Anthropic 3Q26 Profit Over

B: The Anthropic IPO Financials Sneak Peak

Anthropic 3Q26 Profit Over

B: The Anthropic IPO Financials Sneak Peak Anthropic’s Opportunity is Theirs to Lose

JUL 8 https://newsletter.semianalysis.com/p/anthropic-3q26-profit-over-1b-the

When Dario Amodei left OpenAI to start Anthropic in early 2021, the viral release of ChatGPT was over 18 months away and the commercialization of LLMs was practically zero. Just a few short years later, Anthropic and OpenAI combine for ~$100B of ARR and a clear winner emerged in the profitable monetization of AI models in 2026 as Claude Code took the software development world by storm.

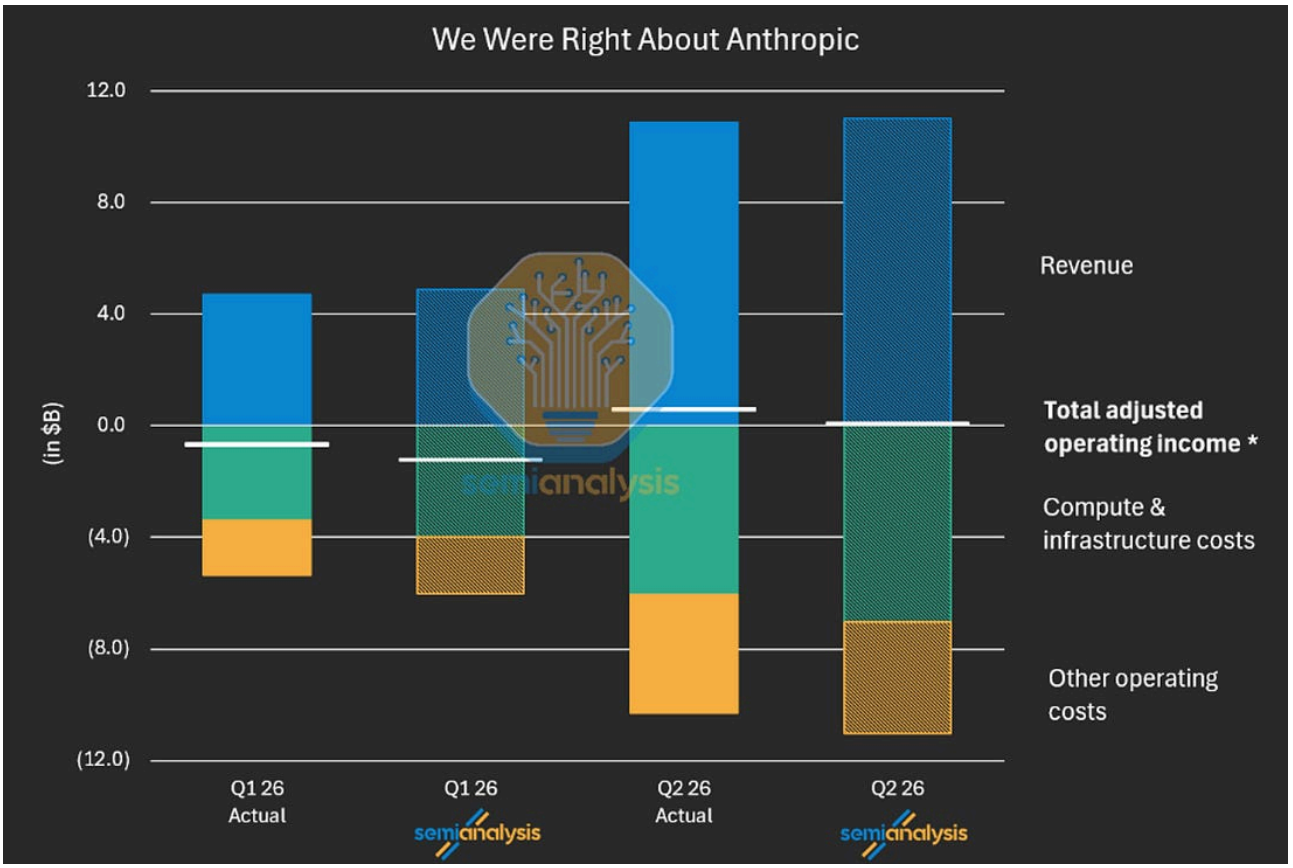
Anthropic confidentially filed for IPO on June 1st. Over 1 month later, equity raises from hyperscalers loom, and a reported OpenAI push out of their own IPO until 2027 has led some to question the ability of the labs to raise. However, Anthropic is the clear clubhouse leader in capturing the B2B market today and is doing so in a profitable manner against an unfocused and money-burning competitor.

QQ:1413221435

With this lead, we expect Anthropic to take advantage of their superior business model and margins to invest in further in new models that help extend their lead and monetization over closed and open source competitors. Anthropic has the ability to truly make OpenAI dance and we see Anthropic as the first $6T company as a base-case possibility if they continue to execute. Pricing power, gross margins, business model, and profitability are all reasons for Anthropic to IPO first and put the impetus on OpenAI to open their financials and raise the necessary capital to compete and fund the massive AI buildout still to come.

We’ve already seen 2 AI Labs IPO this year (Zhipu and Minimax from China), but Anthropic would be the first AI lab of this scale to do so. A confidential filing means there are no public numbers disclosed. Fortunately, the Tokenomics team at SemiAnalysis works to build the financials from the bottom-up by SKU, tier, and customer type. Recently, a WSJ article on Anthropic’s financials confirmed the accuracy of the work our Tokenomics team does across labs and hyperscalers to help investors, corporates, and other stakeholders understand the economics and financials of the AI Ecosystem.

QQ:1413221435



Source: WSJ, SemiAnalysis Tokenomics Model

QQ:1413221435

In this article, we'll dive deep into the work our team has done in the Tokenomics Model and walk you through the financial details of Anthropic, how that compares to OpenAI, where we see the market moving, and implications for the value chain and broader AI market. While other SemiAnalysis work focuses on the technical aspects of AI Labs, this piece will focus solely on the current and future financials, margin economics, and long-term outlook for Anthropic.

If you are interested in our coverage/model of Anthropic, OpenAI, Hyperscalers, and more please reach out to sales@semianalysis.com.

Background of the IPO

Anthropic's capital position is strong. With over $100B raised, a profitable business model (on a non-GAAP basis), and strong gross margins, why go public in the age of seemingly endless VC dollars and companies like Stripe and Databricks choosing to stay out of the public eye? We see three major reasons:

- Access to equity and debt markets (which will become a necessity)
- Audited and public financials help ease enterprises making large contract commitments
- Make the talent war easier by allowing employees to easily sell shares

QQ:1413221435

In the current world of talent deals that rival professional sports and massive and growing compute needs, having liquid public equity and the ability to raise debt and equity quickly makes strategic sense.

Anthropic Funding History			
Round	**Date**	**Amount Raised**	**Post-money Valuation**
Series A	May 2021	$124M	-
Series B	Apr 2022	$580M	-
Google	Feb 2023	$300M	-
Series C	Mar 2023	$300M	$4.1B
Amazon	Sep 2023	up to $4B (initial $1.25B)	-
Google	Oct 2023	up to $2B (initial $500M)	-
Series E	Mar 2025	$3.5B	$61.5B
Series F	Sep 2025	$13B	$183B
Series G	Feb 2026	$30B	$380B
Series H	May 2026	$65B	$965B

Source: SemiAnalysis

QQ:1413221435

We believe these proceeds will be used to fund an ever-growing supply/demand deficit between the compute Anthropic requires running inference while investing in new model training and the supply at which they can acquire it. Like the above market rate SpaceX deal, we expect Anthropic to purchase additional compute from trusted vendors such as Meta, who was recently rumored as of 7/1/26 to explore the business of renting compute to external parties. Our Tokenomics Model shows the stark difference between unconstrained demand for compute from Anthropic (based on Training spend as a % of revenue and Inference ARR per MW) and the supply-based method in our Datacenter Model which tracks the timeline and progress of data center builds.

In fact, we see over 100GW of compute demand between OpenAI and Anrthropic by the end of 2030. This will require over 90GW of net compute additions through 2030 vs 2.5GW added in 2025 and 5GW in 2026. Today, we estimate OpenAI and Anthropic have access to just over 6GW of compute.

During 1Q26, Anthropic users felt this compute shortage. Rate limits, downtime, and other throttling measures made using Claude difficult. As we've written in the past for our Tokenomics subscribers, Anthropic's gross margins in their API business are extremely strong. Access to open capital markets will help Anthropic exploit its superior margins and secure capital at favorable rates to bid up compute to levels rivals will struggle to match.

As we explain in detail in the Business Overview section, OpenAI currently doesn't have the same luxury.

IPO/Capital Raise Market

In June, Alphabet raised $84.75B of equity. Meta has been rumored to raise in the tens of billions range. As we discussed in our FCF to Zero note to institutional subscribers back in April, hyperscalers will not generate sufficient operating cash flow to pay for their capex needs in 2027 and beyond. This funding, along with the rest of the AI buildout, must come from somewhere. Regardless of how cheap it is, capital is always competed for, and supply is sharply and swiftly increasing.

The market to raise equity has quickly shifted, and financing will only become more necessary. Google had not issued equity since 2006. Microsoft has been buying back stock to the tune of $20B/yr for the past 7 years. These are the world's best businesses, and they are telling you the ROIC and demand for serving someone else's AI models and infrastructure is so high that it is worth taking combined hyperscaler FCF into the negative 9 figures in CY2027. We forecast Anthropic and OpenAI's compute needs as ever growing, and they will become an increasing percentage of total hyperscaler and neo-cloud (including SpaceX and Meta) compute as the buildout continues.

