. Claude for Financial Services shipped in July 2025, then Claude Science with an internal drug-discovery program landed in June 2026.

Efficiency-market pricing converges to infrastructure logic. When AI replaces a known cost and the ROI is bounded, buyers want the meter visible, so pricing settles into per-token or per-API-call billing, capped and observable. The competitive layer moves down the stack, toward inference cost, deployment, customization, workflow integration, and the boring instrumentation that lets an enterprise control what its AI actually costs. This is why Databricks raised $3B at a $188B valuation as CEO Ali Ghodsi described enterprises shifting from "tokenmaxxing to valuemaxxing," why OpenRouter hit $50M in annualized revenue in April fielding multi-billion-dollar acquisition interest, and why Spectro Cloud closed an oversubscribed $100M+ Series D building

infrastructure control for enterprise AI. None of these companies bet on the frontier moving. They bet on the efficiency market being the largest segment, and being underserved.

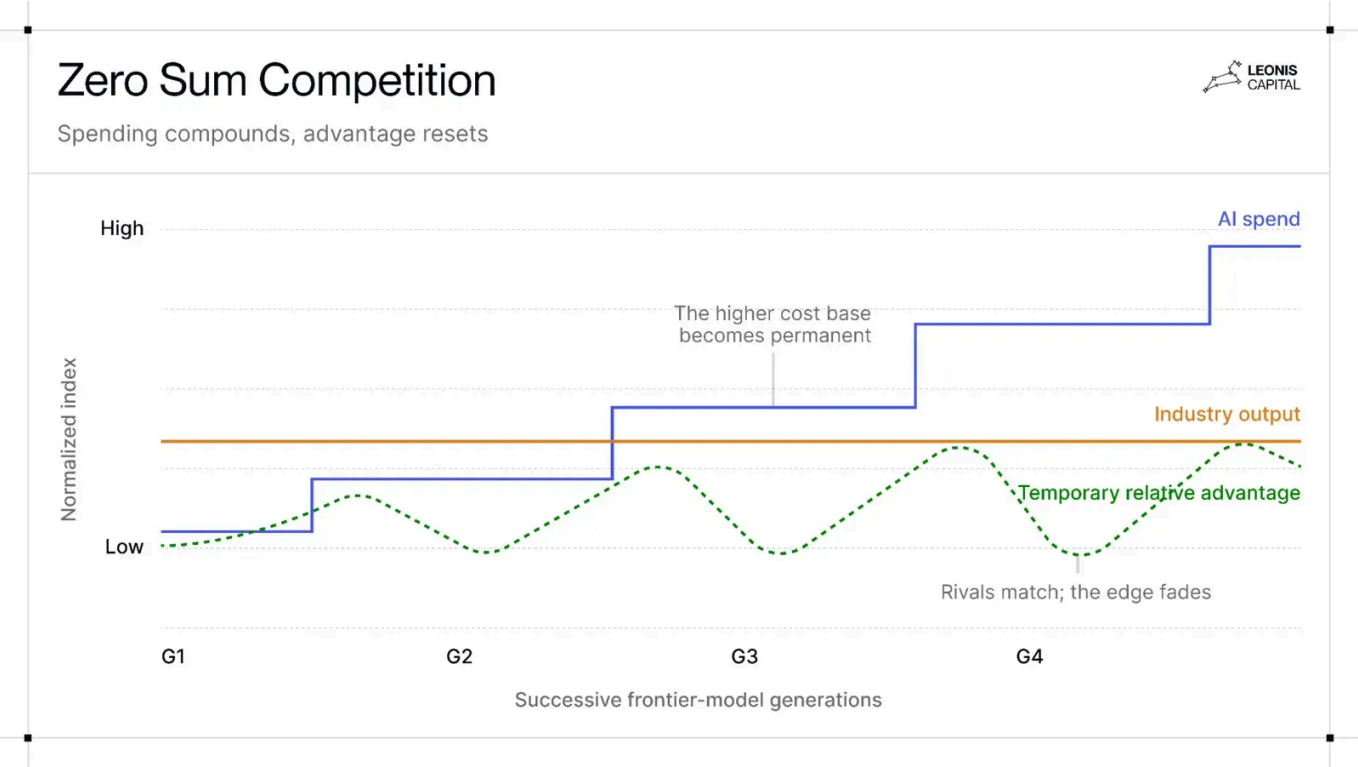
Why Someone Will Always Pay for Frontier Intelligence

If open-source is closing the capability gap, why do frontier models still command a premium?

