Anthropic 3Q26 Profit Over $1B: The Anthropic IPO Financials Sneak Peak

Anthropic 2026 年第三季度利润超 10 亿美元：Anthropic IPO 财务数据抢先看

Anthropic's Opportunity is Theirs to Lose

Anthropic 的机会不容有失

JOEY BROOKHART, CRYSTAL HUANG, AND DYLAN PATEL

乔伊·布鲁克哈特、黄水晶、迪伦·帕特尔

JUL 08, 2026 2026 年 7 月 8 日 · PAID · 已付费

1 2 Share 分享

Introduction 引言

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.

当达里奥·阿莫代于 2021 年初离开 OpenAI 创立 Anthropic 时，ChatGPT 的病毒式传播还要等 18 个月以上，LLMs 的商业化几乎为零。短短几年后，Anthropic 和 OpenAI 合计年经常性收入突破 1000 亿美元，而在 2026 年，随着 Claude Code 席卷软件开发领域，AI 模型盈利化中一个明确的获胜者已经出现。

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.

Anthropic 于 6 月 1 日秘密提交了 IPO 申请。一个多月后，来自超大规模云服务商的股权融资迫在眉睫，同时有报道称 OpenAI 将其 IPO 推迟至 2027 年，这引发了一些人对 AI 实验室融资能力的质疑。然而，在今天，Anthropic 在捕捉 B2B 市场方面显然是绝对的领跑者，并且以一种盈利的方式，对抗着一个缺乏专注且烧钱的竞争对手。

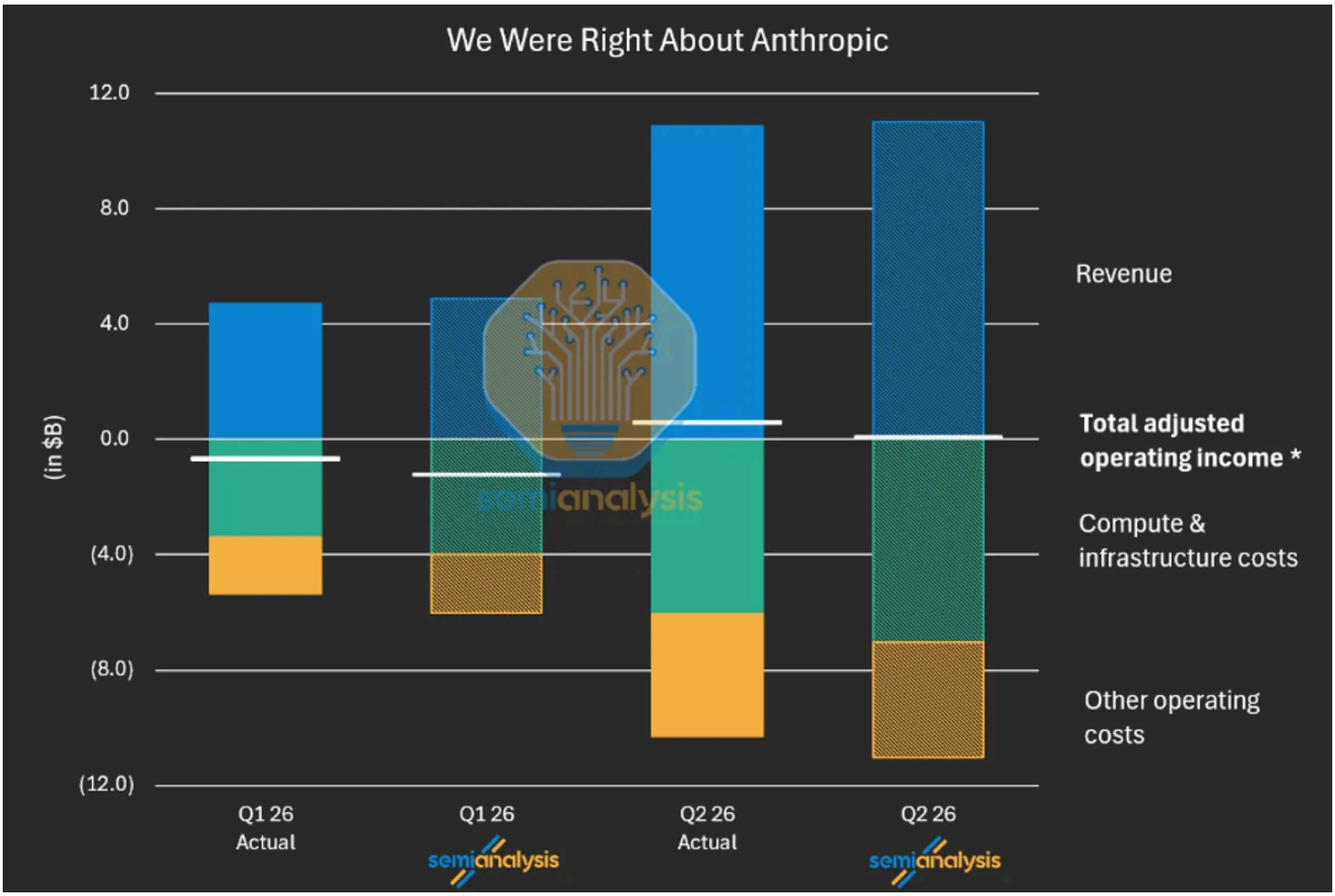
With this lead, we expect Anthropic to take advantage of their superior business model and margins to invest in further in new models that help extend their lead and monetization over closed and open source competitors. Anthropic has the ability to truly make OpenAI dance and we see Anthropic as the 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.

凭借这一优势，我们预计 Anthropic 将利用其卓越的商业模式和利润率，进一步投资开发新模型，从而扩大其对闭源和开源竞争对手的领先优势及变现能力。Anthropic 完全有能力让 OpenAI 疲于应对，我们认为若 Anthropic 持续高效执行战略，其成为首家市值达 6 万亿美元的公司是基本可能。定价权、毛利率、商业模式及盈利能力等综合因素，都将促使 Anthropic 率先启动 IPO，迫使 OpenAI 不得不公开财务数据并筹集必要资金以应对竞争，为即将到来的大规模 AI 基础设施建设提供资金支持。

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.

今年已有两家 AI 实验室（中国的智谱和 MiniMax）申请 IPO，但 Anthropic 将是首个达到如此规模的 AI 实验室进行上市。保密申请意味着没有公开的财务数据。幸运的是，SemiAnalysis 的代币经济学团队致力于按 SKU、层级和客户类型自下而上地构建财务模型。近期，《华尔街日报》一篇关于 Anthropic 财务数据的文章证实了我们的代币经济学团队在跨实验室和超大规模企业中的工作准确性，帮助投资者、企业及其他利益相关者理解 AI 生态系统的经济与财务状况。



Source: WSJ, SemiAnalysis Tokenomics Model

来源：《华尔街日报》、SemiAnalysis 代币经济学模型

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.

本文将深入探讨我们团队在代币经济学模型中的工作成果，带您了解 Anthropic 的财务细节、与 OpenAI 的对比、我们对市场走向的判断，以及对价值链和更广泛 AI 市场的影响。虽然 SemiAnalysis 的其他工作侧重于 AI 实验室的技术方面，但本文将完全聚焦于 Anthropic 当前和未来的财务数据、利润率经济学以及长期前景。

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

如果您对我们的 Anthropic、OpenAI、超大规模服务商等领域的报道或模型感兴趣，请通过 sales@semianalysis.com 联系我们。

Background of the IPO IPO 背景

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:

Anthropic 的资本状况十分稳健。在募集超过 1000 亿美元、实现盈利（按非公认会计准则计算）且保持高毛利率的情况下，为何要在风投资金看似取之不尽、且 Stripe 和 Databricks 等公司选择远离公众视野的时代上市？我们认为主要有三个原因：

- Access to equity and debt markets (which will become a necessity)

  获取股权和债务市场融资（这将成为必要之举）

- Audited and public financials help ease enterprises making large contract commitments

  经审计的公开财报有助于缓解企业在签订大额合同时的顾虑

- Make the talent war easier by allowing employees to easily sell shares

  通过允许员工轻松出售股份，简化人才争夺战

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.

在当下人才争夺激烈堪比职业体育、且算力需求巨大且持续增长的世界里，拥有可流通的公开股权以及快速筹集债务和股本的能力，具有战略意义。

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

semianalysis

Source: SemiAnalysis SemiAnalysis

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.

我们认为，这些资金将用于填补 Anthropic 在推理运算、新模型训练投资与可获取算力之间日益扩大的供需缺口。与上述 SpaceX 市场利率交易类似，Anthropic 预计将从 Meta 等可信供应商处采购额外算力——近期有传言称，Meta 自 2026 年 7 月 1 日起正探索向外部方出租算力的业务。我们的通证经济学模型清晰展示了 Anthropic 算力需求（基于训练支出占收入比及推理 ARR/MW 指标）与数据中心模型追踪的建设进度所反映的供应能力之间的悬殊差距。

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.

