PitchBook
a Morningstar company

Institutional Research Group

Harrison Rolfes
Senior Research Analyst,
Late-Stage Company Research
harrison.rolfes@pitchbook.com

pbinstitutionalresearch@pitchbook.com

Published on July 20, 2026

Contents

LATE-STAGE COMPANY RESEARCH

Anthropic Is Winning on Value, Not Price

Frontier AI's token trap, where the priciest tokens do the cheapest work

Key takeaways

- **The best model for the money is Claude Opus 4.8, priced near the top of the sheet.** At a fixed quality bar, it finishes a representative enterprise agentic task for about $2.56, while the cheapest model finishes the same task for about $5.83, more than twice the completed work per dollar.
- **The price war is a trap.** A cheaper model that fails more often costs more per completed task once human time spent per failure is counted, so every sticker cut can raise the true cost of the work. The buyers celebrating the cuts pay for the failures.
- **This is a segmentation event dressed as a price war.** Meta prices Muse Spark near or below cost to commoditize a market it need not profit from—something only an ad-funded player like Meta can afford to do. OpenAI and Anthropic defend premium tiers. xAI is a forced participant.
- **Grok has the loudest capability and the weakest position.** xAI's mid-July Grok 4.5 refresh, which it says adds native computer use, multimodal input, and a real-time X and Starlink feed no rival can match, makes Grok the best value on paper. It is still the worst place to build, on a segment losing more than $6 billion a year, with no enterprise surface to sell into.
- **The war is subsidized, not won.** Floor prices sit below serving cost, and the roughly $3.5 trillion private-credit market funds the gap, deferring a solvency question straight into Anthropic's October S-1.

The technology behind a token price

Every claim here is measured by one unit: a model reads and writes in tokens, each roughly three to four characters, and a provider charges separately for input tokens (the prompt and context) and output tokens (the generated text), quoted per million tokens (/MTok), split input/output.

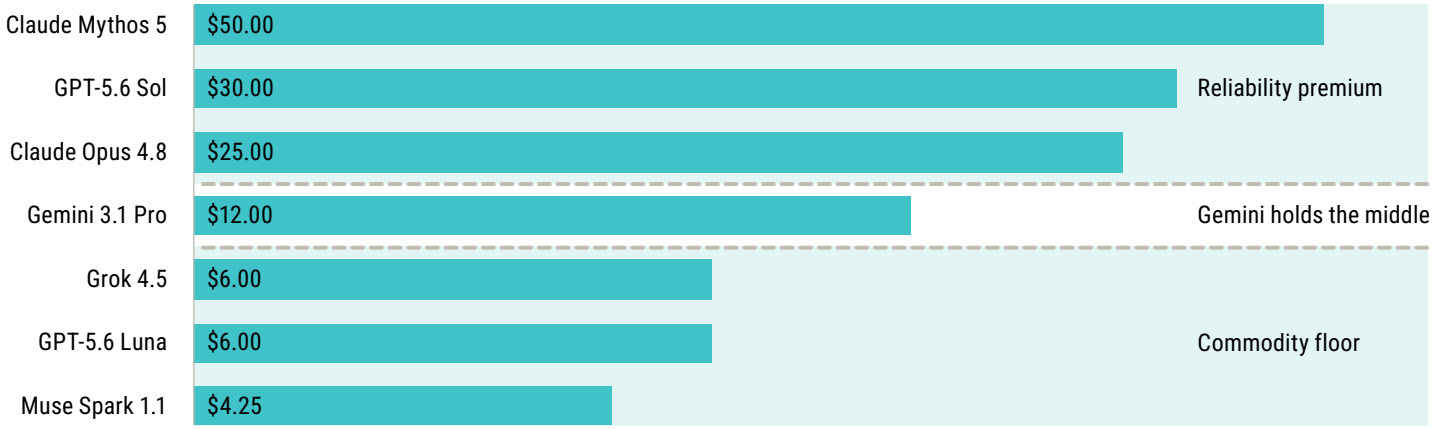
Output is priced far above input because the gap is physics rather than a margin choice. Opus charges 5x more to write than to read, Sol and Luna 6x more. Reading a prompt happens in one parallel pass, the prefill; writing happens one token at a time, and each token demands a full forward pass that reads the entire model and the growing conversation memory out of high-bandwidth memory. Generation is, therefore, sequential and bound by memory bandwidth, which is expensive and serial, whereas reading is cheap.[1] The binding constraint on output cost is the same scarce high-bandwidth memory the build-out is fighting over, so the price sheet and the compute thesis are one story.