Because in expansion markets, the value of one more unit of intelligence never floors. Intelligence is the binding constraint, so any incremental capability translates directly to output, and buyers will pay for the best available model regardless of what open-source can do at the 95th percentile. A hedge fund that spends more on frontier models than the total revenue of a mid-sized software company is not being reckless. It is just buying the input that produces its return. For example, Bridgewater's AIA Macro Fund is an AI-run hedge fund. An AI agent makes the trades, humans oversee. Launched in late 2023 with roughly $2B in capital, the fund had grown to about $4.5B in AUM by mid-2026 with an 11.3% annualized return, matching the firm's flagship human-run Pure Alpha fund in the first half of 2026. CEO Nir Bar Dea has said the AI fund generates "unique alpha uncorrelated to what our humans do." When you can attribute alpha directly to intelligence, model quality is the only variable that matters and the token bill is not a consideration.

Expansion buyers do not, however, buy in isolation. They compete with each other, and the shape of that competition determines how durable the labs' revenue actually is. Two forms of it look identical on an income statement and could not be more different underneath.

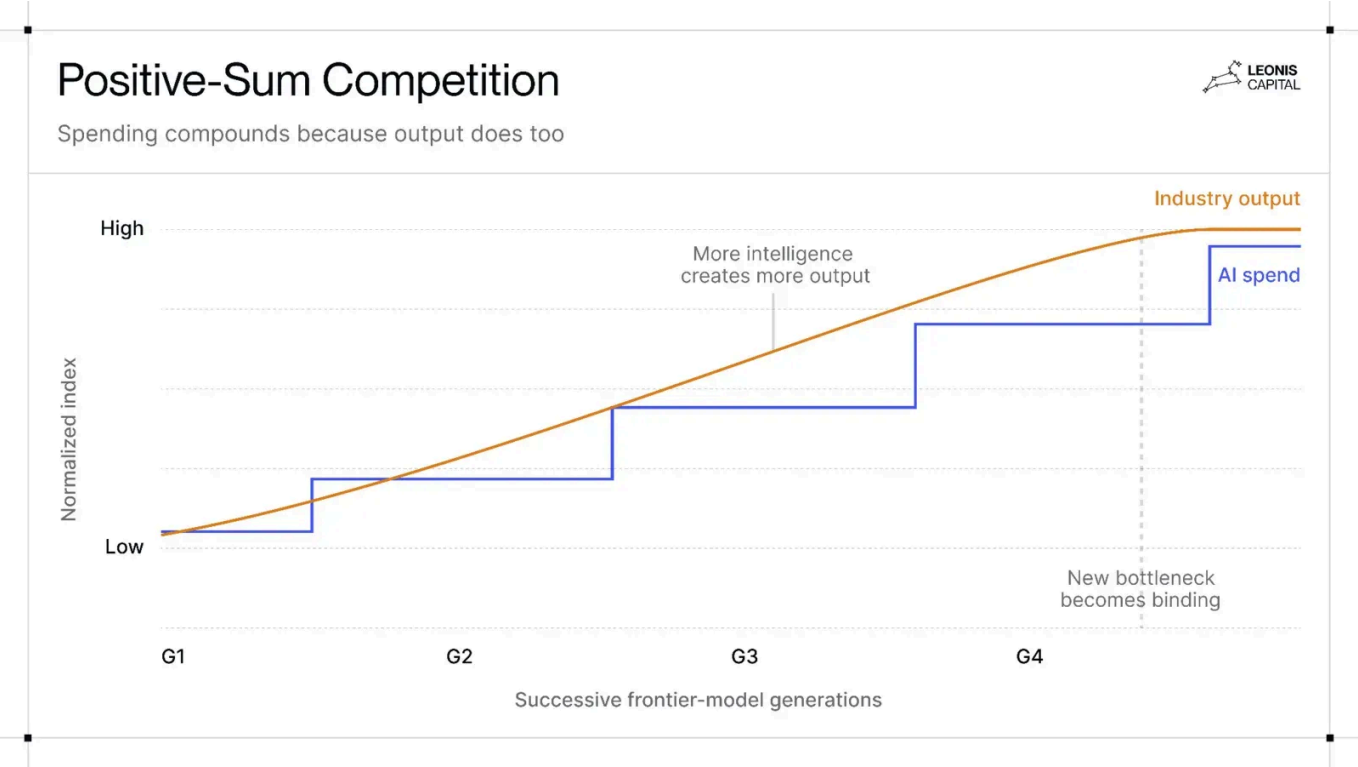
Zero-sum competition: In some expansion markets, better intelligence redistributes a fixed pie. Alpha in hedge funds is the pure case, one fund's excess return is another's underperformance. When one fund adopts frontier models, competitors must match or bleed. Once everyone has matched, the edge nets to zero, the pie has not grown, and the industry's cost base has permanently risen. Same aggregate alpha, minus millions per fund per year, paid to the labs. The consulting version is quieter but identical, if McKinsey adopts the best model, BCG has to, and the total consulting budget does not move by a dollar. The spend is a tax competitors levy on each other, and the model provider collects it.



