

The Price of Intelligence

A price-times-quantity read on the AI model market, and one you can track

Thematic Research | AI Model Economics

Jordan Rohan, Clearmeadow Partners

July 27, 2026

Companion to: Two Regimes, One Repricing

BOTTOM LINE

Tokens are commoditizing quickly, with less intense workloads finding cheaper token sources that appear capable. The implication is not a uniform re-rating of the AI leaders but a more precise lens on their valuations, one in which strategic moat and margin durability, rather than headline growth, are the deciding factors. This note supplies part of that lens, a realized price per token by lab that we carry as a live series rather than a one-off.

WHAT THIS MEANS FOR INVESTORS



1 In healthy markets, open-weight models will gain share. Now we can track it.

The open and open-weight tier has been taking token volume steadily, with Moonshot's Kimi only the latest to reach global recognition. The direction of travel on volume is clear, and it points toward the cheaper end of the stack. Their far lower training and inference costs also give these models an exponentially higher return on capital on any revenue they do capture.



2 The volume mix flipped to open, but revenue is still 90%+ premium.

A quarter ago proprietary models carried about sixty percent of token volume against forty for open weights; today that ratio has reversed. Buyers are routing down the price ladder, which is what pulls the blended token price down even as the premium labs hold their own rates. It also opens the door for a Moonshot or DeepSeek to jump straight into the premium tier, since the proprietary-versus-open distinction means little to token buyers, who care only whether a model solves their hardest problems and runs their most complex routinized operations with sophisticated agents, not how it came to be.



3 For investors, this matters.

The premier labs hold a commanding share of dollars today, but frontier capability and low-cost open-weight rivals trained by distillation are likely to approach parity over time. It is the innovator's dilemma, expressed in tokens.



4 Coding tokens are the exception, and they favor Anthropic, at least for now.

Tokens spent on code demand higher standards and are worth more to enterprise accounts, so that premium is stickier than the rest. We believe this advantages Anthropic relative to OpenAI.

I. Price times quantity

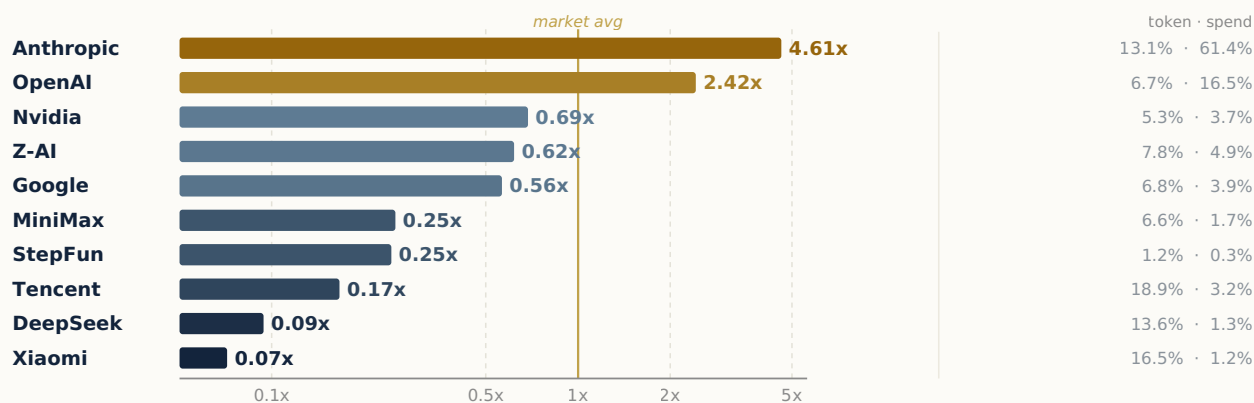
The unit is the token. Because every model is metered per token, the market has for the first time a common denominator: intelligence priced by the million tokens. From public gateway data we observe two shares for each lab, its share of tokens consumed and its estimated share of dollars spent. Their ratio is a realized price per token, and by setting the market average to one, it places every lab on a single scale.

What it is, and is not. This is not a benchmark of capability and it is not company revenue. It is what the open market actually pays per unit, shown as a multiple of the market average. What matters is not the level on any single day but that it can be measured the same way over time. Price times quantity is the oldest idea in markets; for AI it has only recently become possible to measure, and the pieces have sat in separate places, token share at OpenRouter, list prices at Artificial Analysis and Epoch AI, and combined spend at codesota and Tidelines.