实际上，我们预测到 2030 年底，OpenAI 与 Anthropic 的算力需求合计将超过 100GW。这意味着到 2030 年需净新增超 90GW 算力，而 2025 年新增量仅为 2.5GW，2026 年预计达 5GW。当前我们估算 OpenAI 与 Anthropic 可获取的算力仅略超 6GW。

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.

在 2026 年第一季度，Anthropic 用户切实感受到了计算资源短缺。频率限制、服务宕机及其他限流措施让 Claude 的使用变得困难重重。正如我们此前在 Tokenomics 订阅内容中所述，Anthropic 的 API 业务毛利率极为强劲。开放资本市场将有助于 Anthropic 发挥其利润率优势，以优惠利率获取资金，从而将计算资源竞价推升至竞争对手难以匹敌的水平。正如我们在业务概览章节中详细阐述的，OpenAI 目前尚未享有同等优势。

IPO/Capital Raise Market *IPO/资本市场融资*

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.

今年 6 月，Alphabet 完成了 847.5 亿美元的股权融资。据传 Meta 的融资规模也高达数百亿美元。正如我们在 4 月份向机构订阅用户发布的《自由现金流归零》报告中所述，超大规模云服务商在 2027 年及以后将无法产生足够的经营现金流来支撑资本开支需求。这些资金连同人工智能基础设施建设的其余投入，必须另寻来源。无论资本成本多低，资本始终面临竞争，而供给量正急剧且快速地增长。

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.

股权市场风向已迅速转变，融资需求只会愈发迫切。谷歌自 2006 年以来从未发行过股票，而微软在过去七年里每年斥资 200 亿美元进行股票回购。这些全球顶尖的企业正用行动表明：为他人提供 AI 模型与基础设施服务的投资回报率（ROIC）及需求如此之高，以至于到 2027 年，超大规模云服务商的自由现金流（FCF）可能会合计跌至负九位数。我们预测，随着基础设施建设的持续推进，Anthropic 与 OpenAI 的算力需求将持续增长，其在超大规模云服务商及新型云平台（包括 SpaceX 和 Meta）总算力中的占比将不断攀升。

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.

当前正是融资的最佳时机。Meta 即将启动股权融资，其他企业也已先行试水。因此，Anthropic 的融资时间表与紧迫性已然明确：资本，以及获取资本的渠道，将成为其核心需求。

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

or paid subscriber.

SemiAnalysis 是一份由读者支持的刊物。如需接收新文章并支持我们的工作，可考虑成为免费或付费订阅用户。

Anthropic IPO: The Business Overview

Anthropic IPO：业务概览

Product & Financial Background

产品与财务背景

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

如果说 ChatGPT 是消费者的 iPhone 时刻，那么 Claude Code 就是 B2B 领域的转折点。



Claude Code is the Inflection Point

Claude Code 是关键转折点

DOUG O'LAUGHLIN, JEREMIE ELIAHOU ONTIVEROS, AND 3 OTHERS

道格·奥劳克林、杰里米·埃利亚胡·翁蒂维罗斯 等三人

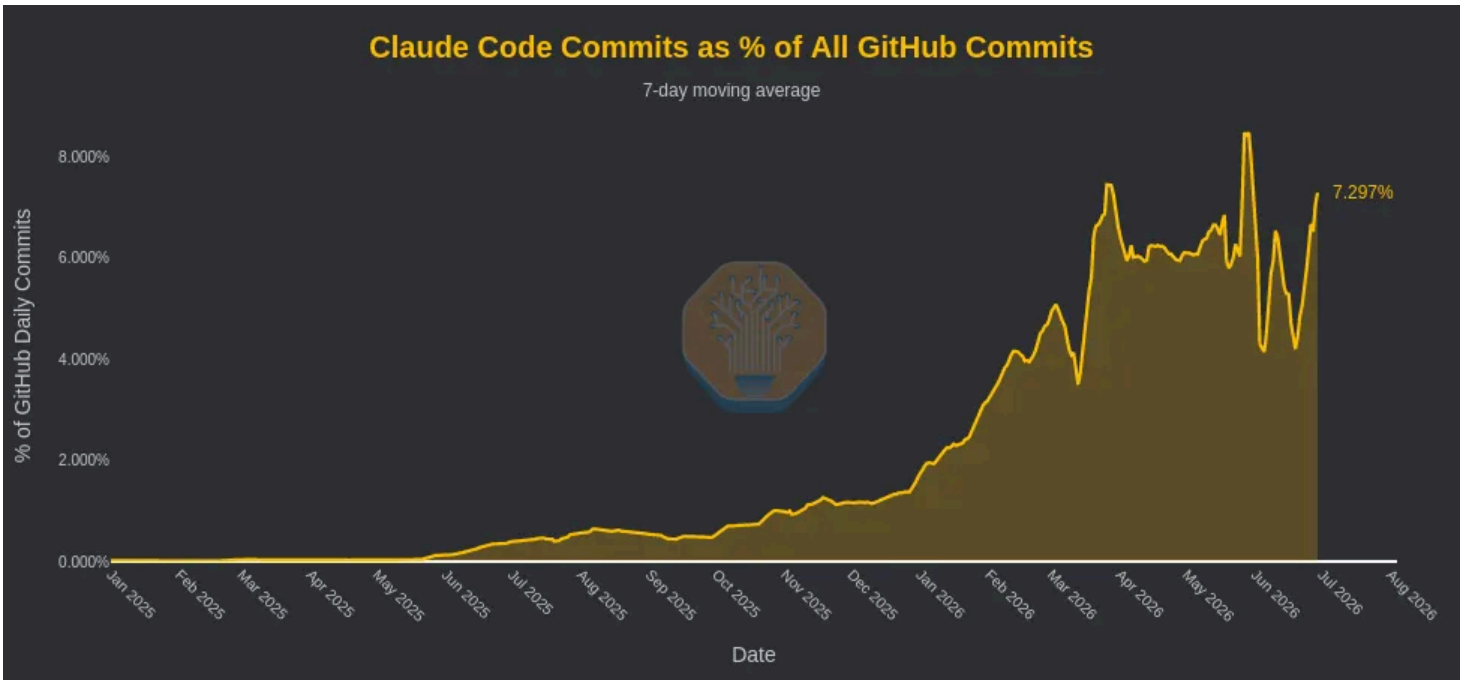
• 2月6日 2 月 6 日

Read full story 阅读完整报道 →

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

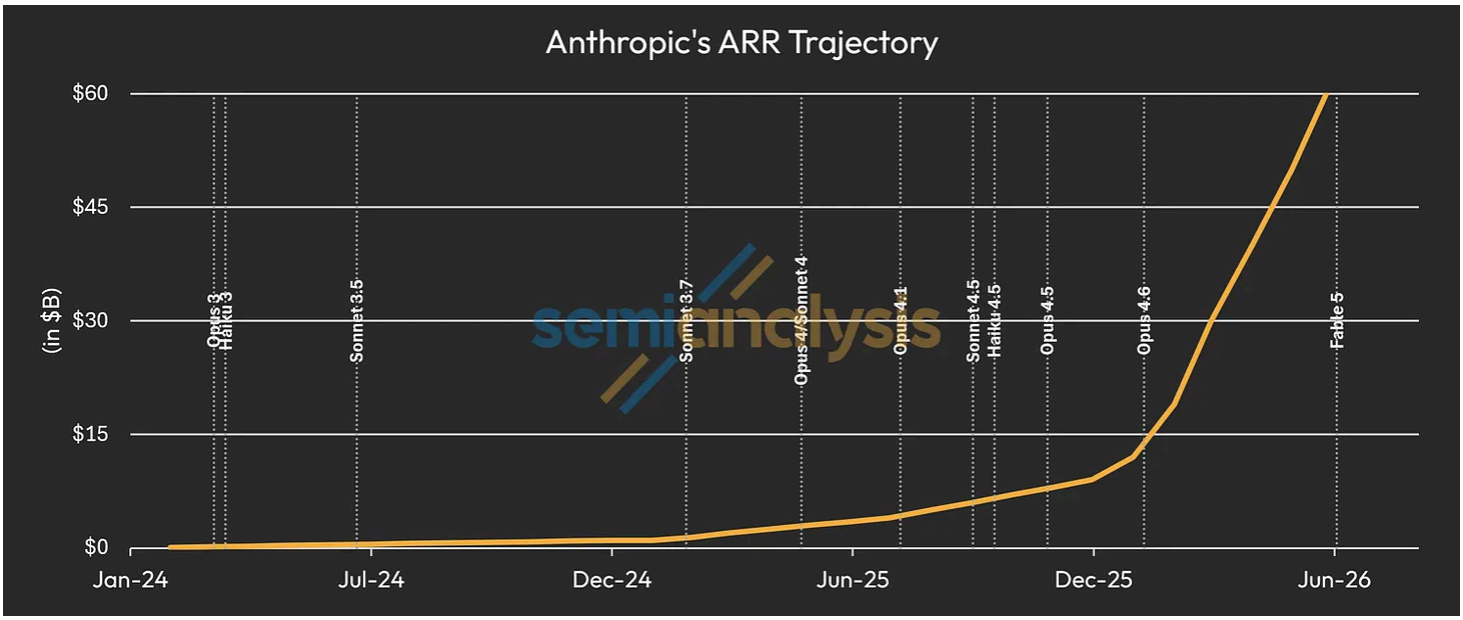
consumer and subscription focused – but things are changing.