Two consequences follow. First, utilization sets the real cost. Providers batch requests so that one expensive read of the weights serves many; a price cut not offset by better hardware or utilization comes out of gross margin, so a lower number is either better economics or a subsidy, which is indistinguishable on a sheet. A long context is not free either, because conversation memory grows with it in that same scarce memory, so Muse Spark's million-token window is costly to use at scale.

Second, the step that reframes the competition is that enterprise work is agentic. An agent plans, calls a tool, reads the result, replans, and repeats dozens of times, each pass re-reading a growing context; consumption multiplies every time it fails and retries. The per-token price, therefore, says almost nothing about what a task costs, because the loop and its failure rate dominate the sticker.

Two crowded ends, one frontier model holding the middle

AI model output list cost/MTok



Sources: PitchBook, Claude, OpenAI, xAI, Google, and Meta • Geography: US • As of July 17, 2026

The unit of account: Value per token

Buyers do not want cheap tokens, but rather the most work completed per dollar. The correct unit is:

> *Cost per completed task = (tokens in x price in + tokens out x price out) x expected attempts to reach success at a fixed quality bar*

A model that succeeds with probability p needs on average 1/p attempts, so one at a quarter of the price that needs three tries is not cheaper. That estimate treats retries as independent; in production, failures are correlated, because a model that misreads a task tends to miss it again, which means the geometric figure flatters the cheap tier, and the premium's true edge is wider than it is in the table below. Our analysis looks at a representative agentic task of 60,000 input and 12,000 output tokens per attempt.

Top AI model cost per completed task

Model	Input cost/MTok	Output cost/MTok	Cost per attempt	Cost per task p=90%	Cost per task p=75%	Cost per task p=65%
Claude Mythos 5	$10.00	$50.00	$1.20	$1.33	$1.60	$1.85
Claude Opus 4.8	$5.00	$25.00	$0.60	$0.67	$0.80	$0.92
Gemini 3.1 Pro	$2.00	$12.00	$0.26	$0.29	$0.35	$0.41
GPT-5.6 Sol	$5.00	$30.00	$0.66	$0.73	$0.88	$1.02
GPT-5.6 Luna	$1.00	$6.00	$0.13	$0.15	$0.18	$0.20
Grok 4.5	$2.00	$6.00	$0.19	$0.21	$0.26	$0.30
Muse Spark 1.1	$1.25	$4.25	$0.13	$0.14	$0.17	$0.19

Sources: PitchBook, Claude, OpenAI, xAI, Google, and Meta • Geography: US • As of July 17, 2026

Note: Prices are published list prices, including for Claude Mythos 5 and Google's frontier Gemini 3.1 Pro, while per-task figures represent PitchBook analysis. Per-attempt token counts are fixed at 60,000 input and 12,000 output. The cost-per-task columns are token-only (retries only) and exclude the $17 failure cost that the body text and the cost-per-completed-task chart apply.