QQ:1413221435

There is no better time than now to raise. Meta will soon raise equity and others have already tapped the markets. As such, the timeline and urgency for Anthropic to raise is clear: capital, and access to it, will be needed.

SemiAnalysis is a reader-supported publication. To receive new posts and support our work, consider becoming a free or paid subscriber.

✓ Subscribed

Anthropic IPO: The Business Overview

Product & Financial Background

If ChatGPT was the iPhone moment for consumers, Claude Code was the inflection point for B2B.

QQ:1413221435

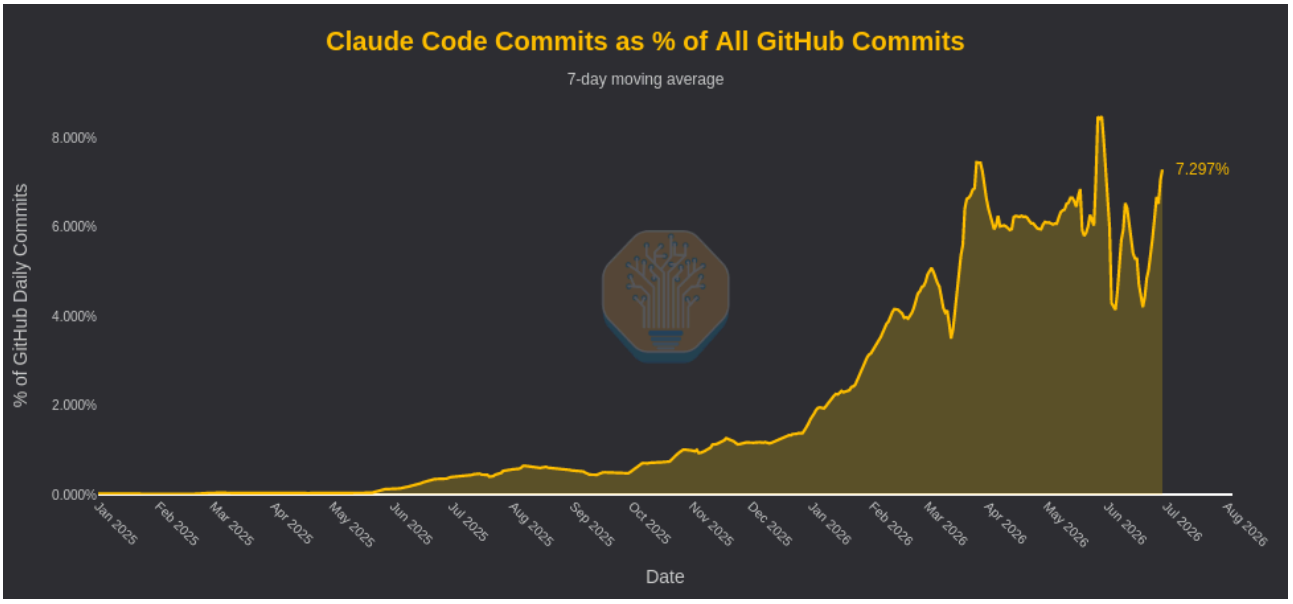


Claude Code is the Inflection Point

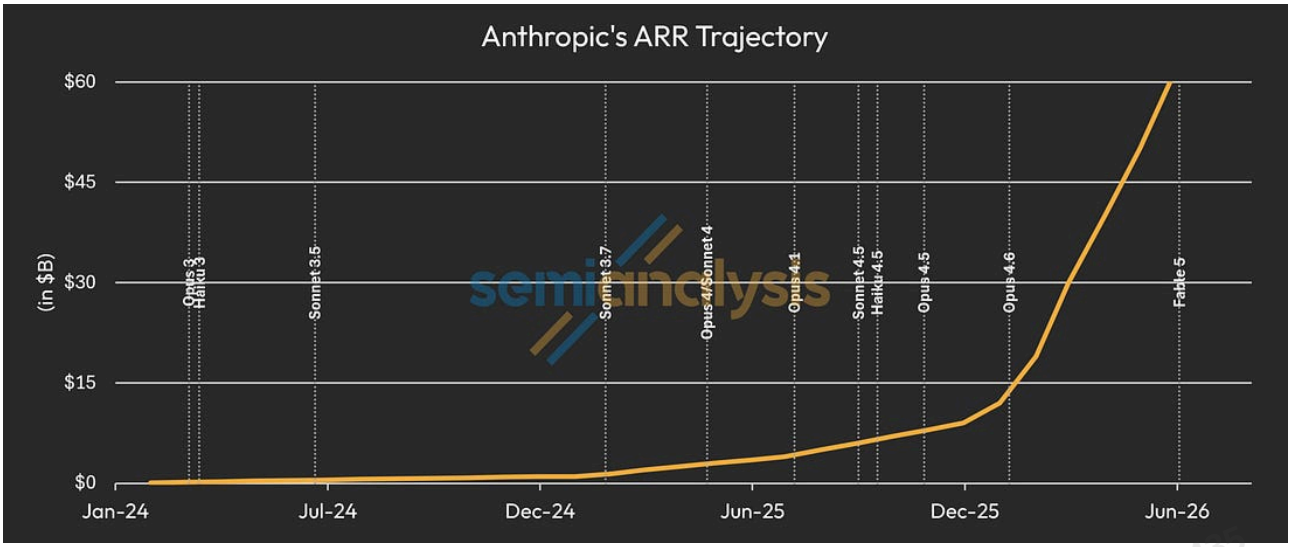
DOUG O'LAUGHLIN, JEREMIE ELIAHOU ONTIVEROS, AND 3 OTHERS·FEB 5

Read full story

Claude Code is now over 7% of total GitHub Commits and has powered the sharp ARR acceleration that Anthropic saw to start 2026. Anthropic ended 2025 with $9B of ARR; today, that number is over $60B. This mix of success in the coding vertical paired with its usage-based pricing model dominates Anthropic ARR today. The company has several different product lines and tiers, but usage-based pricing businesses (which we refer to as API) is 75-85% of its revenue. In contrast, OpenAI has historically been consumer and subscription focused – but things are changing.



Source: SemiAnalysis Research



Source: SemiAnalysis Tokenomics Model

Anthropic's revenue composes several product areas across both subscription and API usage-based pricing. For subscriptions, the company has several business lines:

Consumer, Team, and Enterprise.

Consumer is split into Free, Pro, Max 5x, and Max 20x with pricing on the paid plans $20, $100, and $200/month respectively when billed monthly. Claude Code, Cowork, and features like Claude for Microsoft 365 are included on all paid plans. Our analysis of card panel data suggests that the average paying US consumer is spending around $45/month on Claude. Team subscriptions are for SMBs between 5-150 seats and are $20 or $100/month when billed annually.

Both consumer and team subscriptions have an opaque usage-based system for burning tokens. SemiAnalysis work on the equivalent API token spend showed that subscriptions provide much more usage than the cost for the equivalent task load on API, but Anthropic has the ultimate say on when and how much usage may be throttled.

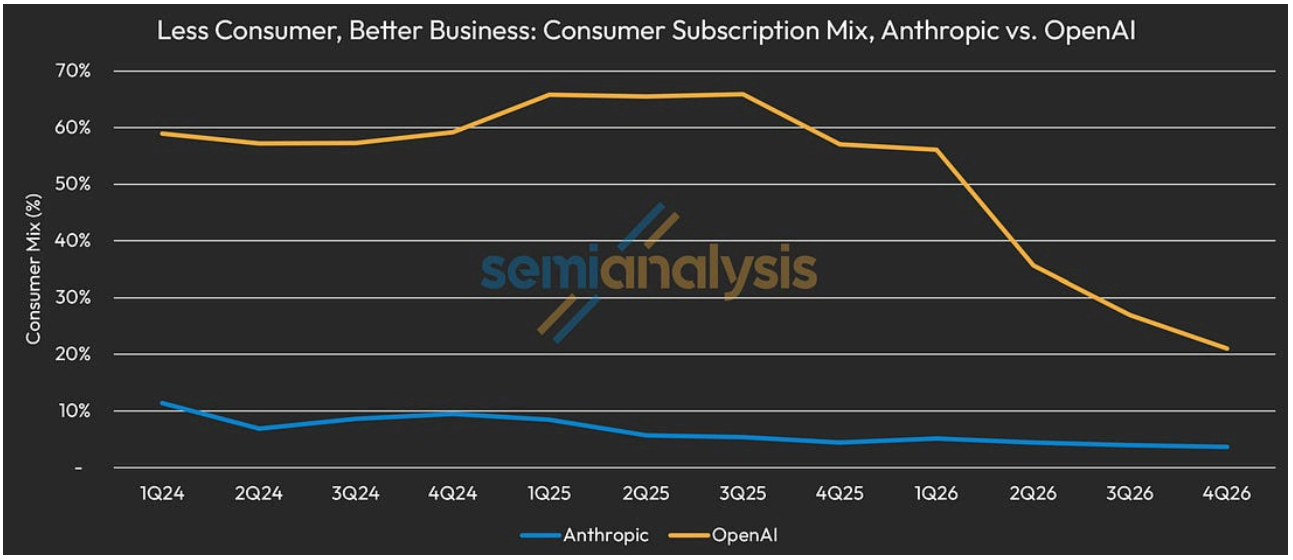
For enterprise subscriptions, however, no usage is included for the subscription price. After 150 seats on the Team plan, you are required to pay $20/month plus the API rates for all usage. The API business is priced on a per million tokens basis. Blended pricing is affected by model use, workload type, and a variety of other factors we cover below.