Exhibit 1. Price per unit of intelligence

realized price per token, market average = 1.00x, log scale

Each lab's share of dollars divided by its share of tokens gives a realized price per token, indexed to the volume-weighted market average. It is among the first standardized price-times-quantity reads of the AI market: price per unit of intelligence multiplied by units consumed, and one we carry as a live series rather than a one-off. Q is token share, \$ is estimated spend share.



A token is one unit, but not all intelligence is the same: a frontier token performs more work per unit, so part of the roughly 64-fold spread is real quality. The business-model overlay carries much of the rest, since enterprise API and subscription pricing sit above free-tier and subsidized open usage, so this measures how a lab charges as much as what its tokens are worth. Inputs are list prices on public-gateway traffic that leans toward developer and open-model use and misses first-party enterprise and consumer, so read levels as directional and the spread as indicative, not as revenue. The parts are not new, since token share has been public at OpenRouter since 2024 and list prices at Artificial Analysis and Epoch AI since 2023, with combined spend reads at codesota and Tidelines; the contribution here is the realized-price normalization and the analytical overlay. Overall token price is down about twelve percent quarter over quarter; a clean year-over-year is not yet available, since the live indices launched in December 2025 and January 2026, though list-price proxies put frontier output near ninety-four percent below its 2023 level. Source: Tidelines (OpenRouter, Vercel AI Gateway), snapshot as sampled; series trackable going forward.

II. Not all intelligence is the same

A token is one unit, but not one thing. A frontier token does more work than a commodity one, so part of the spread is real. The rest is business model: enterprise and subscription pricing sit above free-tier and subsidized open usage, so a lab's price per token reflects how it charges as much as what its output is worth.

The semiconductor parallel. Hold it against semiconductors. That one label runs from a bleeding-edge Nvidia accelerator at tens of thousands of dollars to a commodity analog part at pennies, and no analyst would value the two on one multiple because both are built on silicon. Tokens invite the same error: a frontier coding token and a commodity

open-weight token are both tokens, yet as far apart in price, margin, and durability as a data-center GPU and a jellybean analog chip. The word is shared; the economics are not.

Two tiers, two strategies. The premium tier, the leading profitable lab near four and a half times the market and the incumbent above two, earns on high-value enterprise and coding work the cheap models cannot yet do. The commodity floor, the open and Chinese models, prices at or below cost to win volume, which is why it moves enormous quantity for almost none of the dollars: on disclosed figures the entire Chinese floor together earns well under a billion dollars a year, roughly one percent of a single premium lab's revenue, while carrying the majority of tokens (Sacra; Hong Kong filings). The shift so far is in volume, not dollars, the premium tier still holds over ninety percent of revenue, and the money has not yet followed the tokens.

III. The open-source precedent

Open has beaten proprietary before. The software record is remarkably consistent. Open won the server operating system, where Linux runs almost all cloud instances and every top-five-hundred supercomputer; it won the developer stack, with Git, Kubernetes, Postgres, and Nginx the defaults; and it won machine-learning tooling, where PyTorch and TensorFlow displaced every proprietary framework. The substrate of this entire industry is already open.

But winning share is not capturing value. In almost every case the profit did not stay with the open artifact; it migrated to a complement or a proprietary layer on top. Linux commoditized the operating system and the money went to the clouds that ran it and the chips beneath, with Red Hat monetizing support. Open databases were monetized by managed services, never the free code. Android is the sharpest case: the open system won roughly seventy percent of smartphone volume, yet Apple's iOS keeps most of the industry's profit and Google monetizes Android through proprietary services, not the open base. Open won the volume; proprietary and the layer on top won the dollars.

The model layer looks the same. Open weights are likely to do to models what Linux did to the operating system: commoditize the base and push value to the two protected ends, the compute and memory below and the proprietary outcome, agent, and deployment layer above. Premium labs holding the dollars while the open floor carries the volume is the Apple-versus-Android split in tokens. To capture premium value rather than only supply volume, the record says an open lab must wrap the base in something proprietary, hosting, agents, or deep enterprise integration, which is exactly where Zhipu's roughly forty percent margins come from, local deployments rather than its near-zero-margin API. One difference favors the premium tier and not the model: software is nearly free to run, but an open-weight model still costs a fortune to serve, so even a dominant one keeps feeding the compute layer with every token.

The valuations do not price this. The stakes are large: the two Western leaders are marked near a trillion dollars each on the premise that a frontier model is a durable, high-margin franchise. The precedent says the opposite is the base rate, the commoditized layer rarely keeps the rent. If cheaper open models reach functional parity on the work that matters, which the next section argues is plausible, the standalone model-lab business could prove far less valuable than today's marks imply, with lasting value accruing to the compute beneath and the applications above. That is the risk the valuations do not appear to reflect.

