Anthropic Growth and Bedrock Mix Drive AWS Margins Higher While Peers Lag

Anthropic 增长与 Bedrock 组合推动 AWS 利润率攀升，同行落后

Amazon's Bedrock Mix and Anthropic Deal Terms Combine to Show Greater Operating Leverage

亚马逊 Bedrock 组合与 Anthropic 交易条款共同展现更强运营杠杆

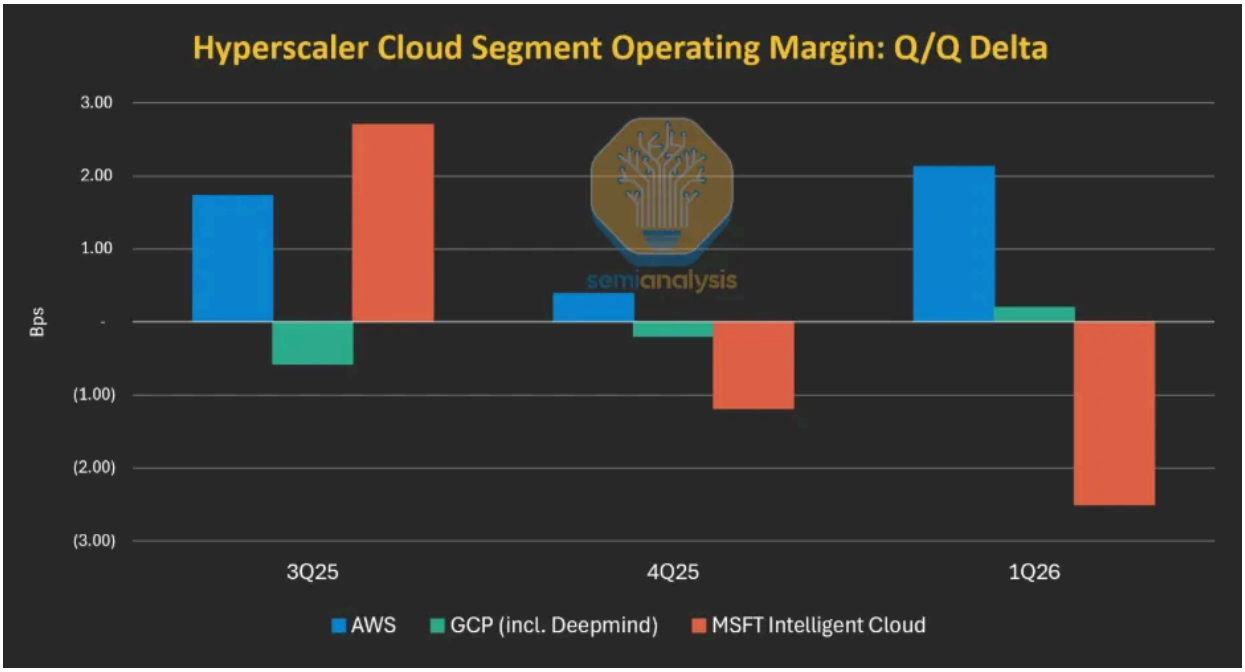
JEREMIE ELIAHOU ONTIVEROS, JOEY BROOKHART, CRYSTAL HUANG, AND DYLAN PATEL

杰雷米·埃利亚乌·翁蒂韦罗斯、乔伊·布鲁克哈特、黄水晶、迪伦·帕特尔

MAY 28, 2026 2026 年 5 月 28 日 · PAID · 付费

While other CSPs have seen declining-to-flat operating margins over the last several quarters, Amazon's AWS margins inflected this past quarter driven primarily by customer spending growth on Claude through Bedrock. AWS' higher share of 3P model API spend, Anthropic/Bedrock deal structure, and Anthropic's ARR outperformance in 1Q26 all contributed to EBIT margins increasing 213bp Q/Q while other CSPs lagged. SemiAnalysis' work in the new Tokenomics 2.0 model shows how AWS has pulled ahead of the pack and found a strong avenue to grow margins. Our model estimates quarterly revenue, profits, ROIC and compute requirements of every single business vertical of hyperscalers and AI Labs, e.g. Gemini API revenue & margins, Microsoft Copilot ARR, OpenAI ChatGPT subscriptions across plans, etc.