QQ:1413221435

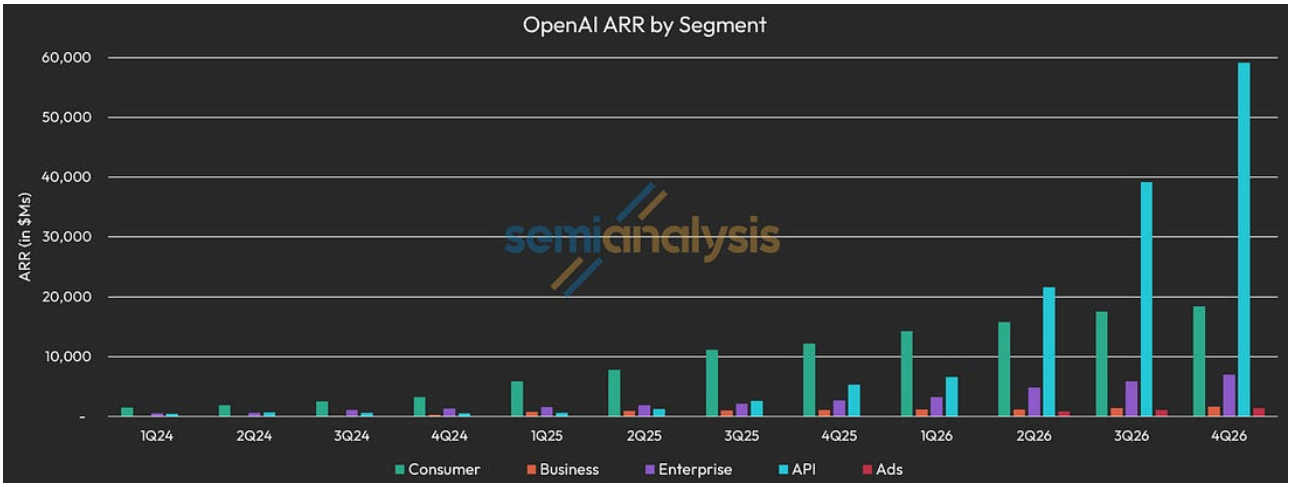
Model	Base Input Tokens	5m Cache Writes	1h Cache Writes	Cache Hits & Refreshes	Output Tokens
Claude Fable 5	$10 / MTok	$12.50 / MTok	$20 / MTok	$1 / MTok	$50 / MTok
Claude Mythos 5 (limited availability)	$10 / MTok	$12.50 / MTok	$20 / MTok	$1 / MTok	$50 / MTok
Claude Opus 4.8	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.7	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.6	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.5	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.1 (deprecated)	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Opus 4 (retired, except on Google Cloud)	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Sonnet 5 through August 31, 2026	$2 / MTok	$2.50 / MTok	$4 / MTok	$0.20 / MTok	$10 / MTok
Claude Sonnet 5 starting September 1, 2026	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Sonnet 4.6	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Sonnet 4.5	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Sonnet 4 (retired, except on Bedrock and Google Cloud)	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Haiku 4.5	$1 / MTok	$1.25 / MTok	$2 / MTok	$0.10 / MTok	$5 / MTok
Claude Haiku 3.5 (retired, except on Bedrock and Google Cloud)	$0.80 / MTok	$1 / MTok	$1.60 / MTok	$0.08 / MTok	$4 / MTok

Source: Anthropic

This API business is where the vast majority of ARR falls today. Subscriptions make up just 15% of total ARR. Compare this to OpenAI and you'll see a stark contrast. During 1Q26, OpenAI was over 65% subscription based. The business vs consumer split is even more stark. Just 5% of Anthropic ARR today comes from Consumer subscriptions vs our estimates of OpenAI around 40% at the end of 2Q26 (the CCO of OpenAI said in a blog that B2B was around 40% of total ARR in 1Q26). Claude Code's success has levered Anthropic to the right tailwinds and business model. As we show in the margin section below, margins on the API business are vastly superior to consumer subscriptions.



Source: SemiAnalysis Tokenomics Model



Source: SemiAnalysis Tokenomics Model

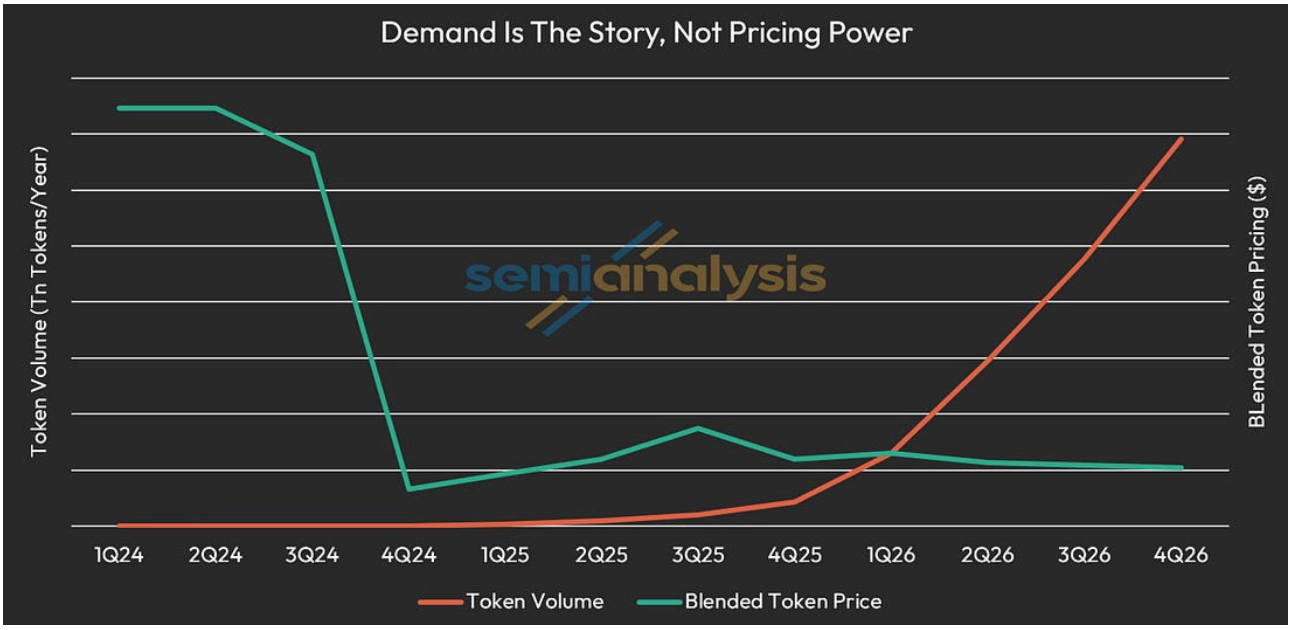
By focusing on API revenue, there is no ceiling on how much revenue can be generated from a single user because it's billed by token rather than by seat. As the same user/customer adopts more agentic workflows, they consume more tokens and generate more revenue without Anthropic needing to add another customer. The disclosures show

this. Anthropic CFO Krishna Rao shared on the Invest Like the Best podcast in early May that Net Dollar Retention, the measure of how much spend has increased at customers who were with the company at least 1 year prior, was 500%. This means $12B of Anthropic's $30B of 1Q26 ARR was driven by the customers who were just $2B of ARR in 1Q25. This means there is still $18B of new customer ARR added over the last year who are still ramping their spend and powering Anthropic's growth higher.

OpenAI is improving, however. The SemiAnalysis Tokenomics team believes that API has been responsible for driving the reacceleration in OpenAI ARR since the March release of 5.5 and Codex and B2B API are now the majority of monthly NNARR. This revenue is higher in margin and could help OpenAI reach profitability sooner than expected.

The AI Lab inference business model is simple. Revenue is dictated by token volume and blended token prices. Token volume has skyrocketed, but blended token pricing has declined over time. With higher token pricing on new models, how has blended pricing gone down? Cache rates, input/output ratios, and agentic vs chat workflow mix are all factors in determining blended token pricing. The Tokenomics Model tracks and projects pricing on a per product (Consumer Max, Free, API, etc.) and per model basis. As more agentic work gets adopted, revenue growth is driven by rising token consumption. Agentic workflows consume significantly more tokens than the traditional chat function, and blended token prices have actually fallen drastically since 2024. The consumption curve, then, is doing all the work in driving revenue growth.