Claude Code 目前已占 GitHub 总提交量的 7% 以上，并推动了 Anthropic 在 2026 年初 ARR 的急剧加速增长。Anthropic 在 2025 年底的 ARR 为 90 亿美元；如今，这一数字已超过 600 亿美元。这种在编码垂直领域的成功，结合其基于使用量的定价模式，目前主导了 Anthropic 的 ARR。该公司拥有多个不同的产品线和层级，但基于使用量的定价业务（我们称之为 API）占其收入的 75% 至 85%。相比之下，OpenAI 历来以消费者和订阅模式为重点——但情况正在发生变化。



Source: SemiAnalysis Research

来源：SemiAnalysis 研究



Source: SemiAnalysis Tokenomics Model

来源: SemiAnalysis 代币经济学模型

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.

Anthropic 的收入由订阅和基于 API 使用量定价的多个产品领域组成。在订阅方面，该公司拥有多条业务线：个人版、团队版和企业版。

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.

消费端分为免费版、Pro 版、Max 5x 版和 Max 20x 版，付费套餐按月订阅的价格分别为每月 20 美元、100 美元和 200 美元。所有付费套餐均包含 Claude Code、Cowork 以及 Claude for Microsoft 365 等功能。我们对银行卡面板数据的分析显示，美国付费用户平均每月在 Claude 上的支出约为 45 美元。团队订阅面向 5 至 150 个席位的中小企业，按年付费时价格为每月 20 美元或 100 美元。

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.

消费端和团队订阅均采用不透明的基于使用量的令牌消耗机制。半分析团队对等效 API 令牌支出的研究显示，相比 API 上的同等任务量成本，订阅提供了更多使用额度，但 Anthropic 对何时以及何种程度的使用可能被限速拥有最终决定权。

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.

然而，对于企业订阅，订阅价格中不包含任何使用额度。在团队计划超过 150 个席位后，您需要支付每月 20 美元的费用，再加上所有使用量的 API 费率。API 业务按每百万个令牌计费。混合定价受模型使用、工作负载类型以及我们下文涵盖的多种其他因素影响。

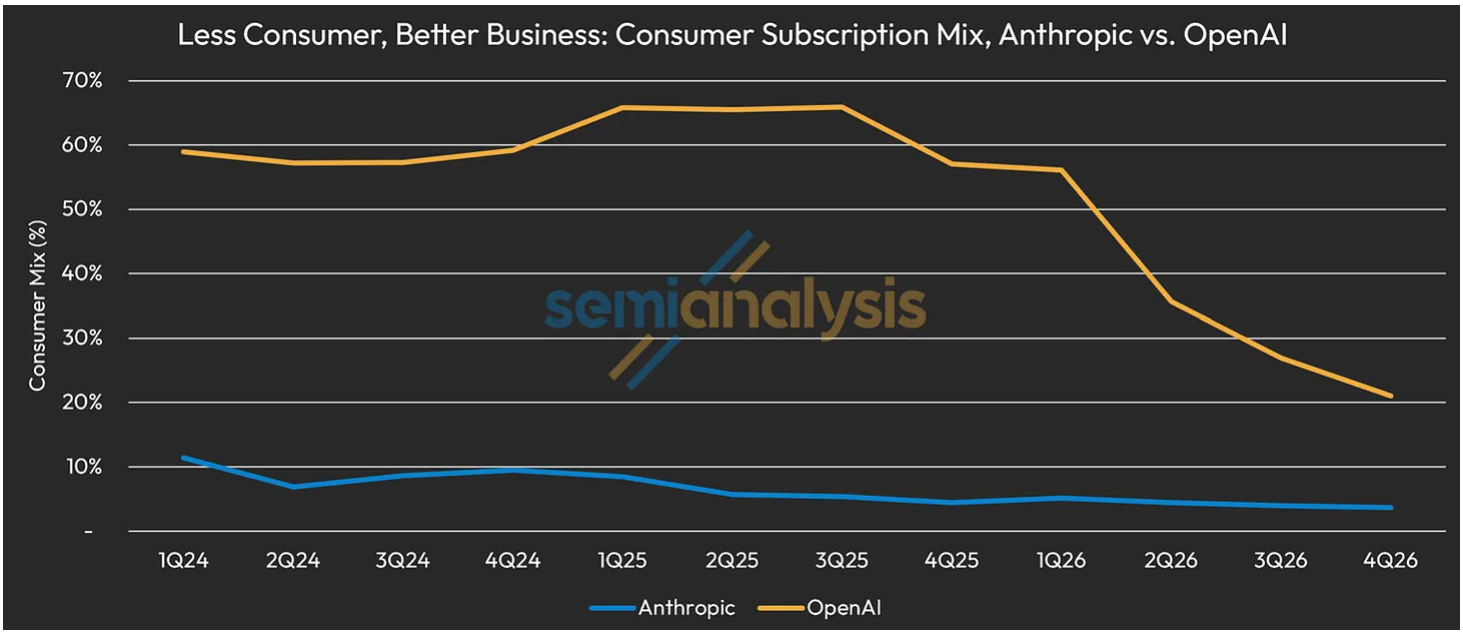
Model	Base Input Tokens	5m Cache Writes	1h Cache Writes	Cache Hits & Refreshes	Output Tokens
Claude Fable 5	$10 / MTok	$12.50 / MTok	$20 / MTok	$1 / MTok	$50 / MTok
Claude Mythos 5 (limited availability)	$10 / MTok	$12.50 / MTok	$20 / MTok	$1 / MTok	$50 / MTok
Claude Opus 4.8	$5 / MTok	$6.25 / MTok	$10 / 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 来源：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

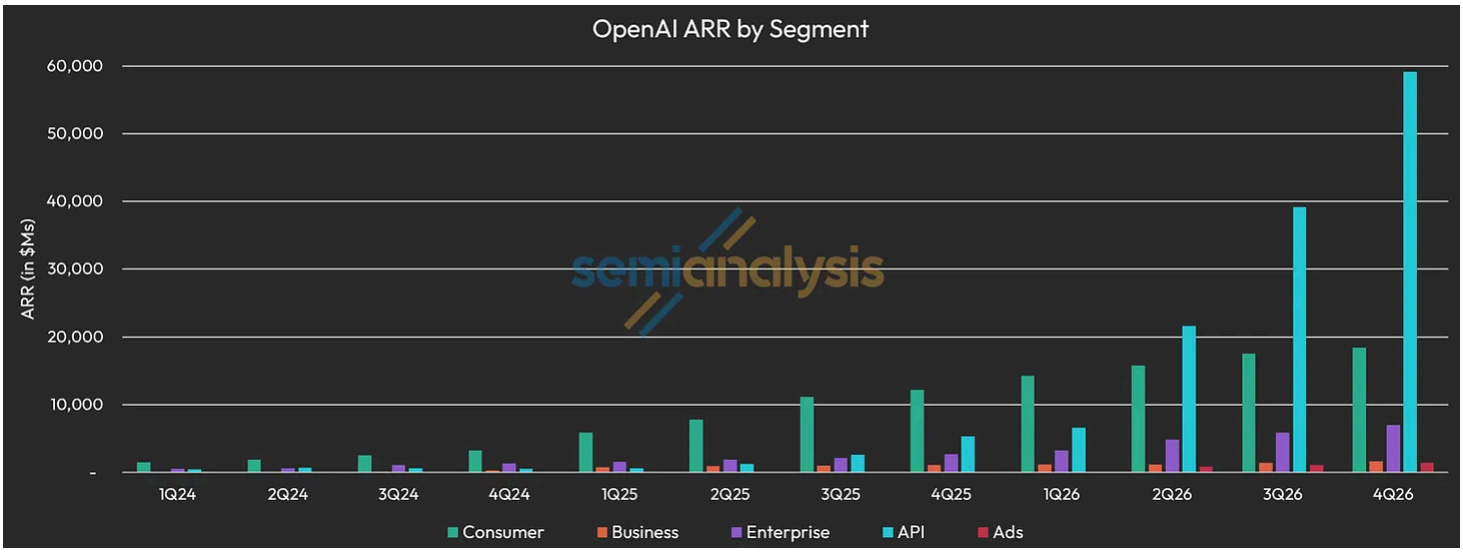
in the margin section below, margins on the API business are vastly superior to consumer subscriptions.

如今，Anthropic 绝大部分年度经常性收入（ARR）均来自 API 业务，订阅收入仅占总 ARR 的 15%。与 OpenAI 对比可见显著差异——2026 年第一季度，OpenAI 超 65%收入源于订阅。企业端与消费端的分化更为悬殊：当前 Anthropic 消费端订阅仅占 ARR 的 5%，而据我们估算，截至 2026 年第二季度末，OpenAI 该比例约为 40%（OpenAI 首席商务官在博客中透露，2026 年第一季度 B2B 业务约占总 ARR 的 40%）。Claude Code 的成功使 Anthropic 顺应市场风向并优化商业模式，如后文利润率章节所示，API 业务的利润率远优于订阅业务。



Source: SemiAnalysis Tokenomics Model

来源: SemiAnalysis 代币经济学模型



Source: SemiAnalysis Tokenomics Model

来源: SemiAnalysis 代币经济学模型

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.