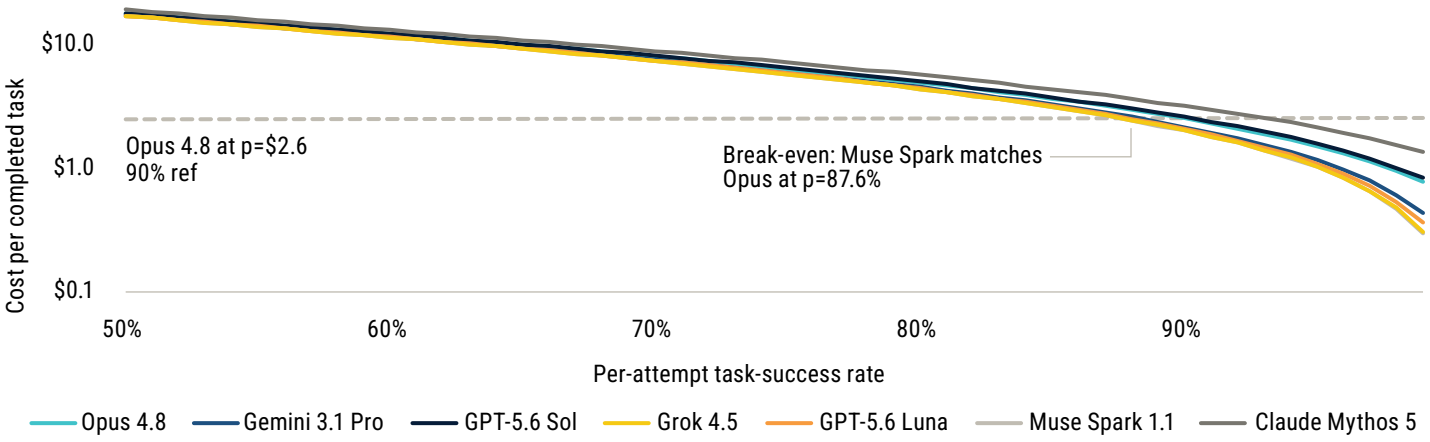
On tokens alone, the cheap tier wins almost mechanically, because Opus at 60 cents an attempt only beats Muse Spark at about 13 cents if it succeeds nearly 5x as often, and no frontier gap is that wide. That is exactly where the token-only view fails, because a failed agentic attempt is not free. It burns reviewer time, breaks the automation downstream, and erodes trust in the seat. The cost of a failure is not one number. It is near zero for failure-tolerant work like drafting and internal search, and hundreds of dollars for a wrong answer that reaches a customer or a regulator. That spread is not noise and is the reason the market segments. Assign a conservative $17 to each failed attempt, about 10 minutes of a fully loaded operator, and the ranking inverts. Opus succeeding nine out of 10 times finishes the task for about $2.56. Muse Spark succeeding three out of four times finishes it for about $5.83, more than

twice as much despite tokens being at a quarter of the price. The cheapest model on the sheet is the most expensive to run, and the near-priciest is the cheapest. That inversion is the thesis of this note. It is also why a lower sticker price often results in a larger bill. A cheaper token invites longer contexts and more agent loops, and the failures it hides are billed to the buyer in labor, not tokens.[2] The same arithmetic reaches the top of the sheet: Mythos 5, at twice Opus' price, still runs near cheapest, the priciest sticker on the sheet finishing among the cheapest to run. It also faces Gemini 3.1 Pro as a challenger. Gemini is a frontier model at $2 and $12, under half Opus' per-attempt cost, and it finishes the reference task for about $2.62—assuming Gemini succeeds about 88% of the time versus Opus' 90%—against Opus' $2.56. Opus keeps the crown by a hair, and only because we assume Anthropic is more reliable. If Google matches it, Gemini is the cheapest way to finish real work.

The conclusion does not rest on the $17 figure. Solve for the failure cost at which the premium breaks even: At a 15-point reliability edge, it pays once a failure costs more than about $2.24, at 10 points more than $3.67, and at five points more than $7.93. Only if failure is free does the cheap tier win, and in production it never is. The premium is defensible whenever mistakes are expensive, which is the reliability-gated work where most enterprise value sits, so value per token belongs to whoever pairs the highest reliability with a defensible price, not the lowest number.

Below roughly 85% success, token price is noise; above it, reliability is the price