QQ:1413221435



Source: SemiAnalysis Tokenomics Model

QQ:1413221435

Period Date End Calendar Period Fiscal Period	Units	Quarterly 2030 Input	3/31/2024 1Q24 1Q24	6/30/2024 2Q24 2Q24	9/30/2024 3Q24 3Q24	12/31/2024 4Q24 4Q24	3/31/2025 1Q25 1Q25	6/30/2025 2Q25 2Q25	9/30/2025 3Q25 3Q25	12/31/2025 4Q25 4Q25	3/31/2026 1Q26 1Q26	6/30/2026 2Q26 2Q26	9/30/2026 3Q26 3Q26	12/31/2026 4Q26 4Q26
Opus 4.6 Fast Mode:														
API revenue	%	10%	-	-	-	-	-	-	-	-	8%	8%	8%	8%
Average input/output mix	ratio	400									200	230	270	290
Average cache hit rate	%	98%									92%	94%	95%	96%
Average cache write rate	%	85%									94%	90%	88%	86%
Blended price per M tok	$										6.44	5.65	5.22	4.84
API tokens produced (annuazlied runrate)	T tokens										316	764	1,341	2,107
API revenue	$M										2,036	4,316	6,996	10,196
Opus 4.6/4.5:														
API revenue	%	76%	-	-	-	-	-	-	-	30%	70%	71%	72%	74%
Average input/output mix	ratio	350								100	200	250	280	300
Average cache hit rate	%	98%								90%	92%	94%	95%	96%
Average cache write rate	%	85%								90%	94%	90%	88%	86%
Blended price per M tok	$									1.30	1.07	0.93	0.87	0.80
API tokens produced (annuazlied runrate)	T tokens									1,787	16,593	41,022	72,691	117,355
API revenue	$M									2,322	17,815	38,305	62,964	94,313
Opus 4.1/4/3:														
API revenue	%		10%	10%	10%	10%	10%	10%	10%	5%	0.9%	0.6%	0.4%	0.2%
Average input/output mix	ratio		2	2	2	2	8	20	30	100	200	220	230	240
Average cache hit rate	%		-	-	50%	60%	70%	80%	85%	90%	92%	92%	92%	92%
Average cache write rate	%		-	-	50%	60%	70%	80%	85%	90%	94%	94%	94%	94%
Blended price per M tok	$		35.00	35.00	31.13	30.20	13.97	8.14	6.29	3.90	3.22	3.19	3.17	3.16
API tokens produced (annualized runrate)	T tokens		1	1	2	3	12	37	82	99	71	102	110	81
API revenue	$M		23	45	66	88	174	305	516	387	229	324	350	255
Sonnet 4.6/4.5/4/3.7/3.5/3:														
API revenue	%	14%	87%	87%	87%	87%	87%	87%	87%	61%	18%	18%	18%	17%
Average input/output mix	ratio	300	2	2	2	2	8	20	30	100	200	220	230	240
Average cache hit rate	%	95%	-	-	50%	60%	70%	80%	85%	90%	92%	95%	95%	95%
Average cache write rate	%	85%	-	-	50%	60%	70%	80%	85%	90%	94%	85%	85%	85%
Blended price per M tok	$		7.00	7.00	6.23	6.04	2.79	1.63	1.26	0.78	0.64	0.53	0.53	0.53
API tokens produced (annualized runrate)	T tokens		29	56	93	127	542	1,627	3,567	6,055	7,111	18,232	29,712	41,099
API revenue	$M		204	394	578	766	1,514	2,649	4,489	4,721	4,581	9,711	15,741	21,667

semianalysis

Source: SemiAnalysis Tokenomics Model, Anthropic Tab

QQ:1413221435

Broadly, the business model of a lab is to invest in building new models, buy or rent compute to train and serve those models, and then sell that compute to customers via tokens. Above we've shown that Anthropic and OpenAI have very different ways of selling that inference compute historically. Anthropic has wholly aligned their business model with high-margin, usage-based pricing while OpenAI is serving costly free and subscription-based users who can run up token usage that hog compute. While OpenAI continues to converge towards Anthropic with incremental ARR being much more B2B and API based, they still have the weight of consumer and subscription-based usage hitting their financials.

Trajectory & and Discussion of Full P&L

The 1Q26 explosion of Claude Code drove Anthropic ARR from $9B at the end of 2025 to $30B in just 1 quarter. In January, the company added $3B of ARR, in February $7B, and in March $11B. While the consumer business benefitted from a bump from the Department of War incident in 1Q, API has driven the show. Still, in consumers we see a strong business for Anthropic. Free MAUs are likely around the 55-60 million point currently, with 9% of total consumer MAUs paying for the service in some form. This contrasts with OpenAI, who has been stuck around 950M WAU since the beginning of the year and boasts a lower paying-to-free user percentage at around 6% currently

QQ:1413221435

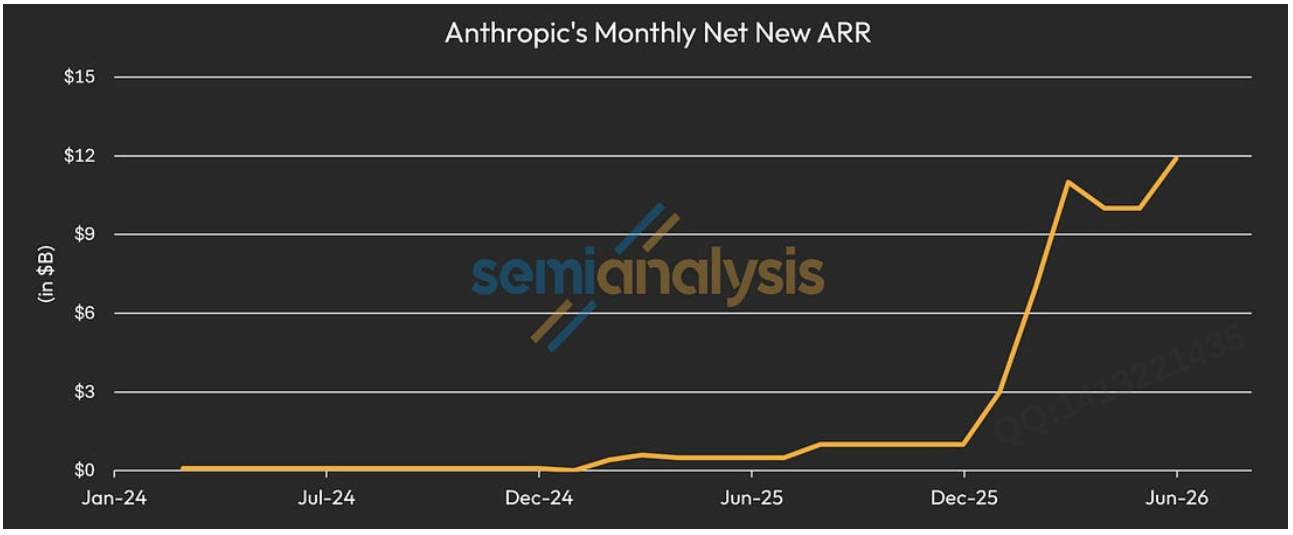
Period Date End			3/31/2024	6/30/2024	9/30/2024	12/31/2024	3/31/2025	6/30/2025	9/30/2025	12/31/2025	3/31/2026	6/30/2026	9/30/2026	12/31/2026
Calendar Period		Quarterly	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26
Fiscal Period	Units	2030 Input	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26
Consumer Subscription Build:														
Consumer MAU, beginning of period	**Users, M**				0.59	2.68	4.44	6.25	8.39	9.92	12.40	39.44	61.27	77.62
net adds in period	Users, M				2.08	1.77	1.81	2.13	1.53	2.48	27.04	21.83	16.35	16.35
Consumer MAU, end of period	**Users, M**			0.59	2.68	4.44	6.25	8.39	9.92	12.40	39.44	61.27	77.62	93.97
Free vs Paying Subscriber Mix:														
Paying Subscriber Attach Rate of MAU	%	9.0%			11.3%	9.8%	10.0%	7.1%	8.5%	7.7%	8.9%	9.0%	9.0%	9.0%
Free Consumer Users, beginning of period	**Users, M**				0.45	2.40	4.05	5.69	7.83	9.14	11.51	36.21	56.21	71.21
net adds in period	Users, M	5.0			1.96	1.64	1.64	2.14	1.31	2.37	24.70	20.00	15.00	15.00
Free Consumer Users, end of period	**Users, M**			0.45	2.40	4.05	5.69	7.83	9.14	11.51	36.21	56.21	71.21	86.21
Paying Consumer Susbcribers, beginning of period	**Users, M**				0.15	0.27	0.40	0.57	0.56	0.77	0.89	3.23	5.06	6.41
net adds in period	Users, M				0.13	0.13	0.17	-0.01	0.22	0.12	2.34	1.83	1.35	1.35
Paying Consumer Susbcribers, end of period	**Users, M**			0.15	0.27	0.40	0.57	0.56	0.77	0.89	3.23	5.06	6.41	7.76
Consumer Subscription Mix Build:														
Paid SKU Mix, % of users:														
Pro, billed monthly	%	70%			95%	95%	95%	89%	85%	84%	80%	78%	75%	72%
Pro, billed annually	%	5%			5%	5%	5%	5%	5%	5%	5%	5%	5%	5%
Max 5x	%	5%			-	-	-	1%	3%	3%	5%	5%	5%	5%
Max 20x	%	20%			-	-	-	5%	7%	8%	10%	12%	15%	18%
		-	-	-	-	-	-	-	-	-	-	-	-	-
SKU Pricing, monthly $:														
Pro, billed monthly	$, actual		20	20	20	20	20	20	20	20	20	20	20	20
Pro, billed annually	$, actual		17	17	17	17	17	17	17	17	17	17	17	17
Max 5x	$, actual		-	-	-	-	-	100	100	100	100	100	100	100
Max 20x	$, actual		-	-	-	-	-	200	200	200	200	200	200	200
Average Users in Period														
Free Users	Users, M				1.43	3.23	4.87	6.76	8.49	10.33	23.86	46.21	63.71	78.71
Pro, billed monthly	Users, M				0.20	0.32	0.46	0.50	0.56	0.70	1.65	3.23	4.30	5.10
Pro, billed annually	Users, M				0.01	0.02	0.02	0.03	0.03	0.04	0.10	0.21	0.29	0.35
Max 5x	Users, M				0.00	0.00	0.00	0.01	0.02	0.02	0.10	0.21	0.29	0.35
Max 20x	Users, M				0.00	0.00	0.00	0.03	0.05	0.07	0.21	0.50	0.86	1.28
Blended Consumer Subscription ARPU, monthly	**$, actual**		20	20	20	20	25	30	35	38	40	45	51	56
months in period	months		3	3	3	3	3	3	3	3	3	3	3	3
Consumer Subscription ARR, end of Period	**$M**		30	35	65	95	170	200	325	400	1,550	2,759	3,911	5,237
Period Subscription Revenue	**$M**				12	20	29	50	69	91	259	565	875	1,195
Revenue by Plan:														
Pro, billed monthly			-	-	12	19	27	30	34	42	99	194	258	306
Pro, billed annually			-	-	1	1	1	1	2	2	5	11	15	18
Max 5x			-	-	-	-	-	2	6	7	31	62	86	106
Max 20x			-	-	-	-	-	17	28	40	124	298	516	765

semianalysis

Source: [SemiAnalysis Tokenomics Model](), Anthropic Tab

QQ:1413221435

This is the stark difference we want to show. OpenAI went after consumer while Anthropic has been strong in the enterprise. While OpenAI has new business lines like Ads in consumer, we believe this is not a material short term to total industry ARR. The opportunity is in B2B, and specifically, in usage-based pricing models. Today, Anthropic is ~75-85% consumption-based API spend which has put Net New ARR in the $10B+/month range as the year has progressed. Fable is the next accelerant of NNARR with higher pricing and a whole host of new workloads and volume expansion being unlocked with the latest model release. We'll revisit this point later in the article, but it's clear that new frontier model releases accelerate ARR growth, and it's imperative for Anthropic to continue to exploit their advantage in agentic API workloads. We believe their P&L and business model allow them to do this.