通过聚焦 API 收入，单个用户能带来的收入没有上限，因为计费方式是按 token 而非按席位。当同一用户/客户采用更多智能体工作流时，他们会消耗更多 token 并产生更多收入，而 Anthropic 无需新增客户。披露信息证实了这一点。Anthropic 首席财务官克里希纳·拉奥在 5 月初的《Invest Like the Best》播客中透露，净美元留存率（衡量与公司合作至少一年的客户支出增长指标）达到 500%。这意味着 Anthropic 2026 年第一季度 300 亿美元 ARR 中，有 120 亿美元来自 2025 年第一季度仅贡献 20 亿美元 ARR 的客户群体。而过去一年新增的 180 亿美元 ARR 客户仍在持续增加支出，推动 Anthropic 实现更高增长。

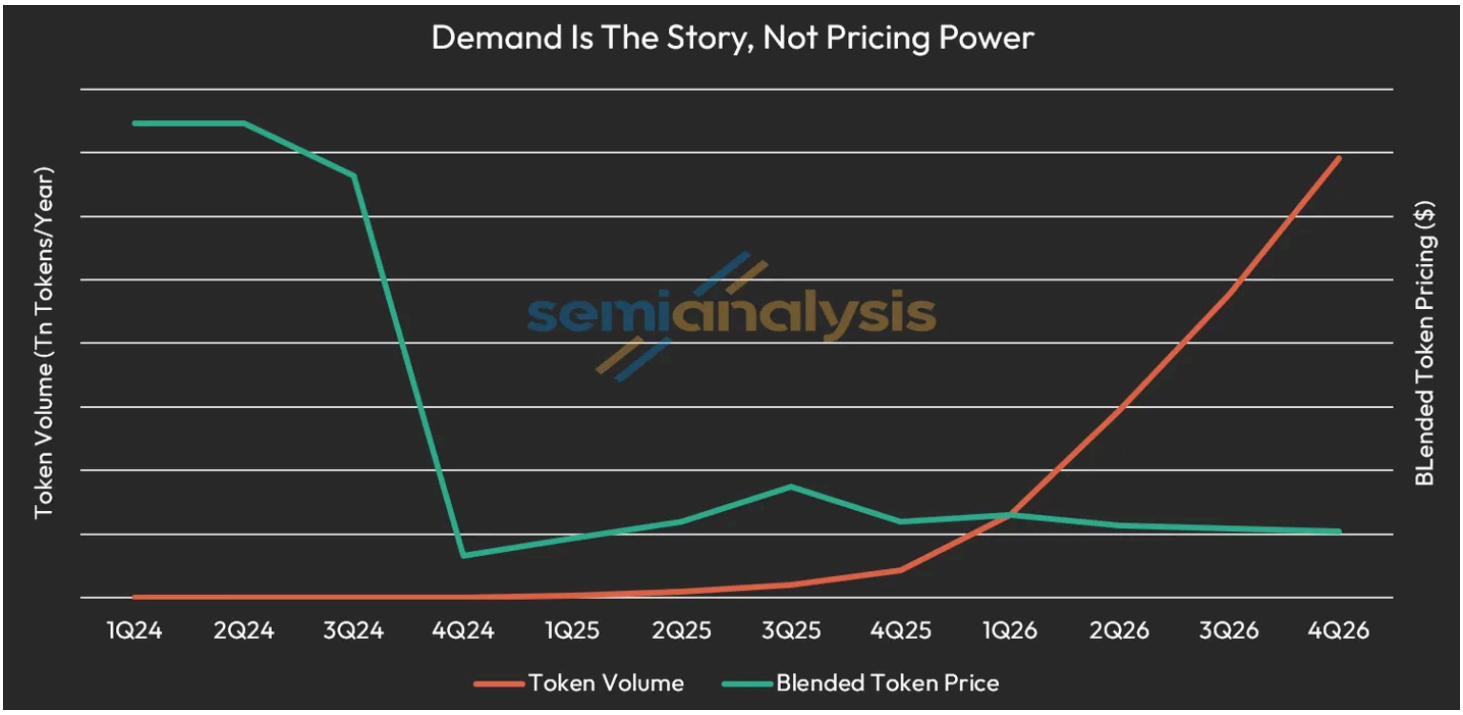
OpenAI is improving, however. The SemiAnalysis Tokenomics team believes that API has been responsible for driving the reacceleration in OpenAI ARR since the March release of 5.5 and Codex and B2B API are now the majority of monthly NNARR. This revenue is higher in margin and could help OpenAI reach profitability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

不过 OpenAI 也在进步。SemiAnalysis 代币经济学团队认为，自 3 月发布 5.5 和 Codex 以来，API 已成为 OpenAI ARR 重新加速的核心驱动力，B2B API 目前占月度新增 ARR 的大部分。这类收入利润率更高，可能帮助 OpenAI 比预期更早实现盈利。

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.

AI 实验室的推理业务模式相当简单。营收由代币量和混合代币价格决定。代币量激增，但混合代币价格随时间推移而下降。既然新模型代币定价更高，混合定价为何会下降？缓存率、输入/输出比率、以及代理与聊天工作流的组合，都是决定混合代币价格的因素。代币经济学模型会跟踪并预测每个产品（如消费者 Max 版、免费版、API 等）和每个模型的定价。随着代理工作被广泛采用，营收增长由代币消耗量的增加驱动。代理工作流消耗的代币量远多于传统聊天功能，而自 2024 年以来，混合代币价格实际上已大幅下降。因此，消费曲线成为推动营收增长的关键因素。



Source: SemiAnalysis Tokenomics Model

来源: SemiAnalysis 代币经济学模型

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

来源：SemiAnalysis 代币经济学模型，Anthropic 选项卡

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.

大体上，AI 实验室的商业模式是：投资建设新模型，购买或租赁算力来训练和运行这些模型，然后通过 Token 的方式将算力出售给客户。上文已显示，Anthropic 和 OpenAI 在推理计算的销售方式上历来截然不同。Anthropic 完全将其商业模式与高利润、基于使用量的定价相结合，而 OpenAI 则服务于成本高昂的免费用户和订阅用户，这些用户可能会消耗大量 Token，占用大量算力。尽管 OpenAI 正逐渐向 Anthropic 的模式靠拢，其增量年度经常性收入（ARR）更多来自 B2B 和 API 业务，但其财务报表仍受到消费者和订阅模式使用量的拖累。

Trajectory & and Discussion of Full P&L

轨迹与完整损益表讨论

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

2026 年第一季度，Claude Code 的爆发式增长使 Anthropic 的年化经常性收入（ARR）从 2025 年底的 90 亿美元飙升至 300 亿美元，仅用了一个季度。一月份，公司新增了 30 亿美元的 ARR，二月份新增 70 亿美元，三月份新增 110 亿美元。虽然消费者业务得益于一季度"战争部"事件的拉动，但 API 才是真正的增长引擎。尽管如此，在消费者领域，我们看到 Anthropic 仍拥有强劲的业务。目前月活跃免费用户（Free MAU）可能稳定在 5500 万至 6000 万左右，其中 9%的消费者月活跃用户以某种形式为服务付费。这与 OpenAI 形成对比，后者自年初以来周活跃用户（WAU）一直停滞在 9.5 亿左右，且付费用户占比更低，目前约为 6%。

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
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