AI model cost per completed task by per-attempt task-success rate

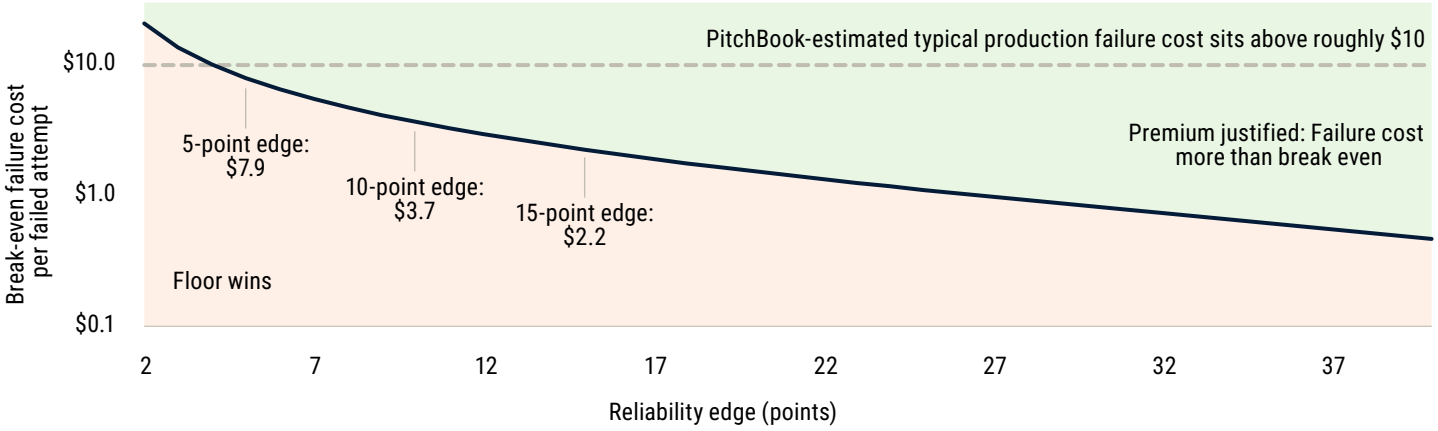


Sources: PitchBook, Claude, OpenAI, xAI, Google, and Meta • Geography: US • As of July 17, 2026

Note: Fully loaded cost per completed task: (cost per attempt + $17 x (1-p))/p. Workload is fixed at 60,000 input and 12,000 output tokens per attempt. The cost-per-task axis uses a log scale.

Past a modest reliability edge, almost any failure cost justifies the premium

Agentic AI task break-even failure cost for reliability premium by reliability edge



Sources: PitchBook, Claude, OpenAI, xAI, Google, and Meta • Geography: US • As of July 17, 2026

Note: Break-even failure cost F = (p_o x c_m - p_m x c_o) / (p_m - p_o), Opus success p_o = 90%. Above the curve, the reliability premium is cheaper per completed task. Prices are vendor list prices; workload and cost inputs are assumptions. The break-even cost axis uses a log scale.*

If independent, nonvendor evaluations that measure cost per completed task rather than price per token show the cheap tier matching premium success rates within two points at scale, the analysis changes.

Margin mechanics: Who can subsidize, and is this a war?

A 4x to 5x output cut destroys model-layer margin. We estimate a frontier model serves at roughly $6 to $8 per million output tokens. At $25, that is about a 70% gross margin. At the $4 to $6 floor, it is negative, roughly -17% at $6 and -65% at $4.25. A 4x cut in one generation outruns the 2x to 3x serving cost decline each generation delivers,[3] so the gap is subsidy, not efficiency, and who can pay it decides the war.

Meta is a structural price warrior because it does not need a model margin. Pricing Muse Spark near or below cost drives the price of the complement toward zero while Meta earns its return elsewhere—on the surface, the ad load, and its own silicon. That is durable because an ad engine funds it rather than outside capital. Commoditizing a market you need not profit from is not winning it.

xAI is the capability story that does not fix the economics. The mid-July Grok 4.5 refresh, which the company says adds native computer use, multimodal input, and a real-time feed from X and Starlink that no rival can replicate, is a genuine jump that makes Grok the value leader on paper.[4] However, it changes nothing. The AI segment turns about $3.2 billion of annual recurring revenue (ARR) against a fiscal year 2025 operating loss we estimate north of $6 billion. xAI is now a SpaceX operating subsidiary following the $250 billion February acquisition, so the losses ride the SpaceX and Starlink cross-flow, a balance-sheet decision rather than a business model. A better, cheaper model does not close a $6 billion loss.