Source: SemiAnalysis Tokenomics Model

Further down the P&L, we have blended gross margins. Labs rent or buy compute from various sources and sell subscriptions or usage-based tokens on top of that. Labs have several levers to pull to get the most out of a given unit of compute they have for inference. First, new models typically have higher prices. Second, different accelerator chips have different token throughputs and can be optimized over time to better serve inference. Tokens are the fundamental unit of measure for AI workloads that LLMs use to process data. Think of a token as a fragment of the query; it could be a word, part of a word, or punctuation mark in a text-based query.

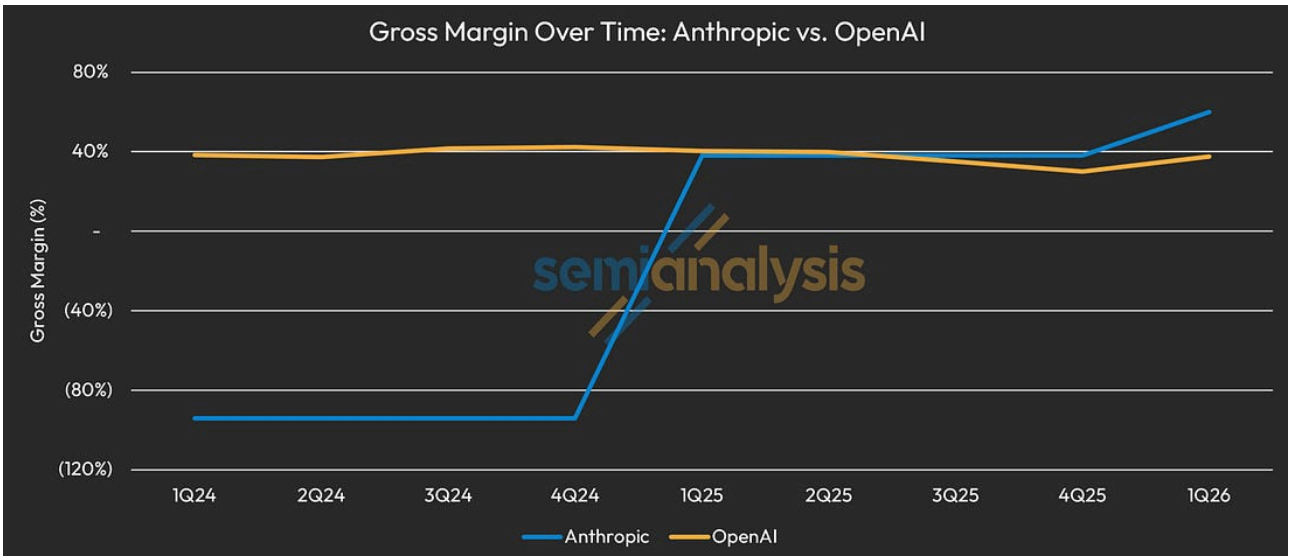
These compute costs are largely fixed per unit of compute, so when you can either: 1) get more tokens through a given unit or 2) higher prices on the tokens you run through a given unit of compute, incremental margins approach 100%. This is why you see blended gross margins today in the mid 60% range vs –94% in 2024.

Labs have made great strides in inference efficiency in the last 12 months, which enabled ARR per MW of compute to drastically rise, and margins followed. Anthropic ARR per MW will reach the $60M range later this year. Just 9 months ago, this number was $16M. This is the powerful margin upside in lab financials at work: token throughput can be improved and better monetized while compute costs are largely fixed, leading to massive margin upside when revenue growth is strong.v

Period Date End Calendar Period Fiscal Period	Units	Quarterly 2030 Input	3/31/2024 1Q24 1Q24	6/30/2024 2Q24 2Q24	9/30/2024 3Q24 3Q24	12/31/2024 4Q24 4Q24	3/31/2025 1Q25 1Q25	6/30/2025 2Q25 2Q25	9/30/2025 3Q25 3Q25	12/31/2025 4Q25 4Q25	3/31/2026 1Q26 1Q26	6/30/2026 2Q26 2Q26	9/30/2026 3Q26 3Q26	12/31/2026 4Q26 4Q26
Implied Training Spend per MW	**$M**	2	**2.21**	**2.43**	**2.21**	**2.80**	**2.82**	**2.55**	**2.09**	**1.50**	2.00	2.00	2.00	2.00
Inference Build:														
Total Revenue	**$M**		**59**	**86**	**158**	**221**	**371**	**691**	**1,191**	**1,876**	**4,890**	**11,513**	**20,125**	**30,194**
y/y %	%						533%	704%	655%	749%	1217%	1566%	1589%	1510%
q/q %	%			47%	84%	40%	68%	86%	72%	57%	161%	135%	75%	50%
Total Cost of Goods Sold, Paid Inference	$M		114	167	306	429	219	408	703	1,107	1,813	4,098	6,863	10,096
y/y %	%						92%	145%	130%	158%	728%	905%	876%	812%
Gross Profit, Paid Inference	**$M**		**(55)**	**(81)**	**(148)**	**(208)**	**152**	**283**	**488**	**769**	**3,077**	**7,415**	**13,262**	**20,098**
y/y %	%													
Gross Margin, Paid Inference	%		(94%)	(94%)	(94%)	(94%)	41%	41%	41%	41%	63%	64%	66%	67%
incremental margins, inference	%													
Total Cost of Goods Sold, Free User Inference	$M						11	21	36	56	143	277	382	472
y/y %	%													
Average Monthly Inference Costs per Free User	$ actual	2.00					0.76	1.02	1.40	1.82	2.00	2.00	2.00	2.00
Total Inference Cost of Goods Sold	$M		114	167	306	429	230	429	739	1,163	1,956	4,375	7,245	10,568
y/y %	%						102%	157%	141%	171%	750%	921%	881%	809%
Total Gross Profit	**%**		**(55)**	**(81)**	**(148)**	**(208)**	**141**	**263**	**453**	**713**	**2,934**	**7,138**	**12,880**	**19,626**
y/y %	%													
Total Gross Margins	**%**	70%	-94%	-94%	-94%	-94%	38%	38%	38%	38%	60%	62%	64%	65%
incremental margins, inference	%						63%	57%	58%	56%	62%	64%	66%	67%
Total Inference Compute	**MW**		**18**	**34**	**50**	**67**	**133**	**233**	**400**	**450**	**1,154**	**2,292**	**3,528**	**4,745**
Inference Blended Outputs:														
Total Revenue Per Inference MW	$M		3.34	2.53	3.14	3.32	2.78	2.96	2.98	4.17	4.24	5.02	5.70	6.36
Average ARR in Period per Inference MW	$M		9.98	11.38	12.56	13.16	11.25	11.79	11.88	16.67	16.90	20.04	22.77	25.41
Total Inference COGS per Inference MW	$M		6.47	4.91	6.09	6.43	1.73	1.84	1.85	2.58	1.70	1.91	2.05	2.23

Source: SemiAnalysis Tokenomics Model, Anthropic Tab

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis Tokenomics Model

OpenAI, again, has a disadvantage against Anthropic. The company is supporting over 900M free users who we estimate cost just under $0.70/month to serve. This is a 20-30% hit to total gross margins historically and means if both OpenAI and Anthropic had $100B of ARR, OpenAI would have $25B less in gross profit. This gross profit can be invested back into the business in the form of training compute/research on the latest model. This leads to a compounding cycle where new models can drive ARR and gross profit higher, leading to more reinvestment in the development of the next model versus competitors