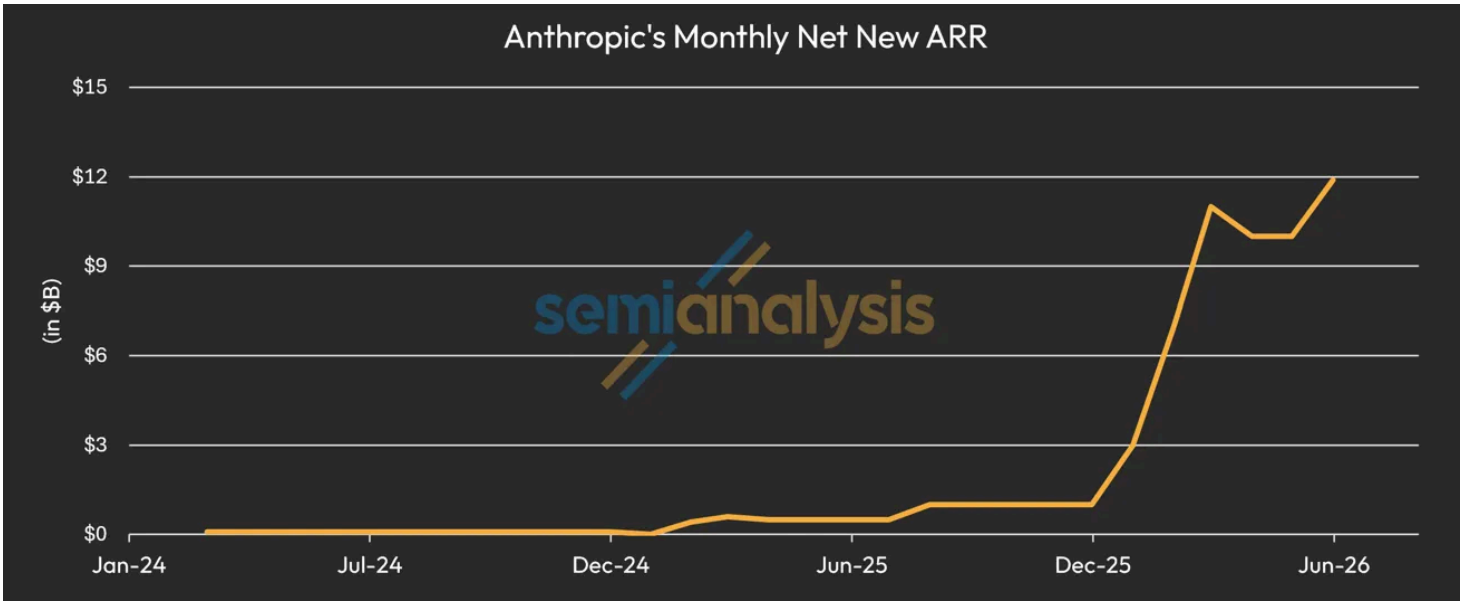
semianalysis

Source: SemiAnalysis Tokenomics Model, Anthropic Tab

来源：SemiAnalysis 代币经济模型，Anthropic 标签

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.

这便是我们想彰显的鲜明反差：OpenAI 主攻消费级市场，而 Anthropic 则深耕企业级领域。尽管 OpenAI 开拓了广告等面向消费者的新业务线，但我们认为短期内其对行业总 ARR（年度经常性收入）的贡献微乎其微。真正的机遇在于 B2B 市场，尤其是基于使用量的定价模式。目前，Anthropic 约有 75%-85%的 API 支出源于消费驱动，随着年内推进，其净新增 ARR 已稳定在每月 100 亿美元以上的区间。而 Fable 作为新一轮增长催化剂，通过更高定价机制，配合最新模型版本解锁的全新工作负载与规模扩展，将成为净新增 ARR 的下一个加速器。本文后续将重提此论点——显然，前沿模型版本迭代能加速 ARR 增长，Anthropic 必须持续巩固其在智能体 API 工作负载领域的优势。我们认为，其损益表结构与商业模式完全支撑这一战略路径。



Source: SemiAnalysis Tokenomics Model

来源: SemiAnalysis 代币经济学模型

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.

再往下看损益表，混合毛利率是关键指标。各实验室通过租赁或自购计算资源，在此基础上销售订阅服务或按用量计费的 Token。实验室有多种手段来优化给定算力的推理效率：首先，新模型通常定价更高；其次，不同加速芯片的 Token 吞吐量存在差异，且可通过持续优化更好地服务推理需求。Token 是 LLM 处理数据时衡量 AI 工作负载的基本单位——可以理解为查询语句的片段，在文本查询中可能是一个词、词的一部分或标点符号。

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.

这些计算成本在单位算力上基本固定，因此当实验室能做到以下两点时——1）在单位算力内产出更多 Token，或 2）提高单位算力上每个 Token 的定价——其边际利润率就会趋近 100%。正因如此，当前混合毛利率维持在 65%左右，而 2024 年同期则为-94%。

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

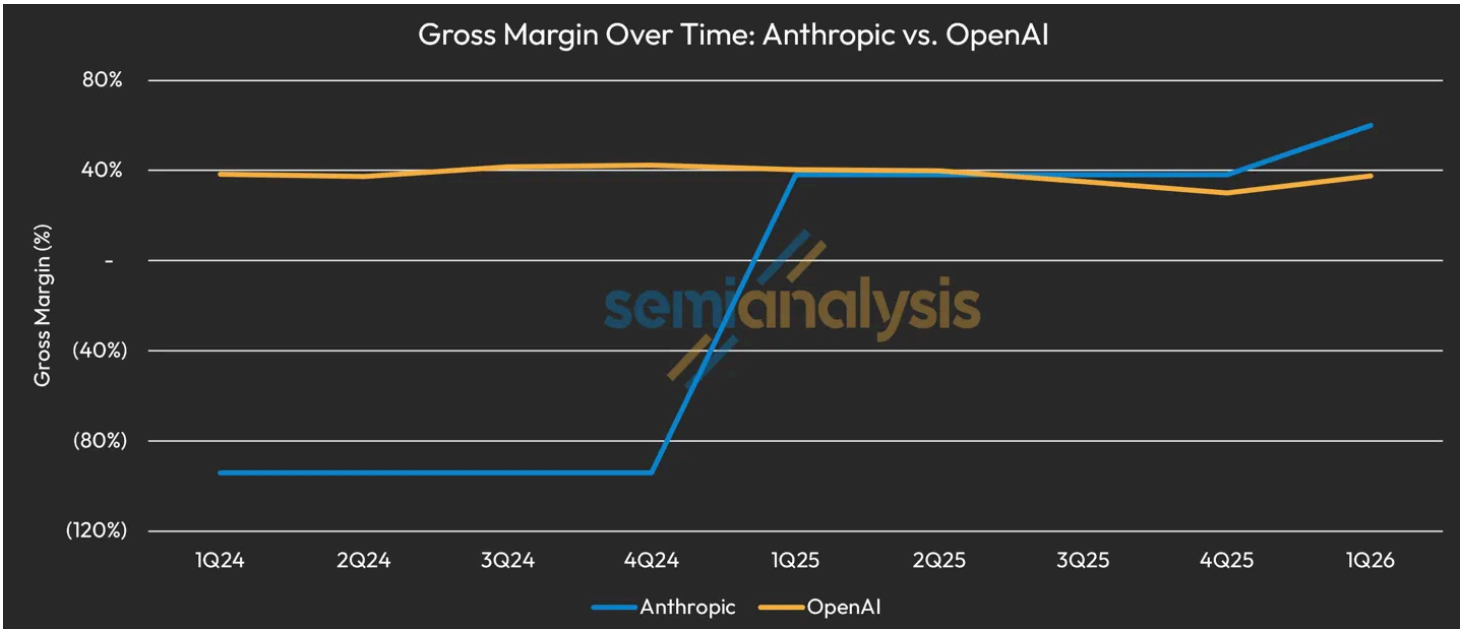
fixed, leading to massive margin upside when revenue growth is strong.v

在过去 12 个月中，各大实验室在推理效率方面取得了显著进展，这使得每兆瓦计算能力的年化经常性收入（ARR）大幅提升，利润率也随之增长。今年晚些时候，Anthropic 每兆瓦计算的 ARR 将达到 6000 万美元左右。而就在 9 个月前，这个数字还只有 1600 万美元。这正是实验室财务数据中利润率强劲增长的体现：在计算成本基本固定的情况下，令牌吞吐量可以持续优化并实现更好的变现，当收入增长强劲时，就会带来巨大的利润率提升空间。

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

来源：SemiAnalysis 代币经济学模型，Anthropic 数据表



Source: SemiAnalysis Tokenomics Model

来源: SemiAnalysis 代币经济学模型

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

OpenAI 再次处于对 Anthropic 的不利地位。该公司需要支持超过 9 亿的免费用户，据我们估算，每位用户的月服务成本略低于 0.70 美元。这在历史上相当于总毛利率的 20%至 30%的侵蚀，这意味着如果 OpenAI 和 Anthropic 均拥有 1000 亿美元的 ARR，OpenAI 的毛利润将比 Anthropic 少 250 亿美元。这部分毛利润原本可以重新投资于业务，用于训练计算资源或最新模型的研发。这将形成一个良性循环：新模型能够推动 ARR 和毛利润增长，进而为下一代模型的开发投入更多资金，从而在与竞争对手的较量中占据优势。

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

来源：SemiAnalysis 代币经济模型，OpenAI 标签页

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.

不过，毛利率可能正在发生变化。Anthropic 以高于市场水平两倍以上的价格租用新算力（参见近期与 SpaceX 的交易），且更多年度经常性收入来自 AWS Bedrock 等渠道，公司需为此支付 20-30%的收入分成。但由于 API 毛利率超过 80%，在可预见的未来，Anthropic 可以轻松维持甚至提升当前约 65%的毛利率水平。

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.

在毛利润之下，AI 实验室需要承担训练/研发成本及其他运营支出。我们最终想表明的是，Anthropic 凭借其高利润率、按使用量计价的商业模式拥有巨大优势，这种模式能够将极高比例的收益再投资于下一代最先进模型的训练，从而不仅拉开与 OpenAI 的距离，更能领先于所有其他封闭实验室和开源竞争对手。我们估计，Anthropic 目前超过 60% 的计算资源专用于训练和内部用途。根据我们的长期预测，Anthropic 至少会将 25% 的收入用于训练计算，到 2030 年，训练与推理计算将分别占公司整体计算资源的 48% 和 52%。Anthropic 目前的再投资率远高于 OpenAI，我们预计其盈利能力将为其打开更多大门，例如获得更低的利率以及以更优条件获取融资，从而继续推进大规模计算基础设施建设，以服务于 2030 年及更远的未来。

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.