过去几个季度，其他云服务提供商的运营利润率持续下滑或持平，而亚马逊 AWS 的利润率却在上一季度出现拐点，主要得益于客户通过 Bedrock 平台对 Claude 的支出增长。AWS 在第三方模型 API 支出中占据更高份额、Anthropic/Bedrock 的交易结构设计，以及 Anthropic 在 2026 年第一季度超预期的年度经常性收入表现，共同推动其息税前利润率环比增长 213 个基点，而其他云服务提供商则表现滞后。SemiAnalysis 在全新的 Tokenomics 2.0 模型中揭示了 AWS 如何领先同行，并找到了提升利润率的强劲途径。我们的模型估算了超大规模云厂商和 AI 实验室每个业务板块的季度收入、利润、投资资本回报率及算力需求，例如 Gemini API 的收入与利润率、Microsoft Copilot 的年度经常性收入、OpenAI ChatGPT 各订阅计划的收入等。

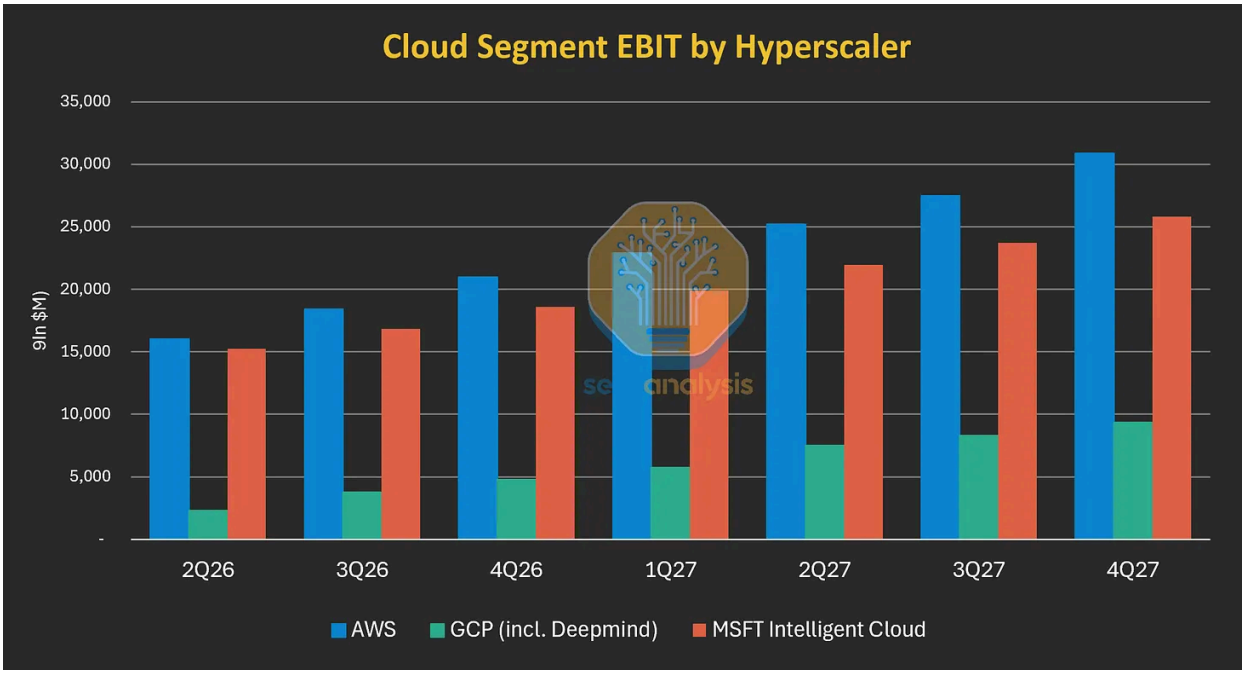


Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis Tokenomics 模型

Although all CSPs are benefiting from increased AI and non-AI revenue, margins are a whole different story. Oracle and Coreweave both disappointed the market with lower-than-expected profits from their cloud arms. Azure is also seeing a downward trend in margins. Google Cloud has had a great upwards climb recently, but margins are inflated since they do not include training costs from DeepMind in the GCP segment. The only CSP with a true rising trend is AWS – a remarkable achievement considering their server depreciation (5yrs) is the lowest of all CSPs.