AI spend ratchets upward with every frontier generation, while competitive advantage resets and industry output remains unchanged.

We have seen this exact dynamic before in high-frequency trading. Through the late 2000s and 2010s, funds poured money into microwave towers, co-location, and dedicated fiber to shave microseconds off their order routing. In 2010, Spread Networks completed an 827-mile fiber-optic cable through the Allegheny Mountains between Chicago and northern New Jersey at a cost of roughly $300M, cutting the round-trip transit time from 17 to 13 milliseconds. Every fund had to match or fall behind, and once they all had, the collective edge was gone and only the bill remained. The durable winners were never the funds. They were the tower operators and the exchanges. And even that edge decayed. Once microwave links made fiber routes obsolete, Spread Networks was sold to Zayo Group for $131M, less than half what it cost to build.

Positive-sum competition: In other expansion markets, better intelligence grows the pie. Drug discovery is the clearest case. If every biotech's hit rate doubles, more drugs exist. Competitors racing each other produce more total cures and more total revenue. The GLP-1 class posted roughly $132B in sales in 2025, up more than 30% year over year, a prize large enough that the R&D and compute needed to chase the next such molecule is rounding error against it. Chip design, materials science, and most software creation work the same way. A better codebase does not (directly) degrade a competitor's codebase. The spend still forces every participant to the frontier, but it compounds into new output. It looks more like investment than tax.



The flywheel compounds: each frontier generation raises AI spend because it also expands industry output, until a new non-intelligence bottleneck becomes binding.

Both look identical on a lab's revenue line today. Both kinds of customers are price-insensitive and both keep renewing, but they age in opposite directions. **Zero-sum spend is extraordinarily sticky as long as the competition runs**, because no participant can stop unilaterally, which is why adtech spending has compounded for two decades even though every advertiser resents it. Zero-sum spend is also fragile to structural breaks. When latency stopped mattering in high frequency trading, exchanges added speed bumps and the race ended. Consolidation, regulation, or a coordinated truce can do the same to any zero-sum AI race. Positive-sum spend is durable for a different reason and ends on a different death. Nothing external can negotiate it away, **but it eventually hits a non-intelligence bottleneck.** Drug-discovery spend climbs until it hits a wall in trial capacity or regulatory barriers, then plateaus no matter how good the models get.

The strategic implication for the frontier labs is that the mix matters more than the total. Two labs with identical premium revenue can own very different businesses if one is monetizing mostly zero-sum races and the other is monetizing positive-sum ones. Nobody has quantified that mix. It is the single number in AI we would most like to see, and we suspect the labs have not computed it either.

Expansion-market pricing also evolves differently. When the buyer's return dwarfs the token bill and the outcome is measurable, **some form of outcome based pricing eventually replaces**

usage pricing: revenue share on drugs discovered, gain share on trading returns, professional-services-style engagements built around results rather than input costs. Not every expansion market gets there because the outcome has to be attributable to the model, but where it does, pricing looks less like infrastructure and more like consulting or services. That is a very different revenue profile from the per-token efficiency market, and it should command very different multiples in the application layer serving each side.

The Question Behind Every AI Dollar

Bring both paradoxes back together and the whole discussion around AI spending looks different.

The market has been asking whether AI is worth what companies are paying for it. That is the wrong question, because it averages two markets that shouldn't be averaged. The right question, for every AI dollar, is **whether intelligence is the bottleneck on what that dollar is trying to produce**. If it is, the frontier is cheap at almost any price. If it isn't, the cheapest adequate model is the only rational choice, and the buyer who pays more is just overpaying.

That reframe is uncomfortable for most participants in the discourse.

It is uncomfortable for the frontier labs, because it says a meaningful portion of their current revenue base is on borrowed time as open-source captures the efficiency workloads underneath it, and their durable business depends on migrating faster into expansion markets than open-source can encroach into their base. It is also uncomfortable for the open-source labs, because it says their strategic prize is the largest but least monetizable segment, and their economics depend on capturing efficiency-market volume at commodity margins rather than on beating closed models on capability.

It is uncomfortable for the enterprises, because it says most of them are efficiency buyers, and any AI line item that cannot be tied to a real bottleneck is theater dressed as strategy. And it is uncomfortable for the discourse itself, because the exciting story of AI transforming every function of every company **was always going to be true in a few places and mostly false in many more.**

The interesting question, for the small share of the market where intelligence really is the bottleneck, has almost nothing to do with cost. It has to do with what becomes possible when the constraint on human output is finally something other than the intelligence available to apply to it. Most of the AI market will be about cost reduction, and that is fine, most markets are. But the piece worth watching is where the value of more intelligence keeps climbing. That is where the frontier lives, and where it justifies itself.