IV. The swing factor: whether the floor climbs

The premium tier keeps its lead only if the tiers stay separate. The one thing that would break it is an open-weight model climbing into the premium tier, and it bears watching, because the premium moat is at bottom a fixed-cost moat, and that cost is falling fast.

Reaching the frontier is enormously expensive. A single frontier run has climbed from roughly five million dollars for GPT-3 in 2020 to sixty to one hundred million for GPT-4 in 2023 and an estimated five hundred million for the 2025 frontier (Epoch AI, Stanford AI Index), and Anthropic alone plans about nineteen billion on training and inference this year. The two Western leaders have raised on the order of three hundred billion between them and will not be profitable for years, OpenAI with cumulative losses near forty billion through 2028. That is the price of finding frontier capability and holding it, most of it spent before a commercial product made sense.

Copying it is nearly free. A fast-follower pays almost none of it: DeepSeek reached GPT-4 and Claude parity on a final run reported near five and a half million dollars, its reasoning layer adding about three hundred thousand more, a few percent of one frontier run and a rounding error against the leaders' cumulative build. The mechanism is distillation, and it is deliberate: the industry builds models that learn from other models' outputs, so the leaders pay to find the capability and the followers copy it for cents on the dollar.

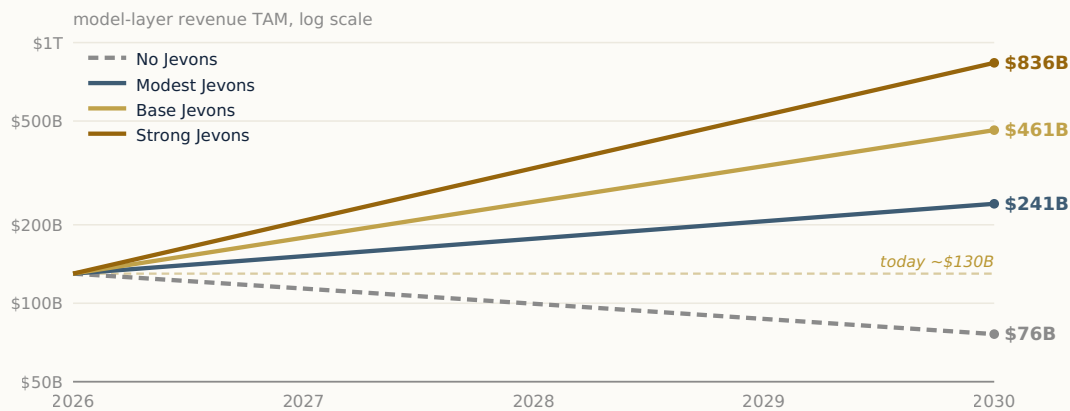
That makes the swing factor a return-on-capital story. A competitor that builds for single-digit millions but charges premium prices earns a return on capital no incumbent can match, and it need only be good enough at the high-value work to charge for it. That is the bet behind the open names now raising at tens of billions: if Moonshot, whose latest model has topped some Western frontier benchmarks and which is reportedly funded toward fifty billion dollars, turns a cheap-to-build, best-in-region model into the AI equivalent of Alibaba, the premium tier's share does not merely erode, it inverts. The real question is not whether tokens commoditize, which they are, but whether the labs that own the dollars can stay a tier above the labs that own the volume.

V. What it could mean by 2030

Unit price and total spending move in opposite directions: the list price of a token is down about twelve percent in a quarter and ninety-four percent since 2023, while total spending keeps climbing as usage explodes, the old pattern of a cheaper resource getting used far more. Anchor today's model-layer revenue near one hundred thirty billion dollars, Anthropic about sixty-nine billion per Yipit and OpenAI about thirty, run both forces to 2030, and the market's size then turns almost entirely on how strong that volume response is, from one that shrinks without it to one several times larger with it.