Period Date End Calendar Period Fiscal Period	Units	Quarterly 2030 Input	12/31/2023 4Q23 4Q23	3/31/2024 1Q24 1Q24	6/30/2024 2Q24 2Q24	9/30/2024 3Q24 3Q24	12/31/2024 4Q24 4Q24	3/31/2025 1Q25 1Q25	6/30/2025 2Q25 2Q25	9/30/2025 3Q25 3Q25	12/31/2025 4Q25 4Q25	3/31/2026 1Q26 1Q26	6/30/2026 2Q26 2Q26	9/30/2026 3Q26 3Q26	12/31/2026 4Q26 4Q26
Inference Margin Build:															
Total Revenue	$M			521	738	963	1,239	1,829	2,651	3,625	4,800	5,877	8,778	13,721	19,132
y/y %	%							251%	260%	277%	287%	221%	231%	279%	299%
q/q %	%				42%	31%	29%	48%	45%	37%	32%	22%	49%	56%	39%
Total Cost of Goods Sold, Paid Inference	$M			313	369	385	496	732	1,061	1,269	1,824	2,057	2,985	4,528	6,122
y/y %	%							134%	188%	230%	268%	181%	181%	257%	236%
Gross Profit, Paid Inference	$M			208	369	578	743	1,097	1,591	2,356	2,976	3,820	5,794	9,193	13,010
y/y %	%							427%	331%	308%	300%	248%	264%	290%	337%
Gross Margin, Paid Inference	%	70%		40%	50%	60%	60%	60%	60%	65%	62%	65%	66%	67%	68%
incremental margins, inference	%							68%	64%	67%	63%	67%	69%	68%	70%
Total Cost of Goods Sold, Free User Inference	$M			8	94	178	218	360	535	1,083	1,536	1,607	1,692	1,780	1,869
y/y %	%							4223%	471%	510%	604%	346%	216%	64%	22%
Average Monthly Inference Costs per Free User	$ actual	0.65		0.02	0.19	0.30	0.29	0.34	0.38	0.59	0.68	0.65	0.65	0.65	0.65
Total Inference Cost of Goods Sold	$M			321	463	563	714	1,092	1,596	2,352	3,360	3,664	4,677	6,308	7,991
y/y %	%							240%	245%	318%	371%	236%	193%	168%	138%
Total Gross Profit	%			200	275	400	525	737	1,055	1,273	1,440	2,214	4,101	7,413	11,141
y/y %	%							269%	284%	218%	174%	200%	289%	482%	674%
Total Gross Margins	%			38%	37%	42%	42%	40%	40%	35%	30%	38%	47%	54%	58%
incremental gross margins	%							41%	41%	33%	26%	36%	50%	61%	68%
margin delta y/y	%							2%	3%	-6%	-12%	-3%	7%	19%	28%
margin delta q/q	%				-1%	4%	1%	-2%	0%	-5%	-5%	8%	9%	7%	4%
Microsoft Revenue Share, Actual ($38B cap)	$M			77	110	139	168	206	248	411	550	647	966	1,509	2,066
Microsoft Revenue Share	$M			77	110	139	168	206	248	411	550	647	966	1,509	2,066
% of total revenue	%			15%	15%	14%	14%	11%	9%	11%	11%	11%	11%	11%	11%
implied Microsoft Inference Compute Routing	%	50%		74%	75%	72%	68%	56%	47%	57%	57%	55%	55%	55%	54%
Implied Inference Metrics:															
Implied Revenue per Inference MW Capacity	$M			5.99	5.56	4.99	5.40	7.01	7.35	5.62	5.94	4.98	7.71	21.24	(469.41)
Implied Total Inference Costs per Inference MW	$M			3.69	3.49	2.92	3.11	4.18	4.43	3.65	4.16	3.11	4.11	9.77	(196.06)
Implied Gross Profit per Inference MW	$M			2.30	2.07	2.07	2.29	2.82	2.93	1.97	1.78	1.88	3.60	11.48	(273.35)

Source: SemiAnalysis Tokenomics Model, OpenAI Tab

Gross Margins, though, may be shifting. Anthropic is renting out new compute at over 2x higher market rates (see recent SpaceX deal) and more ARR is coming from channels like AWS Bedrock where the company pays out a 20-30% revenue share. But with API gross margins in the 80%+, Anthropic can easily hold or improve gross margins from their mid 60% level for the foreseeable future.

QQ:1413221435

Below gross profit, labs have Training/R&D Costs and Other Operating Expenditures. What we ultimately want to show is that Anthropic has a massive advantage with its business model in high-margin, usage-based pricing that can be reinvested at very high levels back into training the next state-of-the-art model and widen its lead against not just OpenAI, but every other closed lab and open-source competitor. At Anthropic, we estimate over 60% of compute today is earmarked for Training and internal purposes. We drive a long-range estimate that Anthropic will spend at least 25% of revenue on training compute which puts the training vs inference compute split at 48% vs 52% of companywide compute in 2030, respectively. Anthropic can currently reinvest at rates much higher than OpenAI and we assume their profitability will open more doors such as lower interest rates and ability to procure funding on favorable terms to continue this massive compute buildout to serve 2030 and beyond.

Anthropic, at $1B of GAAP EBIT in 3Q26 (6% margin) and growing NNARR faster than OpenAI, who is sitting at –100% EBIT margins, is a stark contract. We assume Anthropic likely does not want to be profitable, but compute constraints limit what they can reinvest into training and new compute deals.

In summary, the stark gross and EBIT margin differences provide Anthropic with a clear reinvestment advantage over OpenAI. We further explore this phenomenon in the Compute & Long Run Lab Economics section. Anthropic is showing that AI Labs are wonderful, profitable businesses when run with focus and execution. We expect investors to reward them with ample access to capital to fund the next stage of the compute build-out when they go public. For OpenAI, it is paramount that they continue to flex early Codex and API momentum into better margins they can reinvest in model development and chart a path to better profitability for their build-out.

SemiAnalysis is a reader-supported publication. To receive new posts and support our work, consider becoming a free or paid subscriber.

✓ Subscribed

QQ:1413221435

QQ:1413221435

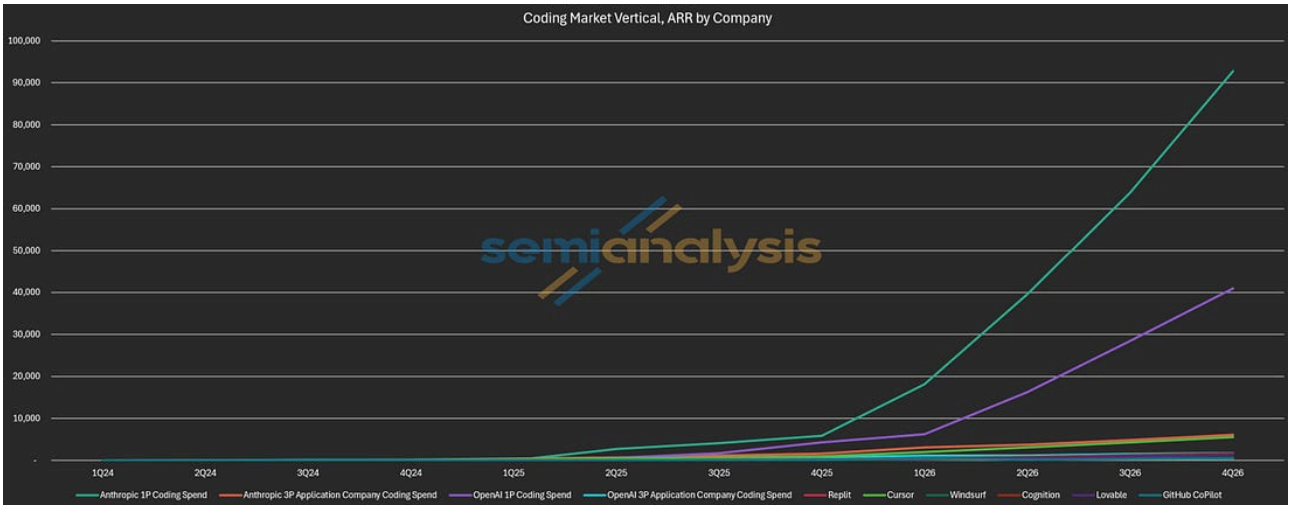
Period Date End			3/31/2024	6/30/2024	9/30/2024	12/31/2024	3/31/2025	6/30/2025	9/30/2025	12/31/2025	3/31/2026	6/30/2026	9/30/2026	12/31/2026
Calendar Period		Quarterly	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26
Fiscal Period	Units	2030 Input	1Q24	2Q24	3Q24	4Q24	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26	4Q26
Consolidated Income Statement:														
Consumers Subscription Revenue	$M		-	-	12	20	29	50	69	91	259	565	875	1,195
Business Subscription Revenue	$M		-	-	6	6	8	14	24	38	98	198	300	411
Enterprise Subscription Revenue	$M		-	-	-	3	8	29	73	134	385	825	1,275	1,725
API Revenue	$M		59	86	140	193	328	598	1,026	1,613	4,149	9,925	17,675	26,863
Total Gross Revenue	**$M**		59	86	158	221	371	691	1,191	1,876	4,890	11,513	20,125	30,194
Cost of Goods Sold, Inference	$M		114	167	306	429	230	429	739	1,163	1,956	4,375	7,245	10,568
Gross Profit	**$M**		(55)	(81)	(148)	(208)	141	263	453	713	2,934	7,138	12,880	19,626
Training Costs	$M		300	350	400	575	750	950	1,150	1,400	2,054	4,605	7,647	10,870
Other OpEx	$M		200	200	250	250	400	600	750	900	2,000	3,000	4,000	6,000
Other Opex as a % of Revenue	%		341%	233%	158%	113%	108%	87%	63%	48%	41%	26%	20%	20%
GAAP Operating Income	**$M**		(555)	(631)	(798)	(1,033)	(1,009)	(1,287)	(1,447)	(1,587)	(1,120)	(467)	1,232	2,756
EBTIT	$M		(255)	(281)	(398)	(458)	(259)	(337)	(297)	(187)	934	4,138	8,880	13,626
Stock Based Compensation														
Implied per MW Metrics:														
Revenue per Inference MW	$M		1.73	2.00	2.62	2.77	3.26	3.95	4.03	3.02	5.63	35.82	(29.61)	(19.62)
Inference COGS per Inference MW	$M		3.35	3.88	5.08	5.37	2.02	2.45	2.50	1.87	2.25	13.61	(10.66)	(6.87)
Training COGS per Training MW	$M		2.21	2.43	2.21	2.80	2.82	2.55	2.09	1.50	2.00	2.00	2.00	2.00
EBTIT per Inference MW	$M		(7.50)	(6.53)	(6.61)	(5.74)	(2.27)	(1.93)	(1.01)	(0.30)	1.07	12.87	(13.06)	(8.85)
EBIT per Total MW	$M		(16.33)	(14.67)	(13.25)	(12.94)	(8.85)	(7.35)	(4.89)	(2.55)	(1.29)	(1.45)	(1.81)	(1.79)
Topline Growth Metrics:														
Consumers Subscription Revenue	%								460%	360%	802%	1032%	1159%	1208%
Business Subscription Revenue	%								308%	559%	1200%	1339%	1163%	996%
Enterprise Subscription Revenue	%									5275%	5033%	2709%	1659%	1184%
API Revenue	%						458%	596%	634%	735%	1167%	1559%	1623%	1566%
Cost Metrics:														
Training Costs % of Revenue	%			407%	253%	260%	202%	137%	97%	75%	42%	40%	38%	36%
Other Opex % of Revenue	%			233%	158%	113%	108%	87%	63%	48%	41%	26%	20%	20%
Total Opex % of Revenue	%			640%	412%	373%	310%	224%	159%	123%	83%	66%	58%	56%
Implied Token Metrics:														
API Revenue per M/API Tokens	$/M Tokens		5.60	4.25	3.87	4.20	2.18	1.38	1.12	0.72	0.62	0.64	0.65	0.65
Total Revenue per M/Tokens	$/M Tokens		5.60	4.25	4.19	4.88	2.17	1.32	1.04	0.75	0.64	0.63	0.66	0.66
Total Inference Costs per M/Tokens	$/M Tokens		10.86	8.25	7.51	8.15	1.35	0.85	0.69	0.45	0.25	0.24	0.23	0.23
EBIT per M/Tokens	$/M Tokens		(53.00)	(31.19)	(19.58)	(19.62)	(5.93)	(2.57)	(1.36)	(0.61)	(0.14)	(0.03)	0.04	0.06