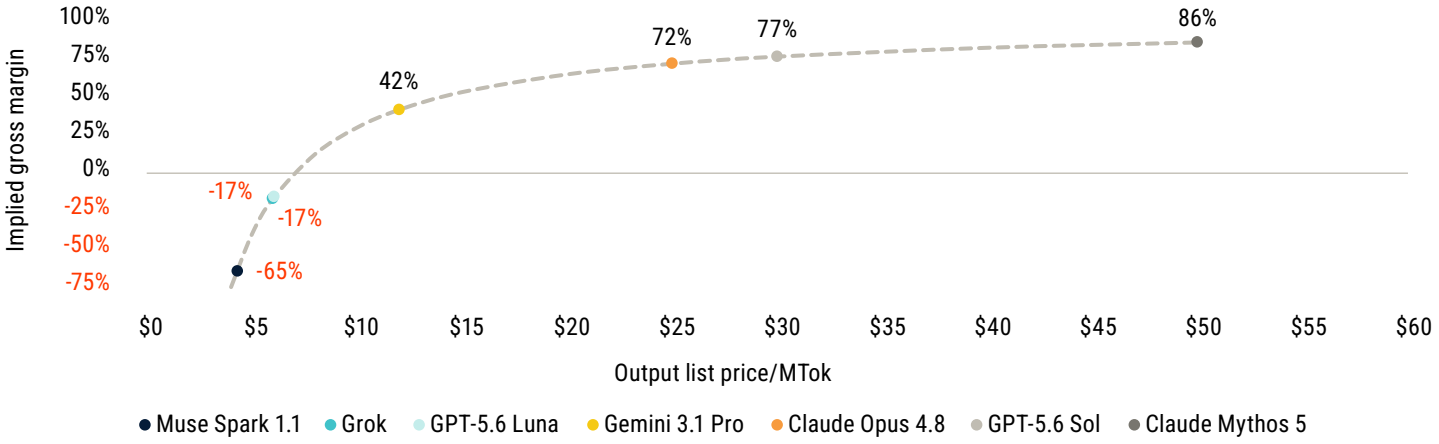
OpenAI is running segmentation, not a race to zero, and the tell is that Sol lists above Opus on output. Luna harvests price-sensitive volume while Sol defends the premium, and with roughly $30 billion in gross run-rate revenue, about 900 million weekly users, and Codex above 1.5 million, OpenAI needs the premium tier because its rich multiple and thin capital efficiency cannot ride a $6 output.

Anthropic defends a two-tier premium and declines to chase. Opus holds at $5 and $25, and Claude Mythos 5 sits above it as a limited-availability flagship priced at $10 and $50, exactly twice Opus on output. Anthropic carries a $965 billion valuation on $47 billion of gross run-rate revenue, the best capital efficiency among the labs, and a reported first quarterly operating profit near $559 million, though that figure excludes stock compensation, which flatters the steady state. Its strategy is now to hold the list price, extend the premium upward, and prove enterprises keep buying it.

Google is the player this war should fear most, because it can fight on both fronts at once. Gemini 3.1 Pro is a frontier model at $2 and $12, funded like Meta's Muse from an advertising engine, distributed like ChatGPT through Search, Workspace, Android, and Cloud, and served on Google's own Tensor Processing Units (TPUs), which give it the best cost structure here with no third-party silicon tax. The tell is that Anthropic's own $34.5 billion June chip-bond raise is earmarked for Alphabet-developed TPUs, so even the premium defender buys its silicon from Google. Meta commoditizes to sell ads but has no frontier model; Google can commoditize and still field one, a stronger hand than any pure premium defender or price warrior alone.

A few dollars separate a money-losing model from an 86%-margin model

Implied agentic AI model-layer gross margin by output list price



Sources: PitchBook, Claude, OpenAI, xAI, Google, and Meta • Geography: US • As of July 17, 2026

Note: Implied model-layer gross margin = 1 - serving cost/output list price, using the $7 midpoint of the $6 to $8 serving band. The price axis uses a log scale.

This is a segmentation play at the top and a real price war only at the floor. Two of the four defend premium tiers. Anthropic pushes the premium upward with Mythos, Meta commoditizes a complement, and xAI follows price without the economics. On our estimate, a slight majority of enterprise agentic dollars flows to reliability-gated work

where the premium is defensible, and the rest to failure-tolerant work that carries most of the volume. The two ends split volume and value in opposite directions, and the value end is Anthropic's.

If Anthropic or OpenAI cut flagship pricing by 15% or more before August 31, 2026, the segmentation read fails, and this becomes a price war. Our analysis shows a 70% probability that Anthropic holds its price.