Anthropic 在 2026 年第三季度实现了 10 亿美元的 GAAP 息税前利润（6%利润率），且净新增年度经常性收入增速超过 OpenAI——后者息税前利润率为-100%，这形成了鲜明对比。我们推测 Anthropic 可能并不追求盈利，但算力限制制约了其能再投资于训练和新的算力交易的能力。

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.

总之，毛利润率和息税前利润率的显著差异，使 Anthropic 相比 OpenAI 拥有明显的再投资优势。我们将在《算力与长期实验室经济学》部分进一步探讨这一现象。Anthropic 证明，AI 实验室只要专注运营并高效执行，就能成为利润丰厚的优质业务。我们预计，当它们上市时，投资者将给予其充足的资本渠道，以资助下一阶段的算力建设。而对 OpenAI 来说，关键在于继续将早期 Codex 与 API 的动能转化为更优利润率，从而将资金再投资于模型开发，并为其算力建设铺就一条通往更优盈利能力的道路。

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

SemiAnalysis 是一份由读者支持的刊物。如需接收新文章并支持我们的工作，可考虑成为免费或付费订阅用户。

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

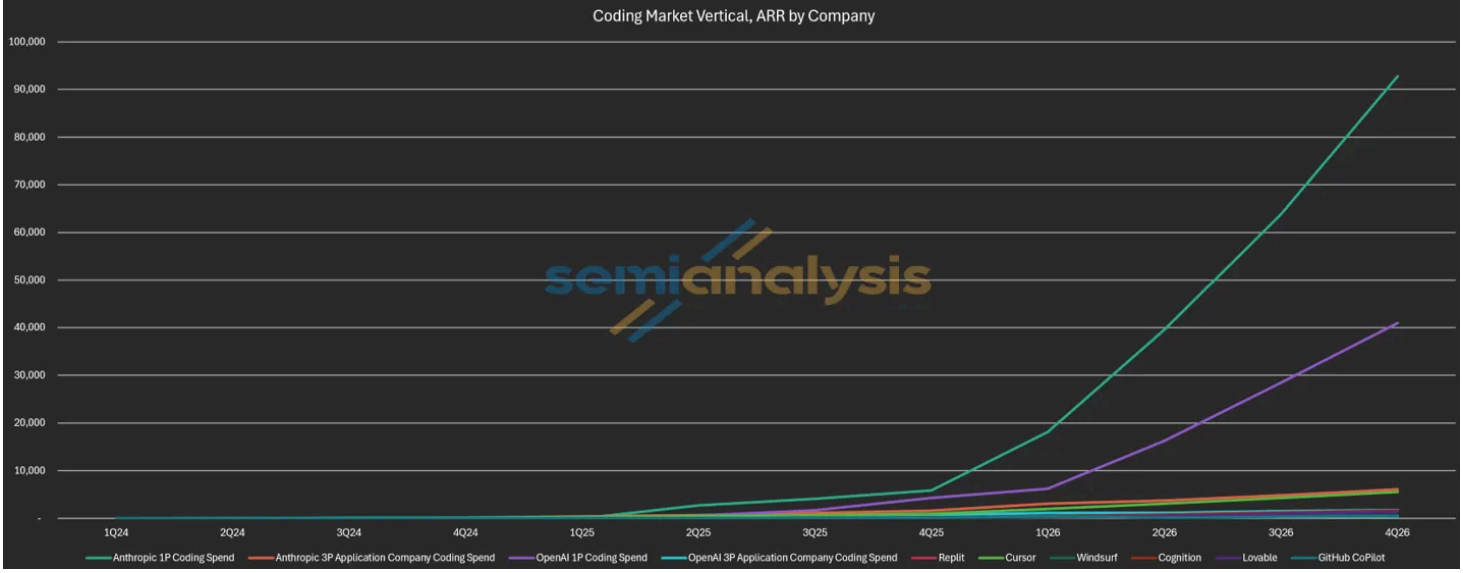
来源：SemiAnalysis 代币经济模型，Anthropic 选项卡

Coding Market, Large Customers, and Other

编程市场、大客户及其他

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.

当前，实验室收入主要由大型客户中的高价值用户驱动。也就是说，支出范围较窄且高度集中于编码工作流，但其基础足够广泛，我们相信 Meta 是最大的客户，尽管其收入仅占 Anthropic 总收入的 3-5%。Tokenomics 团队估计，目前实验室超过 65%的 ARR 来自编码用例，而 Cursor、Cognition、Loveable、Replit 等其他编码初创公司，截至 2026 年第二季度，其 ARR 合计增加了 60 亿美元。



Source: SemiAnalysis Tokenomics Model, CodingMarket Tab

来源：SemiAnalysis Tokenomics 模型，编码市场选项卡

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.

尽管包装类公司数量众多，但随着行业增长范围的扩大，它们目前仅占实验室总 ARR 的 10-15%。大型企业客户数量众多，但支出规模仍在逐步提升。Meta 作为 tokenmaxxing 最受瞩目的案例，很可能是 Anthropic 最大的客户，其支出规模高达九位数。

TaaS, Hyperscalers, And Sales

TaaS、超大规模云服务商与销售

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.

近期，按量计费的代币服务业务模式迅速崛起。对客户而言，这意味着通过外部供应商 API 购买模型，而非直接从实验室获取。客户可以购买该供应商提供的任意模型，并通过 API 用量进行消耗。我们的分析显示，Anthropic 目前通过 AWS Bedrock 和 Azure Foundry 等间接渠道服务了 15%-20%的年度经常性收入（ARR），而一个季度前这一比例仅为 5%-10%。

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:

随着 Claude 在 Bedrock、Foundry 和 Gemini Agent 企业平台（原 Vertex）上架，三大超级云服务商拥有大型企业青睐的三大优势：

- Access and flexibility to spend on any model

  可灵活访问任意模型并自由分配支出

  - 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

    如果 OpenAI 在六个月内为特定工作流开发出更优模型，客户可以轻松转移支出，不会受限于单一模型供应商

- Ability to use committed usage and ELAs towards any service from the Cloud provider, including any model within a TaaS product

  能够将承诺使用量和企业许可协议（ELA）用于云服务商的任何服务，包括 TaaS 产品中的任何模型

- Existing relationships with strong security and compliance frameworks that enterprises trust

  与现有安全合规框架深度绑定，这些框架深得企业信任

  - No need to qualify a new vendor directly in a 6–12-month RFP

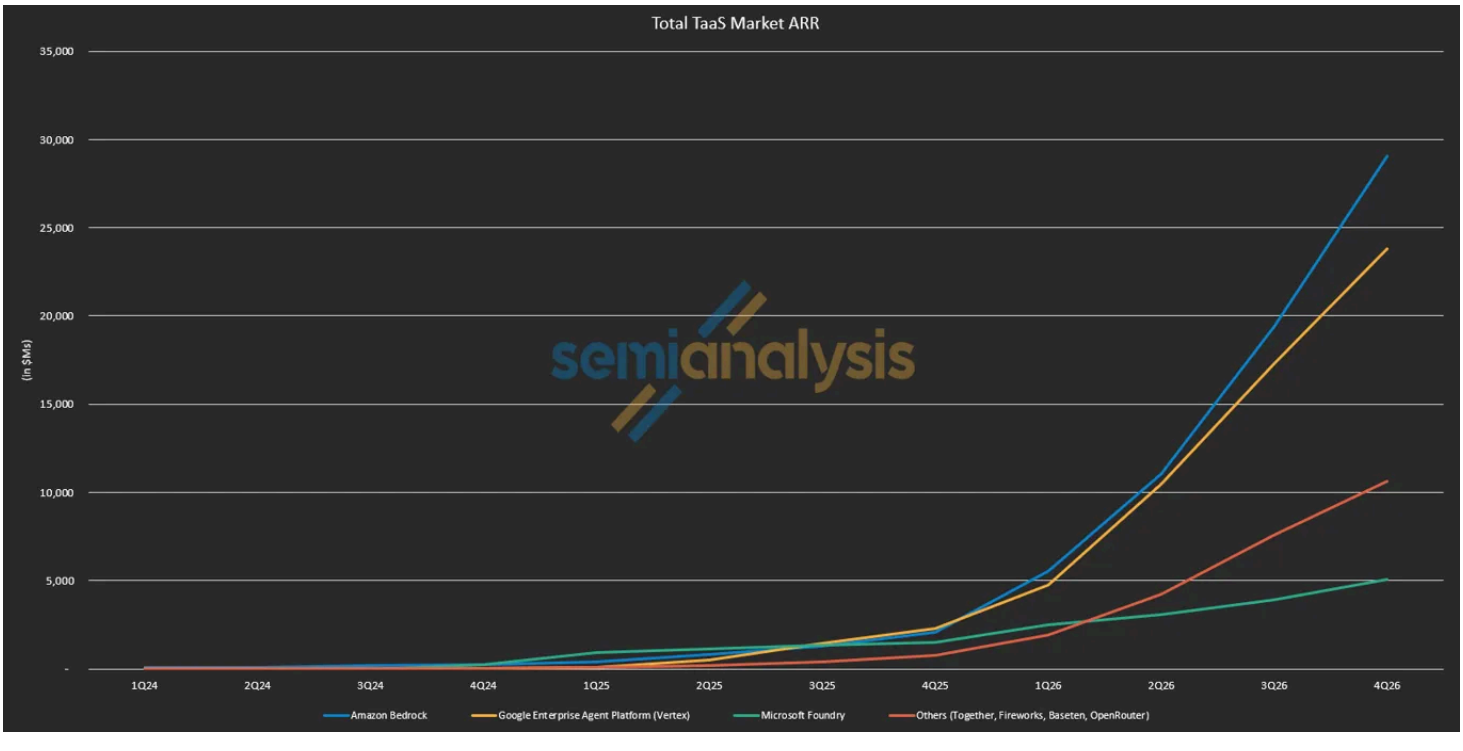
    无需在 6-12 个月的招标流程中重新认证新供应商

- Better latency for international workloads (Anthropic routes all direct inference queries back to the US)