Source: SemiAnalysis Tokenomics Model, Anthropic Tab

QQ:1413221435

Coding Market, Large Customers, and Other

Power users at power customers drive lab revenue today. That is, spend is narrower and heavily concentrated in coding workflows, but is broad based enough that we believe Meta is the largest customer while only being 3-5% of total Anthropic revenue. The Tokenomics team estimates that over 65% of Lab ARR today is from coding use cases and a host of other coding startups such as Cursor, Cognition, Loveable, Replit, and more compose a additional $6B of ARR as of 2Q26.



Source: SemiAnalysis Tokenomics Model, CodingMarket Tab

QQ:1413221435

While wrapper companies are numerous, they represent less than 10-15% of total lab ARR today as industry growth has broadened. Large enterprises customers are numerous

but still largely ramping spend. Meta, the highest profile case of tokenmaxxing, is likely Anthropic's largest customer and spends in the 9 figures.

TaaS, Hyperscalers, And Sales

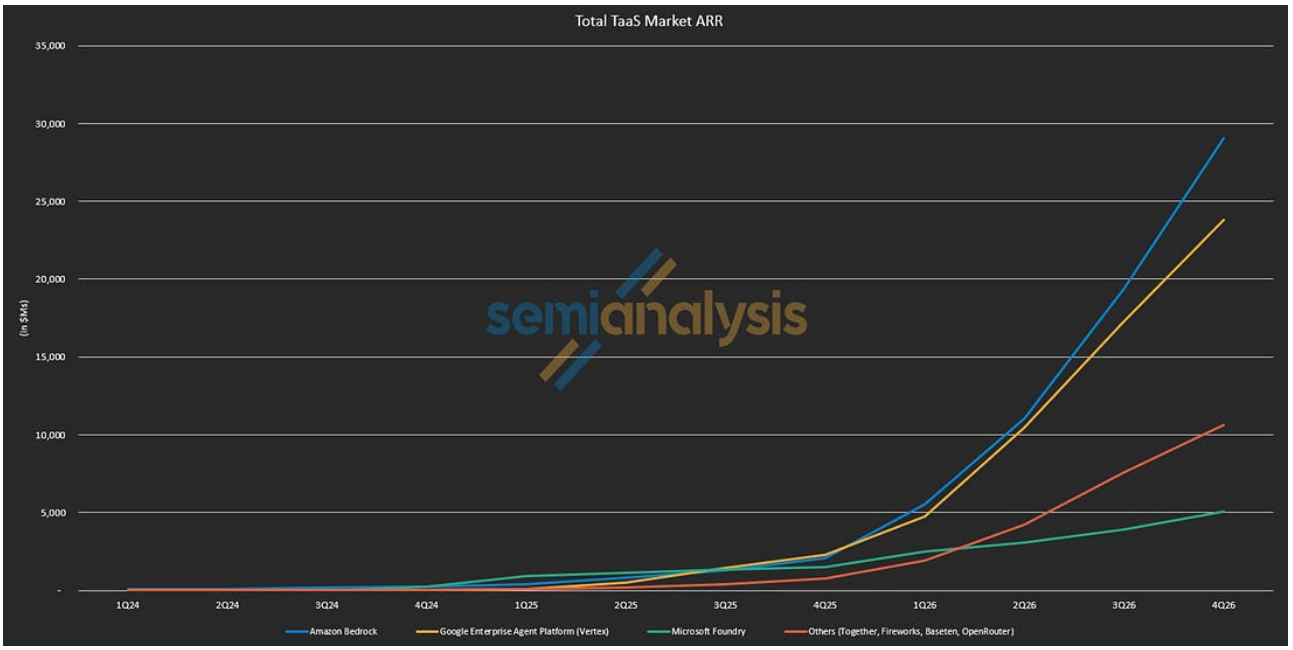
Recently, the token-as-a-service business model has taken off. For customers, this means buying a model through an external provider API rather than directly from the lab. Customers can purchase any model available through that provider and consume it via API usage. Our work suggests that Anthropic now serves 15-20% of ARR through indirect channels such as AWS Bedrock in our Azure Foundry vs just 5-10% one quarter ago.

With Claude available on Bedrock, Foundry, and Gemini Agent Enterprise Platform (formerly Vertex), the big 3 hyperscalers have 3 things going for them that large enterprises prefer:

- Access and flexibility to spend on any model
    - If OpenAI has the better model in 6 months for a given workflow, the customer can easily switch spend and is not locked into one model provider
- Ability to use committed usage and ELAs towards any service from the Cloud provider, including any model within a TaaS product
- Existing relationships with strong security and compliance frameworks that enterprises trust
    - No need to qualify a new vendor directly in a 6–12-month RFP
- Better latency for international workloads (Anthropic routes all direct inference queries back to the US)

So, is this better for Anthropic if 1) enterprises can easily switch models; and 2) you're paying a revenue share for giving them that privilege? We think yes. TaaS players give labs capital flexibility with more favorable IaaS fee terms than directly serving inference and allow labs to tap into these massive businesses' customer bases of large enterprises and cloud estates. Enterprise sales are difficult. This is where the Hyperscalers have an advantage: hundreds of billions of annual enterprise customer spend make it easier for Anthropic and OpenAI to sell into the base. In many cases, it's preferable to give a 20% revenue share to a hyperscaler for their distribution advantages versus paying a BDR, enterprise sales rep, sales engineer, customer success team to manage the land, expand, and retention.

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis Tokenomics Model, TaaS Market Tab

QQ:1413221435

Enterprise adoption of the TaaS business has been strong. As of 2Q estimate the market at $28B of ARR with 85% of the market share between the big 3 clouds. Because of the above factors, we believe TaaS will be a powerful business model going forward and Anthropic ARR to increasingly compose more indirect relationships like these.

Risks & Upside

We see several risks that investors should watch for:

- Top Line Risks
  - Tokenbudgeting Concerns
  - OSS and Price Cuts
    - OpenAI has been rumored to cut token pricing to try to win back spend
  - Competitive Coding Models from Google and Meta
  - New Model Blockades
    - A US regulatory regime that blocks model releases and cedes AI leadership and allows open source to catch up to current Fable model capabilities almost fully eliminates Anthropic's inherent business advantages
  - Compute Shortages
    - Where will the supply come from?
- Margin Risks
  - Margins from more TaaS mix

QQ:1413221435

Tokenbudgeting is the latest concern in lab ARR trajectories. When our team looks at widely publicized cases of Tokenbudgetting such as Coinbase, we see a company who just laid off 14% of their workforce due to a massive cyclical bear market in their end market rather than a pure look at ROI. As we touched on in our recent TokenMaxxing vs TokenBudgeting note, Global 2000 use of AI is still at a miniscule revenue base relative in their share of total lab ARR against their broader spend on IT. Anthropic's own help pages the company states that the average Claude Code enterprise user spends just $150-250/month and that 90% of users spend less than $30/day. We'd like to reiterate: the tokenmaxxing stories are at the tail of the distribution, and most users at the tail that continue to spend large amounts are enjoying incredible ROI on their spend.



TokenBudgeting: Our Conversations with Enterprises on Token Spend

CRYSTAL HUANG, JOEY BROOKHART, AND DYLAN PATEL·JUN 30

Read full story

The runway for breadth and depth of use within the enterprise remains wide and unpenetrated today. Enterprise decision makers like the predictability of knowing what they're going to pay. We expect usage limits/budgets to be in place for many roles, but as ROI increases, we expect those per seat budgets to increase and for top performers, who generate enormous ROI, to be given almost unlimited budgets.

Even if tasks begin to shift to the open source or a lower tier than the latest SOTA model, token pricing and volume at Anthropic will stay elevated if it remains a 2-horse race at the frontier. Token pricing deflation will need to be driven by an increase in SOTA frontier models, not a shifting of workloads to lower priced models given the natural competitive dynamics of customers. Why would a management team want their best users using the last generation of intelligence when competitors are using the latest? If the recent release of Sonnet 5 showed anything, it's that even the lower tier closed models actually have the power to command higher pricing than previous generations. We chalk this up to the overwhelming demand factor. Today, the market is so narrow in scope at the frontier level, that new workflows in the other 95%+ of the economy (i.e. non-coding) are bound to take the torch and power Anthropic ARR higher through pricing and volume that keeps frontier models at the forefront of workload mix.