The barbell and the squeezed middle

Repricing events rarely kill the top or bottom of a market. They hollow out the middle. The vulnerable layer is undifferentiated mid-tier access, neither the cheapest useful inference nor the most reliable premium surface. It reprices to commodity within quarters, because Luna and Muse Spark now define good enough at the floor, while Opus and Mythos set a reliability bar that the middle cannot clear. Capital and revenue drain to the ends. Gemini 3.1 Pro complicates this because it sits in the middle that the note calls hollow, but it is a differentiated frontier model, not the undifferentiated mid-tier that dies. A frontier model at a mid-price is not the squeezed middle but a strong value position on the sheet.

Model resellers and gateways earned the spread between list price and enterprise inertia, and that spread has gone negative now that the floor undercuts their markup. Legacy vertical software sold a workflow moat by the seat, and that umbrella collapses the moment an agent that finishes the same task for a few dollars runs the workflow itself. The commodity model tier and the seat-priced software middle are one trade—capacity charging more than it costs to replace.

Here is the test: If a well-known reseller or gateway reprices to a thin usage-based margin, shuts down, or sells below its last private mark by the end of this year, the squeeze is confirmed.

Market implications

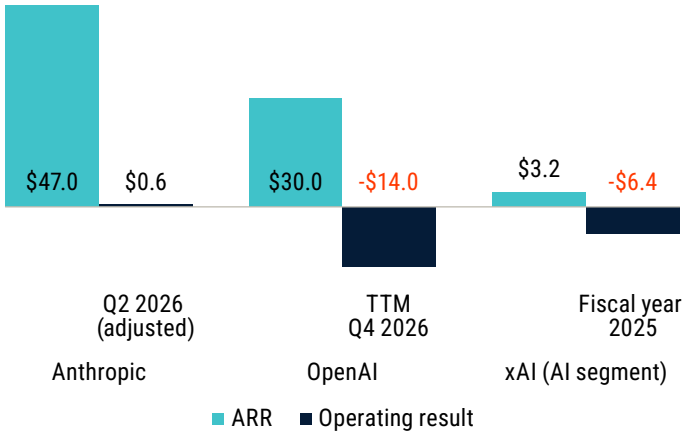
Cheaper inference is accretive to buyers and the application layer and dilutive to the model layer, because every dollar cut from output pricing lands with whoever owns the workflow, not the lab that cut it. A valuation dislocation follows: Anthropic trades near 20.5x revenue, $965 billion on $47 billion of gross run-rate, against OpenAI's 28.4x, $852 billion on $30 billion of gross run-rate, so the market pays more for the weaker position. If the thesis survives the S-1, that gap is the trade.

The build-out is insulated from the price war. High-bandwidth memory, wafer fabrication equipment, datacenter power, cooling, and permitted baseload are priced on capacity scarcity, not model margin, the same memory-bandwidth wall that sets output pricing. A model-layer price war does not cancel compute demand at the volumes the cheap tier chases; it raises it. The complex faces a financing shock rather than a pricing one.

Model-layer economics versus build-out economics

The model layer can already pay its own way

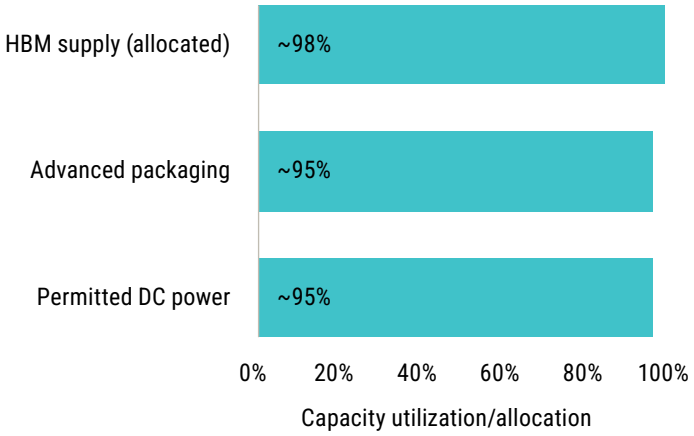
AI lab ARR versus latest operating result ($B)