国际工作负载延迟更优（Anthropic 将所有直接推理查询路由回美国处理）

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.

那么，这对 Anthropic 是否更有利？前提是：1）企业能轻松切换模型；2）你需为这种灵活性支付收入分成。我们认为是的。TaaS（模型即服务）参与者为实验室提供了资本灵活性，其 IaaS（基础设施即服务）费用条款比直接提供推理服务更优惠，并让实验室得以接入这些巨头企业庞大的客户群和云资产。企业级销售难度极大，而超大规模云厂商在这方面具有优势：数千亿美元的企业客户年支出，使 Anthropic 和 OpenAI 更容易触达潜在客户。在许多情况下，将 20%的收入分成给超大规模云厂商以换取其分销优势，比自建 BDR（业务拓展代表）、企业销售代表、销售工程师和客户成功团队来管理获客、扩张与留存更为明智。



Source: SemiAnalysis Tokenomics Model, TaaS Market Tab

来源：SemiAnalysis 代币经济学模型，TaaS 市场表格

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.

企业对 TaaS 业务的采纳势头强劲。据二季度估算，该市场规模已达 280 亿美元年化经常性收入（ARR），其中三大云巨头占据 85%市场份额。基于上述因素，我们认为 TaaS 将成为未来极具潜力的商业模式，而 Anthropic 的 ARR 中将有越来越大的比重来自这种间接合作关系。

Risks & Upside 风险与上行空间

We see several risks that investors should watch for:

我们认为投资者应关注以下几项风险：

- Top Line Risks 收入端风险
  - Tokenbudgeting Concerns 代币预算方面的隐忧
  - OSS and Price Cuts 开源模型与降价策略
    - OpenAI has been rumored to cut token pricing to try to win back spend

      据传，OpenAI 正在削减 Token 定价，试图赢回客户支出
  - Competitive Coding Models from Google and Meta

    来自谷歌与 Meta 的竞品编程模型
  - New Model Blockades 新模型壁垒
    - 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

一种禁止模型发布、放弃 AI 领导地位、并允许开源模型追赶当前 Fable 模型能力的美国监管体制，几乎完全消除了 Anthropic 固有的商业优势。

- Compute Shortages 算力短缺
  - Where will the supply come from?

    供应将从何而来?

- Margin Risks 利润率风险
  - Margins from more TaaS mix

    更多 TaaS 组合带来的利润率提升

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.

Token 预算规划已成为实验室年度经常性收入（ARR）趋势中的最新关注点。当我们团队审视像 Coinbase 这样广为人知的 Token 预算规划案例时，我们发现这家公司裁员 14%的主要原因是其终端市场出现了大规模周期性熊市，而非单纯基于投资回报率的考量。正如我们在近期关于 TokenMaxxing 与 TokenBudgeting 的研报中所指出的，全球 2000 强企业在 AI 方面的支出，相较于其整体 IT 投入，在实验室 ARR 中所占的营收基数仍然微乎其微。Anthropic 官方帮助页面显示，其 Claude Code 企业用户的平均月支出仅为 150-250 美元，且 90%的用户每日支出低于 30 美元。我们在此重申：Tokenmaxxing 的故事属于分布曲线的尾部，而那些持续高额支出的尾部用户，其投入正收获着惊人的投资回报率。



TokenBudgeting: Our Conversations with Enterprises on Token Spend

Token 预算管理：我们与企业关于 Token 花费的对话

CRYSTAL HUANG, JOEY BROOKHART, AND DYLAN PATEL

黄水晶（CRYSTAL HUANG）、乔伊·布鲁克哈特（JOEY BROOKHART）与迪伦·帕特尔（DYLAN PATEL）

7月1日 · 7 月 1 日

Read full story 阅读完整报道 →

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.

企业在使用广度和深度上仍有广阔且未被充分开发的空间。企业决策者倾向于偏好可预测的成本结构。我们预计多数岗位将设置使用限额/预算，但随着投资回报率的提升，这些单座预算也将相应增长，而能创造巨大投资回报率的顶尖人才将获得近乎无上限的预算支持。

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.

即便任务开始向开源模型或低于最新 SOTA 水平的模型转移，只要前沿领域仍保持双雄争霸格局，Anthropic 的代币定价与使用量仍将居高不下。代币定价通缩必须通过前沿 SOTA 模型数量的增加来驱动，而非依靠客户自然竞争动态将工作负载转移至低价模型——当竞争对手正在使用最新智能时，哪个管理团队会让优质用户使用上一代智能？若以近期发布的 Sonnet 5 为例，即便较低层级的闭源模型实际上也具备比前代产品更高的定价能力。我们将此归因于压倒性的需求因素。当前前沿市场格局过于狭窄，经济体系中 95%以上的其他领域（即非编码领域）的新工作流必将接棒，通过定价与使用量双轮驱动，让前沿模型始终占据金字塔顶端，助推 Anthropic 的 ARR 再创新高。

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 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.

其他实验室的竞争，最有可能对编程市场造成冲击。Meta 超级智能和谷歌 DeepMind 新模型的成功，将成为决定 Token 价格下行以及随之而来的毛利率下降的关键因素。如果 Meta 和谷歌能将编程市场变成四强争霸的局面，Token 定价显然会受到影响。这是否意味着 Anthropic 的 NNARR 会转为负值？可能不会。最新、最优秀的智能永远有其价值，工作负载的扩张和 Anthropic 的新模型将迫使谷歌和 Meta 必须追赶下一轮技术前沿。Anthropic 乃至所有主要实验室，始终需要在闭源与开源领域竞相超越对手。因此，只要新模型能力持续提升，且 SOTA（最先进水平）与开源能力之间的差距足够大，Anthropic 的定价权和使用量增长法则就能保持优势。

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

or paid subscriber.

SemiAnalysis 是一份由读者支持的刊物。如需接收新文章并支持我们的工作，可考虑成为免费或付费订阅用户。

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.

接下来谈风险：更多《寓言》式的政府干预可能会拖累前沿实验室。当政府和安全专家限制最新模型的使用人群时，模型的投资回报率会怎样？Anthropic 最大的风险是否来自其本国政府？虽然监管锁定对某些企业有利，但当你拥有资金充裕的超大规模实验室在奋力追赶，而中国实验室又在蒸馏你的模型时，情况就截然不同了。前沿实验室希望延长开源及其他实验室达到同等模型能力水平所需的时间。政府在变现前沿技术上的拖延，可能会动摇美国在人工智能领域的领先地位，以及美国在全球科技领域更广泛的主导权。尽管我们认为目前这种风险很小，但仍敦促监管机构审慎权衡利害关系。

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

而在增长方面，我们看到了多重利好因素：

- New Customers Ramping Seats and Usage throughout 2H26 and 27

  新客户在 2026 年下半年及 2027 年持续增加席位与使用量

- Cyber with Mythos/Fable to be the next major vertical to power growth

  网络安全领域，Mythos/Fable 将成为下一个主要的增长方向。

- With growth ramp faster the Coding

  随着增长加速，编程领域表现尤为突出

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.

我们的内部观点是，无论是从每百万 Token 的定价组合角度，还是从用例扩展的角度来看，Fable 都将有助于推动 Anthropic 在 2026 年下半年将月度 NNARR 超越当前每月 100 亿美元的增长速度。

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, 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.