Other lab competition, as such, have the most power to cause disruption to the coding market. Meta SuperIntelligence and Google DeepMind's new model success will be the deciding factor here on forcing token pricing down, and with it, gross margins. If Meta and Google can turn coding into a 4-horse race, token pricing will obviously be affected. Does this mean Anthropic NNARR will turn negative? Probably not. The latest and best intelligence will always have a price; workload expansion and new models from Anthropic

QQ:1413221435

will mean Google and Meta need to catch up on the next thing. Anthropic, and any major lab, will always be competing to outrun other labs across closed and open source. Thus, if new model capabilities remain up-and-to-the-right and that time between SOTA and open-source capabilities is wide, pricing power and the power laws of usage remain in Anthropic's favor.

SemiAnalysis is a reader-supported publication. To receive new posts and support our work, consider becoming a free or paid subscriber.

✓ Subscribed

Next on risks, more Fable style government delays hurt frontier labs. What happens to model ROIC when governments and safety experts limit who can use the latest models. Is Anthropic's biggest risk their own government? While regulatory lock-in can help some businesses, this just isn't the case when you have well-funded hyperscaler labs playing catch-up and Chinese labs distilling your models. Frontier labs want to increase the time it takes for open source and other labs to get to an equivalent model capability level. Delays at monetizing the frontier by governments risk unseating US dominance in the AI space and the US' broader global leadership in technology. While we view this as a small risk today, we urge regulators to be mindful of what is at stake.

QQ:1413221435

However, on the growth side we see a combination for upside through:

- New Customers Ramping Seats and Usage throughout 2H26 and 27
- Cyber with Mythos/Fable to be the next major vertical to power growth
    - With growth ramp faster the Coding

Our house view is that Fable from both a pricing mix per M Tokens and expanded use case perspective will help accelerate Anthropic monthly NNARR above its current $10B/month pace in 2H26.

The number of customers spending over $100k/year on Claude has grown 7x in the past year, and customers spending over $1M /year up roughly 42x over the past two years. When we look at the SaaS and Cloud optimization phase in 2022 and 2023, we see the hit to multiples and growth rates that can occur. Many of these businesses like Snowflake, Datadog, and Cloudflare traded at ridiculous multiples (50x forward revenue or more!) and growth sharply decelerated. We don't see the major risks to that sort of optimization in the short term. The Global 2000/Fortune 500, who make up a large amount of total IT spend,

QQ:1413221435

are so nascent today in their usage of frontier AI products that the ramp period in terms of seat deployments and per seat usage are major factors in 2H26 and 2027 NNARR growth. In addition, new use cases are being unlocked every day, and we expect what Claude Code did for growth will be repeated at a higher and faster pace with Mythos and cybersecurity. Healthcare, finance, biotech, and more verticals will all have their moments of massive TAM workload expansion and aggressive growth rates up those s-curves.

At SemiAnalysis, these risk and reward factors lead us to conservatively model NNARR acceleration into 2H26 and 2027. Demand for $300B of year end ARR is easily possible in 2027 if Anthropic can accelerate to just $15B/month of NNARR.

Stead State Margins

Anthropic's long-term margins will reflect the best-of-breed nature of their products and business models. Like the best software companies of yesteryears, 30-40% EBIT/FCF margins are possible. With gross margins moving into the mid 75s, training/R&D at 48% of total compute, and Other OpEx at 20% of revenue, AI labs are an excellent business model, and Anthropic will continue to show why.

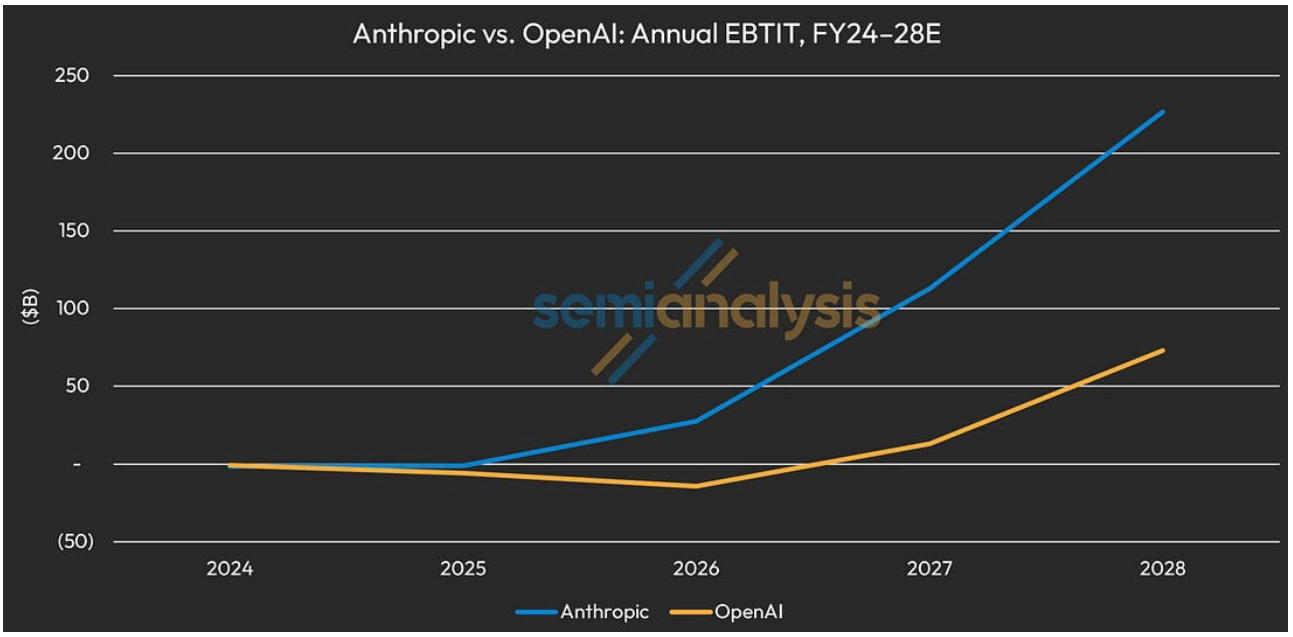
QQ:1413221435

Compute & Long Run Lab Economics

Anthropic's advantage lies in its gross margins (which itself is a function of their API usage revenue mix). Labs will want to show some profitability when public, but Anthropic itself has a higher revenue base and much higher gross margins to plow into building the next SOTA model. On a dollar basis, we see Anthropic with $160B of capital after COGS to plow into reinvestment in 2027 vs $92B at OpenAI.

Earning Before Training, Interest, and Taxes (EBTIT) is the new default metric for lab reinvestment and the best source of truth of where cash economics sit for the inference business. Anthropic has excelled here at showing that AI can be wildly profitable with 36% EBTIT margins in 2Q26. However, the broader implication is that every dollar of EBTIT above OpenAI, DeepMind, Meta SuperIntelligence, and the entire neolab and opensource community helps strengthen their leadership in model capabilities. This model capability lead increases the time for non-frontier models to catch up and is a source of pricing power for workloads that happen at the frontier. This cumulative EBTIT $ advantage of $250B vs OpenAI through 2028 is the key reason we believe Anthropic should IPO first and procure as much compute as possible and as quickly as possible.

QQ:1413221435



Source: SemiAnalysis Tokenomics Model

QQ:1413221435

We see no blip in the up-and-to-the-right nature of model improvements, so why stop? Cumulative training costs of $8B since 1Q24 have yielded a $60B ARR business today – an extremely impressive ROIC. It is not hard to imagine what the stacking of s-curves in new verticals such as cybersecurity, biotech/healthcare, finance, and more will do the lab ARRs as model capabilities improve and the flywheel of high-margin inference API revenue will provide the funding for widening the intelligence gap for the next generation of models. Anthropic has the opportunity to invest the most in these new verticals and attack these markets first if they flex their capital advantage.

However, this is only possible if there is compute to serve it. We forecast labs at over 100GW of unconstrained compute demand in 2030. Capital will need to meet demand and that's the broader story of this IPO: $100B+ of equity issuance this year by some of the best businesses ever (hyperscalers) is just a precursor to the capital needed over the next 2-3 years to continue this massive buildout to serve demand. Anthropic's and OpenAI's IPOs' are a turning point for the broader ecosystem requiring trillions per year in access to financing.

Conclusion & Valuation

Anthropic has built an enviable business. The company is the clear winner in the AI Lab race across the two most important financial metrics: growth rate and profitability. However, OpenAI is catching up in the areas that matter. Codex reception and our enterprise spending checks on API have been overwhelmingly positive since the release of 5.5. OpenAI's ability to continue to accelerate NNARR via their API business will be the deciding factor in how big a lead Anthropic can take.

QQ:1413221435

Thus, we believe Anthropic should attempt to IPO first. Their financials are better, their business model is better, and they can continue to press that advantage and force capital their way. The impetus is then on OpenAI to raise equity with worse financials and dollar growth, so Anthropic should keep their foot on the gas.

Our base case for valuation is 20x 2027, ending ARR of $300B (~15B/month of NNARR in 2027). This would put Anthropic's enterprise value at $6 trillion – making it the world's largest company.

We update our Anthropic, OpenAI, and hyperscaler estimates monthly down to the business line, token, margin, and compute level every month. To get access to our estimates, research, and opinions, please reach out to sales@semianalysis.com

QQ:1413221435

QQ:1413221435