尽管所有云服务提供商（CSP）都受益于 AI 和非 AI 收入的增长，但利润率却呈现出截然不同的局面。甲骨文（Oracle）和 Coreweave 均因云业务利润低于预期而令市场失望。Azure 的利润率也呈下降趋势。谷歌云（Google Cloud）近期虽大幅攀升，但其利润率存在虚高——因为 GCP 板块并未计入 DeepMind 的训练成本。在所有 CSP 中，唯一呈现真实上升趋势的是亚马逊云服务（AWS）——考虑到其服务器折旧年限（5 年）为所有 CSP 中最短，这一成就尤为瞩目。



Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济学模型

The Amazon Story & Background

亚马逊的故事与背景

We believe that Amazon's margin success rests on a differentiated strategy that will be exploited further in coming quarters and years. The firm was late to wake up to the AI opportunity (we were the first to call out their leadership loss in 2023). Two years later, we were again the first to call out their change in trajectory, an upcoming revenue acceleration, when all the market was labelling them as an AI loser.

我们认为，亚马逊在利润率上的成功源于其差异化战略，这一优势将在未来几个季度乃至数年得到进一步发挥。该公司在 AI 机遇面前曾反应迟缓（我们是首家指出其 2023 年领导力下滑的机构）。两年后，当市场普遍将其视为 AI 领域的落后者时，我们再次率先指出其发展轨迹的转变——即将到来的营收加速增长。

Amazon's AI Resurgence: AWS & Anthropic's Multi-Gigawatt Trainium Expansion

亚马逊 AI 复兴：AWS 与 Anthropic 的多吉瓦 Trainium 扩张

JEREMIE ELIAHOU ONTIVEROS, DYLAN PATEL, AND 2 OTHERS

杰雷米·埃利亚胡·昂蒂维罗斯、迪伦·帕特尔等 2 人

• 2025年9月4日 2025 年 9 月 4 日

Read full story 阅读完整报道 →

Now, we see a new era for AWS where the firm combines accelerating revenue growth AND outperforming margins. Amazon brings a unique combination of the following:

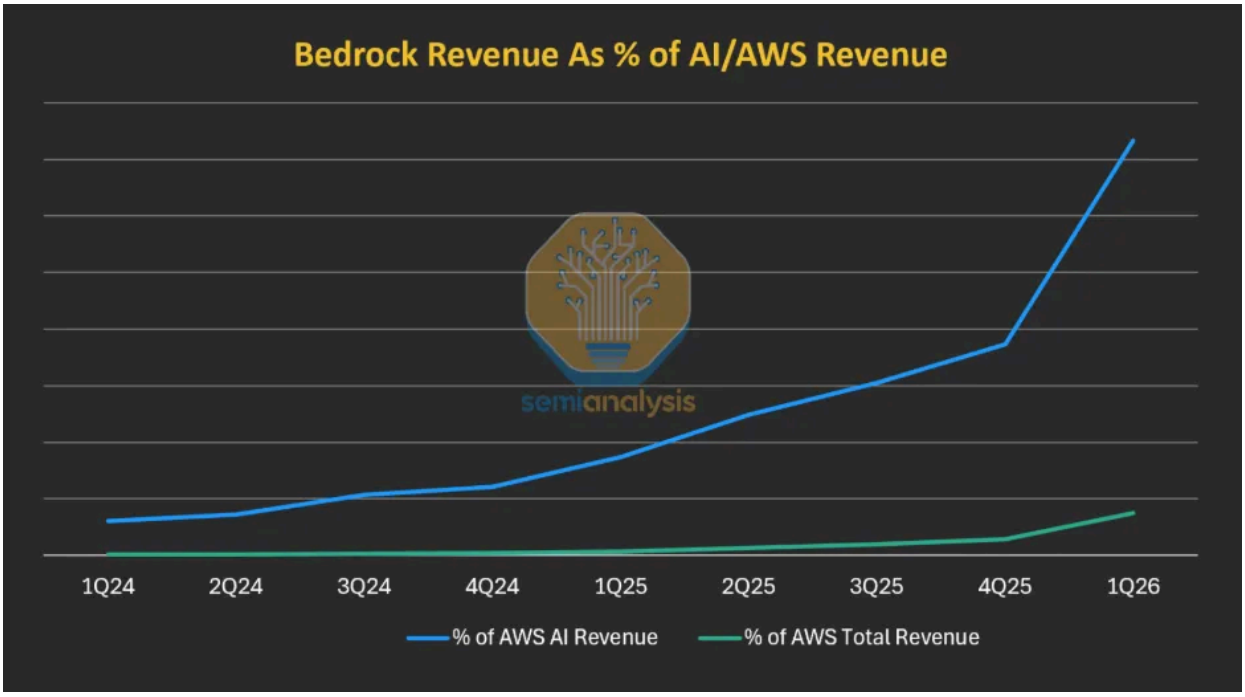
如今，我们见证了 AWS 的新纪元——这家公司正将加速的收入增长与卓越的利润率相结合。亚马逊具备以下独特的组合优势：

- Risk appetite: winners in the AI infrastructure landscape are not afraid of putting their balance sheet at work. As our Datacenter Industry Model demonstrates, Amazon has secured more power than any other cloud provider besides Google, understanding before others that energy drives market share in this constrained environment, and that requires capital and multibillion dollar PPAs.