Sources: PitchBook, company disclosures • Geography: US • As of July 17, 2026

The build-out underneath runs near full capacity

Indicative capacity utilization/allocation by constraint



Source: PitchBook estimates • Geography: US • As of July 17, 2026

Financing makes the war possible. The roughly $3.5 trillion private-credit market is the marginal source of AI capital, and labs price below serving cost because debt funds the gap. Anthropic's $34.5 billion chip-bond stack, earmarked for Alphabet-developed TPUs, is the template,[5] and xAI's cross-subsidy, now folded inside SpaceX, is the same in a corporate wrapper. A war paid for by credits ends when the credit reprices, not when a margin breaks, so the covenant is the market's appetite for negative-margin token sales, tying the price war to the coming IPO.

The IPO is the pressure test. OpenAI's September filing front-runs it, but Anthropic's October S-1 is the one that forces open what private markets have never seen: model-layer gross margin, mix, and any enterprise attach. Filing at a $965 billion mark while a rival prices comparable-claim inference at a quarter of the rate is the hardest setup and cleanest validation at once. If the price holds and attaches through the IPO window, the best-value thesis is proven. The gross margin page, as opposed to any benchmark, is the filing's most repriced disclosure.

If an AI-exposed private-credit vehicle discloses markdowns or tighter terms on chip-backed paper by the start of 2027, the financing-fragility read is confirmed. A second large chip-bond stack pricing at or inside June's terms would refute it.

Consumer implications

Consumers almost never pay per token. They pay for a flat subscription near $20 a month, attention and data on an ad-funded free tier (Meta's model), or nothing visible when AI is folded into a subscription they hold. Cheaper inference makes the free tier good enough, so capability inflates there while willingness to pay compresses above it, mirroring the enterprise story. Enterprises pay for reliability they can measure, but consumers will not pay for quality they cannot perceive. Distribution decides this one, because a user will not switch assistants over a difference they cannot feel.

If a consumer provider materially raises free-tier capability or discloses paid conversion or average revenue per user (ARPU) under pressure by the end of November this year, the pass-through will be visible. We expect free-tier expansion first.

The surface decides

The war looks like a pricing table, but it is settled on the desktop and the phone. OpenAI's ChatGPT Work agent and Codex application, and Anthropic's Claude Cowork compete for the enterprise-agent seat, the permission-controlled surface through which agents reach into a company's systems; whoever owns it owns task routing. That is where the best-value argument becomes a moat. The surface owner routes silently, sending the high-stakes step to Opus or Mythos and the trivial one to the floor, capturing both ends of the barbell. Google is a third front, folding Gemini into Workspace and Search, but the enterprise-agent seat, the permissioned surface where high-stakes work runs, is still contested between OpenAI and Anthropic. This is why Grok's refresh does not win. A better cheap model with a unique data feed is still only a model, and xAI owns no seat to route it through and no cash to outlast the incumbents. Cheap tokens win benchmarks and spot volume, which churn on the next price sheet. Seat reliability, admin tooling, and audit trails cannot be reprocured quarterly. The seat, not the sticker, is the moat.

If the first disclosed enterprise-agent or attach metric arrives by the end of September 2026, the surface thesis becomes measurable. Trivial seat counts at both OpenAI and Anthropic would mean the surface war is earlier than we think.

The verdict: Best model, best value

Judge the labs on the qualities that compound over time, and the answer is not close.

AI lab competitive scorecard on pricing, model, moat, and unit economics

Lab	Pricing for value	Model	Moat	Unit economics	Verdict
Anthropic	Best value: cheapest per completed task	Best value: Opus 4.8; Best model: Mythos 5	Reliability, the Cowork seat, a defended premium	Best capital efficiency; first quarterly profit	Winner
OpenAI	Split: Sol premium, Luna at the floor	Strong and broad	Distribution: roughly 900 million weekly users	Rich multiple on thin efficiency	Closest rival
Google	Frontier value: Gemini 3.1 Pro at $2/$12	Strong frontier: Gemini Pro and Flash	Search, Workspace, Cloud, and its own TPUs	Best cost structure; ad and cloud funded	Dark horse
Meta	Cheapest sticker, dearest per completed task	Reported strong agent	Advertising engine and owned surfaces	No model margin needed	Owns the floor
xAI	Value leader on paper only	Grok 4.5 refresh; agentic, real-time data	Real-time X and Starlink data; no seat	Losing more than $6 billion per year	Broken economics

Sources: PitchBook, Claude, OpenAI, xAI, Google, and Meta • Geography: US • As of July 17, 2026
Note: The scorecard is PitchBook analysis; the Meta and xAI model claims are the vendor's own.