Exhibit 2. The dollar market to 2030

model-layer revenue, three Jevons cases

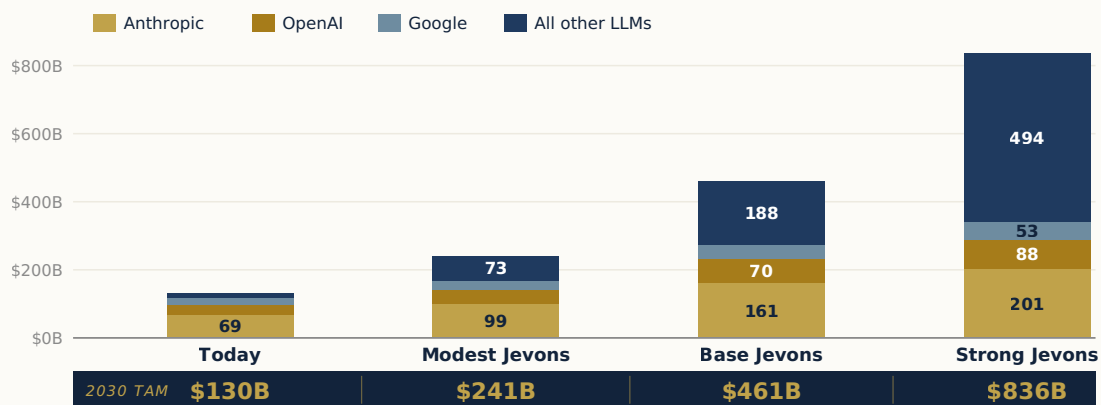


Illustrative, not a forecast. Today anchored at ~\$130bn model-layer revenue (Anthropic ~\$69bn per Yipit July 2026, OpenAI ~\$30bn per Sacra, Google and others the balance). The token mix skews toward open by 2030, more so under stronger Jevons (twenty-eighty to five-ninety-five), which pulls the blended price down; volume grows at zero, forty, seventy-five, and one hundred twenty percent a year. Market size is volume times blended price. Even modest volume growth outruns the price decline; only no growth shrinks the market.

But the premium labs capture a smaller share the stronger the boom, since the stronger the volume the more of it is cheap open-weight tokens, so every named lab grows in dollars while the open and commodity tier grows fastest and, under strong Jevons, becomes the single largest owner of the 2030 market.

Exhibit 3. Where the 2030 revenue lands

model-layer revenue by owner, \$bn



2030 model-layer revenue split by owner under three Jevons cases, with today for reference. Anthropic, OpenAI, and Google are held at their current shares within the premium tier; all other LLMs is the open and commodity floor. Every premium name grows in absolute dollars as Jevons strengthens, but the all-other slice grows far faster, from about seventy billion under modest Jevons to nearly five hundred under strong, where it exceeds Anthropic and becomes the single largest owner. The strongest-volume world hands most of the new revenue to the cheapest tokens. Illustrative, tier-level, not company forecasts; anchored to Anthropic ~\$69bn (Yipit, July 2026) and OpenAI ~\$30bn (Sacra) today.

VI. Caveats, and how to track it

Four things to keep in mind. It uses published list prices, not the money labs actually collect, so it overstates revenue after enterprise discounts. It is built on public gateways, which lean heavily toward developer and open-model use and miss first-party enterprise and consumer traffic, along with usage bundled inside the big cloud providers. It mixes quality with how a lab chooses to charge, as above. And it is a snapshot of a fast-moving market. None of that undermines the measure; all of it argues for tracking it over time rather than trading a single reading.

How often to refresh it. The token and spend data update daily and weekly from public gateways, so the price-times-quantity read can be recomputed each month and carried as an ongoing series. That is how we intend to hold it: not as a one-time call, but as a gauge, the first that lets an investor watch the price of intelligence the way they watch any other price.

ANALYST CERTIFICATION. The analyst responsible for this report, Jordan Rohan, certifies that the views expressed accurately reflect his personal views about the subject securities and issuers, and that no part of his compensation was, is, or will be directly or indirectly related to the specific recommendations or views expressed herein.

This report is published by Clearmeadow Partners, LLC and is intended for institutional investors, including hedge funds and family offices. It is provided for informational purposes only and does not constitute investment, legal, tax, or accounting advice, nor an offer or solicitation to buy or sell any security. Price targets and scenario levels are estimates derived from stated valuation methodologies and are subject to change without notice. Certain current market prices referenced herein are indicative and should be confirmed against a primary market source prior to any transaction. Ratings and valuation levels for individual securities, including NVDA (NEUTRAL) and other names referenced, reflect published Clearmeadow research as of the dates indicated. Clearmeadow Partners, its principals, and clients may hold positions in the securities discussed. Private-company valuations, including any reference to a leading frontier-model developer or a private launch and orbital-compute issuer, are estimates and postulates and do not reflect any nonpublic information. Forward-looking statements involve risks and uncertainties, and actual results may differ materially. Past performance does not guarantee future results. This material may not be redistributed without the written consent of Clearmeadow Partners, LLC.