  风险偏好：AI 基础设施领域的赢家从不畏惧动用资产负债表。正如我们的数据中心行业模型所示，亚马逊已获得除谷歌外所有云服务商中最多的电力资源，它比同行更早意识到：在资源受限的环境下，能源决定市场份额，而这需要资本投入和数十亿美元的购电协议（PPA）。

- Business Model: Amazon is the only CSP with token-as-a-service being the dominant share of its AI business, while all others are focused on multi-year IaaS deals. That demonstrates a higher risk appetite but also a better understanding of the unique opportunity provided by Bedrock, as we'll detail below.

  商业模式：亚马逊是唯一一家以"代币即服务"（Token-as-a-Service）作为 AI 业务主导份额的云服务商，而其他厂商均聚焦于多年期 IaaS 合同。这既体现了更高的风险偏好，也彰显了对 Bedrock 平台独特机遇的深刻理解，下文将详细阐述。



Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济学模型

- Scale & speed: No other provider will build more capacity than AWS in 2025, 2026 and 2027, as per our Datacenter Industry Model. AWS dwarfs rivals. Not only did the CSP procure a lot of power, it has also executed much more rapidly than peers and is rolling out a new datacenter design that will exacerbate its speed advantage.

  规模与速度：根据我们的数据中心行业模型，2025 年、2026 年和 2027 年，没有其他云服务提供商会比 AWS 建设更多容量。AWS 远超竞争对手。该云服务商不仅采购了大量电力，其执行速度也远快于同行，并且正在推出一种新的数据中心设计，这将进一步扩大其速度优势。

- Vertical integration: we were first to cover the growing CPU constraints in December 2025 and explain the coming CPU surge driven by Reinforcement Learning and inference. Amazon is the best positioned CSP, with its custom chip Graviton providing better economics than merchant solutions. In the AI Accelerator market, Amazon hopes to replicate the same success, and is seeing good results with Trainium. As explained in our Deep Dives, Trainium is attractive for inference and RL workloads.

  垂直整合：我们率先在 2025 年 12 月报道了日益增长的 CPU 限制，并解释了由强化学习和推理驱动的 CPU 需求激增。亚马逊是定位最佳的云服务商，其自研芯片 Graviton 相比商用解决方案提供了更优的经济效益。在 AI 加速器市场，亚马逊希望复制同样的成功，并且 Trainium 已初见成效。正如我们的深度研究报告所述，Trainium 对推理和强化学习工作负载极具吸引力。

Let's dig in, starting by covering the economics and market drivers of Bedrock, Amazon's most differentiated product. We then dive into their datacenter footprint relative to others, and then at the end of the report dive into the outlook for other CSPs.

让我们深入分析，首先从亚马逊最具差异化的产品 Bedrock 的经济效益和市场驱动因素入手。接着，我们将探讨其数据中心规模与其他厂商的对比，最后在报告结尾部分分析其他云服务商的前景。

Amazon Bedrock Deep Dive 亚马逊 Bedrock 深度研究

Bedrock is an AWS service that enables customers to choose their favorite LLM among many options, benefit from AWS security and compliance and unified billing (among other things) and run AI workloads. This market, which we call “API endpoints”, has many competitors, including Microsoft Foundry and Google Gemini Enterprise Agent Platform (previously Vertex), as well as many providers focused on open-source models such as TogetherAI, Fireworks, Baseten, etc.

Bedrock 是 AWS 的一项服务，允许客户从众多选项中选择自己偏好的 LLM，同时享受 AWS 的安全合规、统一计费（以及其他功能）并运行 AI 工作负载。这个我们称之为"API 端点"的市场拥有众多竞争者，包括 Microsoft Foundry 和 Google Gemini Enterprise Agent Platform（前身为 Vertex），以及许多专注于开源模型的提供商，如 TogetherAI、Fireworks、Baseten 等。

Endpoints typically claim to be differentiated through the following items:

端点通常通过以下方面宣称自身差异化：

- Model library breadth: providers like to flex the number of LLMs their platform has available.