Anthropic owns both ends of the value frontier. The best model outright is Claude Mythos 5, the limited-availability flagship above Opus on capability and, at $10 and $50, exactly twice Opus' output price. At the top of the risk curve, where a mistake costs the most, its reliability edge earns that premium. Once the edge over Opus reaches about three points, Mythos is also the cheapest way to finish the work. The best value on the market is Claude Opus 4.8, which finishes a representative task for about $2.56 despite a near-top sticker, though Google's Gemini 3.1 Pro is now within a whisker. Either way, the company is Anthropic.

Is Opus actually best, or an artifact of our reliability assumption? Against Sol, the answer is airtight. Opus wins at every equal level of reliability because Sol prices output higher, so Sol takes the crown only by being materially more reliable, which nothing suggests. Against Google, it is not airtight. Gemini 3.1 Pro's tokens cost less than half of Opus' tokens, so at equal reliability, Gemini is cheaper. Opus holds the value crown only if Anthropic's reliability edge over Google clears about two points, a thinner and more contestable claim than the rest, because Google's frontier model is genuinely frontier. Against the failure-tolerant cheap tier, Opus still wins on a small edge. The controversial part is what it says about the market. The companies winning the price-cut headlines sell the worst value, and the value crown is now a photo finish between Anthropic and Google.

What would change the call

The verdict breaks if OpenAI turns its distribution lead into the enterprise seat before Anthropic secures it, or if Anthropic cuts its flagship prices and concedes the value argument it must sell. The segmentation read breaks if agents self-correct cheaply, collapsing the cost of failure and the premium with it, or if Meta or a capability-rich xAI builds a credible enterprise surface. The value call breaks if Google's reliability matches Anthropic's, which would hand Gemini 3.1 Pro the value crown outright and its TPU cost structure the durability to keep it. We are watching six dated markers: Anthropic's flagship pricing by August 31, the first enterprise-agent seat or attach metric by September 30, the first consumer free-tier or ARPU signal by November 30, chip-backed private-credit terms through 2027, whether Anthropic's $34.5 billion chip-bond converts into delivered Alphabet TPU capacity by December 31—which would confirm that Google's own silicon underwrites even its closest rival and hardens the cost-structure edge behind the value call—and above all, Anthropic's October S-1, whose gross margin page converts most of this note into established fact.

References

1: "Compare Input vs Output Token Pricing: Why Output Costs More and How to Budget For It," Amnic, July 2, 2026.

2: "LLM Inference Prices Have Fallen Rapidly but Unequally Across Tasks," Epoch AI, Ben Cottier, et al., March 12, 2025.

3: "The LLM Cost Paradox: How "Cheaper" AI Models Are Breaking Budgets," IKANGAI, Zoe Spark, August 21, 2025.

4: "Introducing Grok 4.5," xAI, July 16, 2026.

5: "Apollo Wraps Up $35 Billion Chip Deal for Anthropic," Bloomberg, Paula Seligson, et al., June 8, 2026.

PitchBook provides actionable insights across the global capital markets.

Additional research:

OpenAI: Waiting for $1 Trillion

Download the report here

Ranking the AI Giants: A New Framework for the Frontier Five

Download the report here

PitchBook Insights is an online compendium of in-depth data, news, analysis, and perspectives that shape the private capital markets.

PitchBook subscribers enjoy exclusive access to a comprehensive suite of private market insights, including proprietary research, news, data, tools, and more on the PitchBook Platform.

Nizar Tarhuni
Executive Vice President of Research and Market Intelligence

Paul Condra
Senior Director, Global Head of Private Markets Research

Report created by:

Harrison Rolfes
Senior Research Analyst

Jenna O'Malley
Senior Graphic Designer

Learn more about PitchBook's Institutional Research team.

Click here for PitchBook's report methodologies.