过去一年，年消费超 10 万美元的 Claude 客户数量增长了 7 倍，年消费超 100 万美元的客户在过去两年间更是激增约 42 倍。回顾 2022 至 2023 年的 SaaS 与云优化阶段，我们看到了倍数与增长率可能遭受的冲击。当时的 Snowflake、Datadog、Cloudflare 等企业，其交易倍数高得离谱（远期营收可达 50 倍以上！），而增长速度却急剧放缓。短期内我们并未看到类似优化对行业构成重大威胁。在全球 IT 支出占比极高的全球 2000 强/财富 500 强企业中，前沿 AI 产品的使用仍处于萌芽阶段，因此席位部署量与人均使用量将是 2026 下半年及 2027 年 NNARR 增长的关键驱动因素。此外，新应用场景正不断涌现——我们预计 Claude Code 对增长的推动作用，将在 Mythos 和网络安全领域以更高强度、更快速度重现。医疗、金融、生物科技等垂直行业，都将迎来其 TAM 工作量的大规模扩张，并在增长曲线上呈现激进的上扬态势。

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.

在 SemiAnalysis，这些风险与回报因素使我们保守地预测 NNARR 将在 2026 年下半年及 2027 年加速增长。若 Anthropic 能将月均 NNARR 提升至 150 亿美元，2027 年实现 3000 亿美元的年终 ARR 需求并非难事。

Stead State Margins 稳定状态利润率

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.

Anthropic 的长期利润率将体现其产品与商业模式的最佳水准。如同昔日最优秀的软件公司，实现 30-40%的息税前利润/自由现金流利润率是可行的。随着毛利率攀升至 75%中段、训练/研发支出占总计算成本的 48%、其他运营支出占营收的 20%，AI 实验室已展现出卓越的商业模型，而 Anthropic 将不断证明这一点。

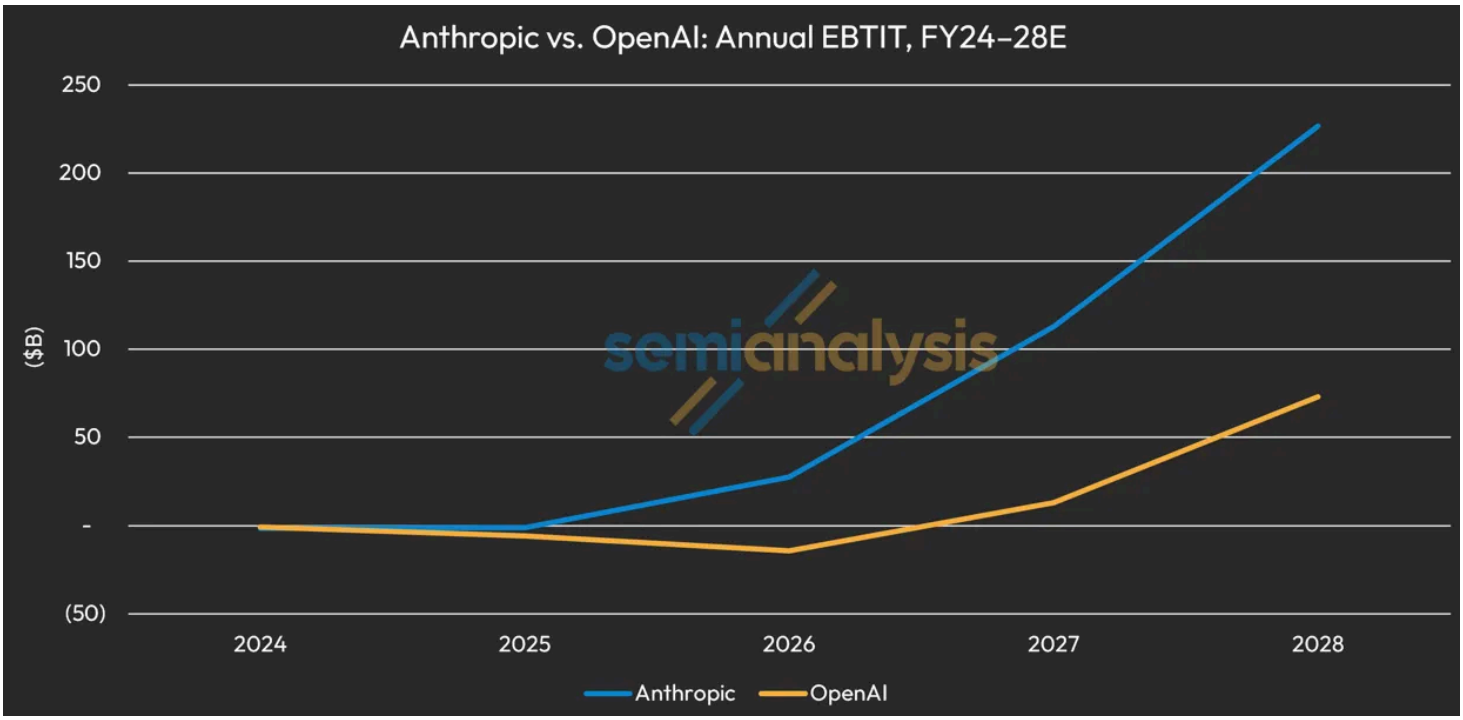
Compute & Long Run Lab Economics
计算与长期实验室经济学

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.

Anthropic 的优势在于其毛利率（这本身取决于其 API 使用收入的组合结构）。实验室在上市时需展示盈利水平，但 Anthropic 拥有更高的营收基础和显著更高的毛利率，可用于投入下一代最先进模型的研发。从绝对值来看，我们预计 Anthropic 在 2027 年扣除销售成本后拥有 1600 亿美元的可再投资资本，而 OpenAI 则为 920 亿美元。

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.

训练、利息与税收前收益（EBTIT）是衡量实验室再投资的新基准指标，也是了解推理业务现金经济状况的最佳数据源。Anthropic 在此项指标上表现出色，2026 年第二季度实现 36%的 EBTIT 利润率，证明 AI 可以带来丰厚利润。然而，更深远的意义在于：每超越 OpenAI、DeepMind、Meta 超级智能以及整个新实验室和开源社区一美元的 EBTIT，都有助于巩固其在模型能力上的领先地位。这种模型能力优势既延长了非前沿模型的追赶周期，也赋予了处理前沿工作负载时的定价权。截至 2028 年，相对于 OpenAI 累计高达 2500 亿美元的 EBTIT 优势，正是我们相信 Anthropic 应当率先 IPO，并尽可能快速获取最大算力的关键原因。



Source: SemiAnalysis Tokenomics Model

来源: SemiAnalysis 代币经济学模型

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.

我们看到模型改进的持续上升趋势毫无停滞迹象，既然如此，为何要停下？自 2024 年第一季度以来，累计 80 亿美元的训练成本已催生出一个年经常性收入达 600 亿美元的业务——这是一项极其惊人的投资资本回报率。不难想象，随着模型能力不断提升，网络安全、生物科技/医疗健康、金融等新垂直领域的 S 曲线叠加效应将如何推动实验室的年经常性收入增长，而高利润推理 API 收入的飞轮效应将为下一代模型拓宽智能差距提供资金支持。Anthropic 若善用其资本优势，将有机会在这些新垂直领域进行最大规模投资，并率先抢占这些市场。

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.

然而，这一切的前提是有足够的算力来支撑需求。我们预测，到 2030 年，各大实验室对算力的无约束需求将超过 100GW。资本必须满足这一需求，而这正是此次 IPO 更宏大的叙事：今年，一些有史以来最优秀的企业（超大规模云服务商）将发行超过 1000 亿美元的股票，这仅仅是未来 2-3 年内为满足需求而持续大规模建设所需资本的序曲。Anthropic 和 OpenAI 的 IPO 是整个生态系统的转折点，该生态系统每年需要数万亿美元的融资渠道。

Conclusion & Valuation 结论与估值

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.

Anthropic 打造了一家令人羡慕的企业。在增长率和盈利能力这两个最重要的财务指标上，该公司是 AI 实验室竞赛中的明显赢家。然而，OpenAI 正在关键领域迎头赶上。自 5.5 版本发布以来，Codex 的反馈以及我们对 API 企业支出的调研结果都极为积极。OpenAI 通过其 API 业务持续加速 NNARR 的能力，将决定 Anthropic 能够保持多大的领先优势。

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.

因此，我们认为 Anthropic 应当率先尝试 IPO。其财务状况更优，商业模式更佳，且能持续扩大这一优势并吸引资本。这样一来，OpenAI 便面临以更差的财务表现和增长数据来筹集股权的压力，因此 Anthropic 应保持全力推进的态势。

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.

我们的估值基准设定为 2027 年预期年化经常性收入（ARR）3000 亿美元的 20 倍（即 2027 年每月净新增 ARR 约 150 亿美元）。这将使 Anthropic 的估值达到 6 万亿美元——成为全球市值最高的企业。

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

我们每月按业务线、代币、利润率和算力层级更新对 Anthropic、OpenAI 及超大规模云服务商的估算。如需获取我们的估值模型、研究报告及观点见解，请发送邮件至 sales@semianalysis.com。

Recommend SemiAnalysis to your readers

向您的读者推荐 SemiAnalysis

Bridging the gap between the world's most important industry, semiconductors, and business.

连接全球最重要的半导体产业与商业之间的桥梁。

Recommend 推荐

Previous 上一篇

Discussion about this post

关于此帖的讨论

Comments 评论 Restacks

Write a comment...

© 2026 Dylan Patel · Privacy · Terms · Collection notice
Substack is the home for great culture