  模型库广度：提供商喜欢展示其平台可用的 LLM 数量。

- Price: some providers have a differentiated inference stack, or cost structure, which enables them to offer more attractive prices while keeping margins viable.

  价格：部分提供商拥有差异化的推理栈或成本结构，使其能够在保持可行利润率的同时提供更具吸引力的价格。

- Interactivity: some vendors have better metrics, e.g. higher token throughput, lower TTFT, etc.

  交互性：部分供应商拥有更优的指标，例如更高的 Token 吞吐量、更低的 TTFT（首 Token 生成时间）等。

While some of these criteria do matter, they miss the single most important differentiator: access to frontier LLMs. As demonstrated by our Tokenomics Model,

Frontier LLMs make up the vast majority of AI API industry revenue.

尽管这些标准中的某些确实重要，但它们忽略了最关键的区别：对前沿 LLMs 的访问权限。正如我们的 Tokenomics 模型所证明的，前沿 LLMs 占据了 AI API 行业收入的绝大部分。

In the API endpoint market, this is the massive advantage that AWS, Microsoft, and Google boast over everyone else. For a long time, AWS had access to Claude, Google to both Claude and Gemini, and Microsoft to OpenAI models. Recently, AWS gained OpenAI access, while Microsoft gained Claude. No other CSP currently has the ability to sell OpenAI, Claude, and Gemini tokens.

在 API 端点市场中，这正是 AWS、微软和谷歌相较于其他所有厂商拥有的巨大优势。长期以来，AWS 拥有 Claude 的访问权限，谷歌同时拥有 Claude 和 Gemini，而微软则拥有 OpenAI 模型。最近，AWS 获得了 OpenAI 的访问权限，而微软也获得了 Claude 的权限。目前，没有其他 CSP（云服务提供商）能够同时销售 OpenAI、Claude 和 Gemini 的 Token。

Having access to these models is one thing, but building a substantial business around them is another. AI Inference notably has huge compute needs. To understand this, let's dig into the economics of Bedrock, Vertex, and Foundry.

拥有这些模型的访问权限是一回事，但围绕它们建立实质性的业务则是另一回事。AI 推理尤其具有巨大的计算需求。要理解这一点，让我们深入探讨 Bedrock、Vertex 和 Foundry 的经济学原理。

Token-as-a-Service Platform Economics

Token 即服务平台的商业模式

The economics of TaaS platforms are very different if they own the IP (or can freely use it, eg open source), versus if they "distribute" the IP:

TaaS 平台的商业模式存在显著差异，取决于平台是拥有知识产权（或可自由使用，如开源模型），还是仅作为知识产权的"分发渠道"：

- IP ownership: the economics are the same as that of an AI Lab. The cloud/token vendor has a fixed cost which is the infrastructure. That cost is largely driven by GPU depreciation, margin of the CSP, datacenter costs, and electricity costs.

Revenue is a function of tokens sold: to make good money, token pricing needs to be high enough, and hardware must be well utilized. Volumes and pricing must be large enough to absorb fixed costs and make some margin.

知识产权自有模式：其商业模式与 AI 实验室相同。云服务/代币供应商的固定成本来自基础设施，主要由 GPU 折旧、云服务提供商利润率、数据中心成本和电力成本构成。收入取决于代币销售量：要获得可观利润，代币定价需足够高，且硬件利用率必须充分。交易量和定价需达到足够规模，才能覆盖固定成本并产生利润空间。

- Model distribution: in this example, Amazon sells Claude tokens to an Amazon customer that pays an AWS bill. However, the seller of record is Anthropic. Public terms on Bedrock are clear: the product is sold by Anthropic, and use of the model is governed by Anthropic's terms. However, customers are invoiced by AWS, and the terms state that the model is "Deployed on AWS". In practice, this means that:

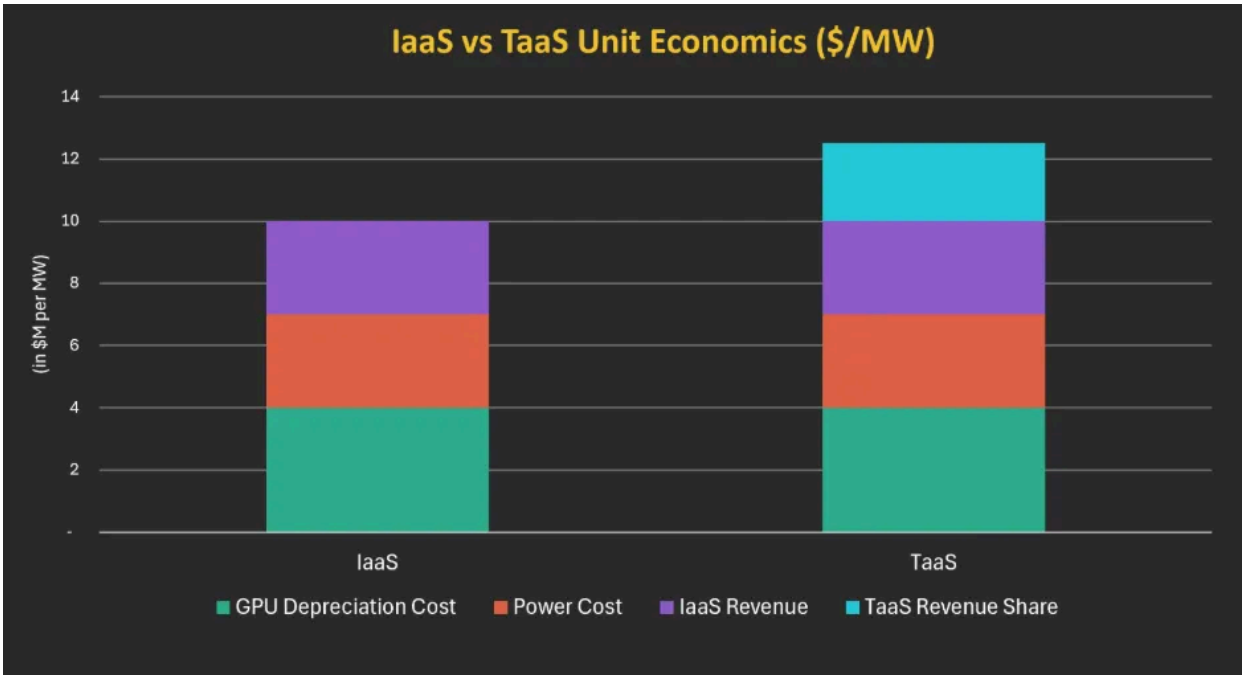
模型分发模式：在此场景下，亚马逊向支付 AWS 账单的客户销售 Claude 代币，但实际销售方为 Anthropic。Bedrock 平台的公开条款明确说明：产品由 Anthropic 销售，模型使用受 Anthropic 条款约束。然而，客户由 AWS 开具发票，条款注明模型"部署于 AWS 平台"。实际操作中这意味着：

- As Seller, Anthropic books full revenue of the sold tokens.

作为卖方，Anthropic 将已售代币的全部收入入账。

- As computer and marketplace provider, AWS gets both an infrastructure fee (akin to an EC2 IaaS fee) and a distribution or revenue share fee. The latter is what boosts margins and makes selling Claude on Bedrock a highly attractive business to AWS.

作为计算与市场平台提供商，AWS 既能获得基础设施费用（类似于 EC2 的 IaaS 费用），也能获得分销或收入分成费用。后者正是提升利润率的关键，使得在 Bedrock 上销售 Claude 对 AWS 而言成为一项极具吸引力的业务。



Source: SemiAnalysis Datacenter Model

来源：SemiAnalysis 数据中心模型

Anthropic was the first AI Lab to roll this out with AWS and Google, recently followed by OpenAI with AWS. Security is of maximum importance: ideally, the CSP doesn't have access to model weights, or a very limited one, but can still use the model and run it on its infrastructure.

Anthropic 是首个与 AWS 和 Google 推出该模式的 AI 实验室，随后 OpenAI 也与 AWS 跟进。安全性至关重要：理想情况下，云服务提供商无法访问模型权重，或仅能有限访问，但仍能使用模型并基于其基础设施运行。

For the AI Lab, there are two main benefits:

对于 AI 实验室而言，主要有两大优势：

- Access to the CSP's customer base

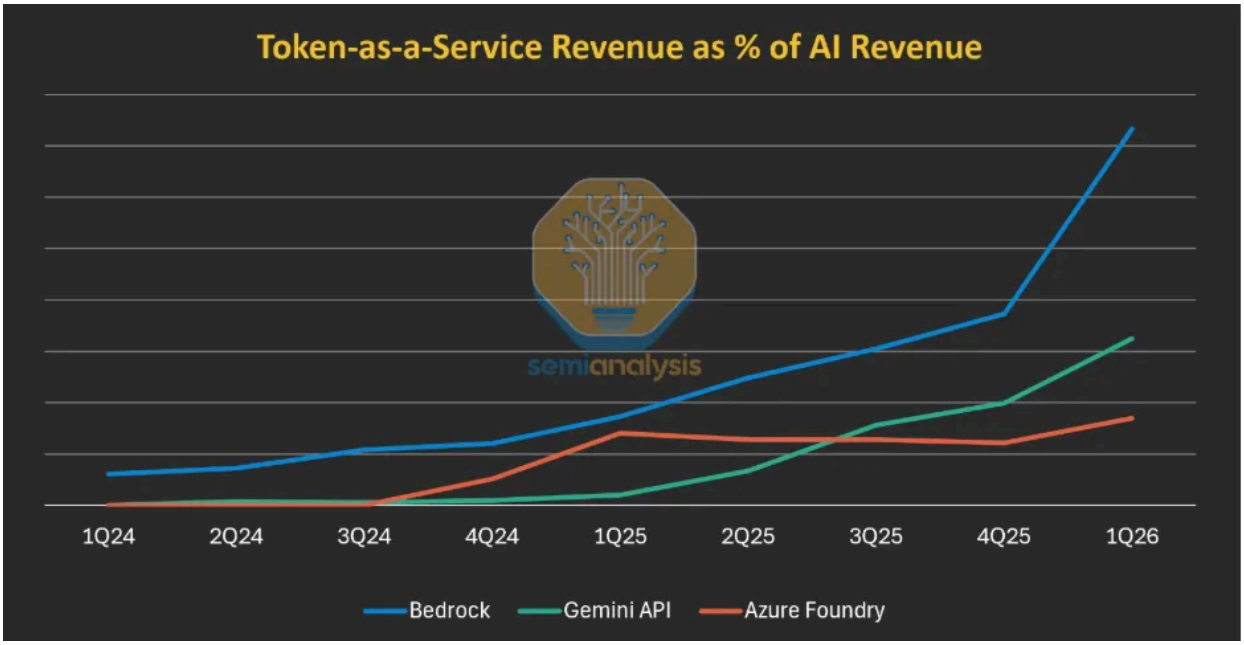
  接入云服务商的客户群体

- Access to compute capacity without the need of an expensive multi-year IaaS contract, but at a higher cost

  无需签订昂贵的多年期 IaaS 合同即可获得算力资源，但成本相对更高

For the CSP, this means more risk relative to a standard GPU IaaS deal, since revenue is not guaranteed via a 5yr IaaS take-or-pay deal. However, the margins are significantly more attractive. And Amazon with its deal structuring and execution has been the main beneficiary of its bet on Bedrock, dwarfing the size of their rival platforms.

对云服务商而言，这意味着相比标准 GPU IaaS 交易承担更多风险——因为无法通过五年期照付不议的 IaaS 合同锁定收入。但利润率却显著更具吸引力。而亚马逊凭借其交易架构设计与执行能力，已成为押注 Bedrock 平台的最大赢家，其平台规模远超竞争对手。



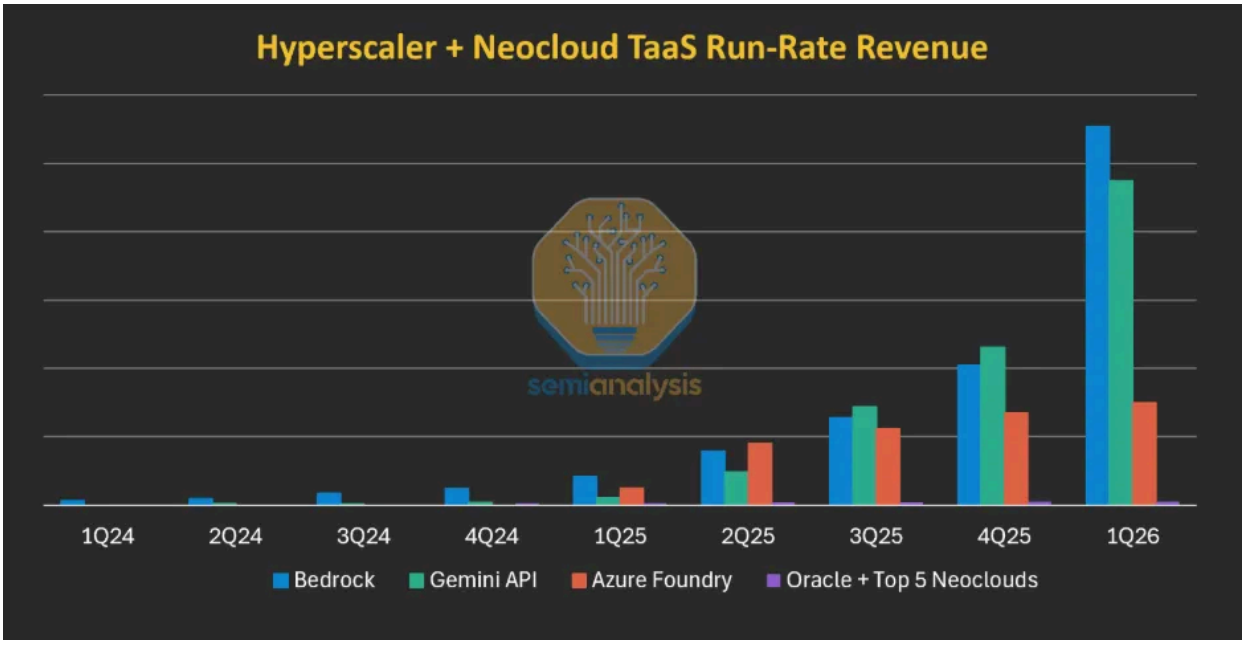
Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济学模型

And this is the main difference for the AMZN, MSFT, and GOOGL against ORCL and neoclouds. Their Token as a Service (TaaS) businesses are $10B+ in ARR today. Contrast this with neoclouds and Oracle at practically nothing. As we show in the Bedrock/Anthropic deal section, margins on these TaaS deals are significantly higher than being a wholesale seller of AI IaaS rental compute. The margin mix and distribution advantage these hyperscalers now possess is significantly widening the

gap amongst the top hyperscalers and the rest of the field.

这正是亚马逊（AMZN）、微软（MSFT）和谷歌（GOOGL）与甲骨文（ORCL）及新云服务商之间的主要区别。它们的“代币即服务”（TaaS）业务目前年化经常性收入（ARR）已超过 100 亿美元。相比之下，新云服务商和甲骨文的该项收入几乎为零。正如我们在 Bedrock/Anthropic 交易部分所展示的，这些 TaaS 交易的利润率远高于作为 AI 基础设施即服务（IaaS）算力的批发商。这些超大规模云服务商如今在利润率组合和分销渠道上的优势，正显著拉大它们与行业其他参与者之间的差距。



Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济学模型

Vertical Integration Positions AWS for Industry-Leading Margins

垂直整合使 AWS 在利润率方面处于行业领先地位

Having a greater business mix towards Bedrock is a great combination with Amazon's bet on custom silicon. In prior reports, we've covered Trainium2 and 3 in great depth, arguing that the chips have leading perf/TCO in certain scenarios where memory bandwidth is most valuable, such as high-batch inference and RL.

提高 Bedrock 在业务组合中的占比，与亚马逊押注定制芯片的战略形成了绝佳组合。在之前的报告中，我们已深入探讨过 Trainium2 和 3，认为这些芯片在内存带宽最具价值的特定场景（如高批量推理和强化学习）中，拥有领先的性价比（perf/TCO）。

In Bedrock, the underlying hardware is abstracted from customers; they only see tokens. This means that an inference-optimized chip is a natural fit with Bedrock, with the main caveat being that porting models can be longer than with Nvidia GPUs, but it's typically just a matter of days. This means that Trainium has a natural offtake in the form of Bedrock, and Amazon is taking full advantage of this, per the CEO:

chip business.

除了 AI 加速器，CPU 在前沿 LLM 的训练和推理中也日益重要。在 CPU 领域，垂直整合是更大的差异化优势，亚马逊第五代定制芯片 Graviton 处于行业领先地位。我们曾在此处对 CPU 进行过全面深度分析，并指出 Graviton4 和 5 为亚马逊带来了卓越的每性能总拥有成本优势。Graviton4 将作为头节点集成到 Trn3 中，并单独用于处理强化学习工作负载和智能体任务。亚马逊目前与 Anthropic、OpenAI 和 Meta 达成了大规模的 CPU 及 Graviton 交易。向 Bedrock 客户推销 Graviton 对亚马逊而言也将更加容易。在我们的代币经济学模型中，我们量化了这一非 AI 芯片业务对 AWS 收入增长的贡献。

CPUs are Back: The Datacenter CPU Landscape in 2026

CPU 回归：2026 年数据中心 CPU 格局

GERALD WONG AND DYLAN PATEL

GERALD WONG 和 DYLAN PATEL · 2月10日 2 月 10 日

Read full story 阅读完整报道 →

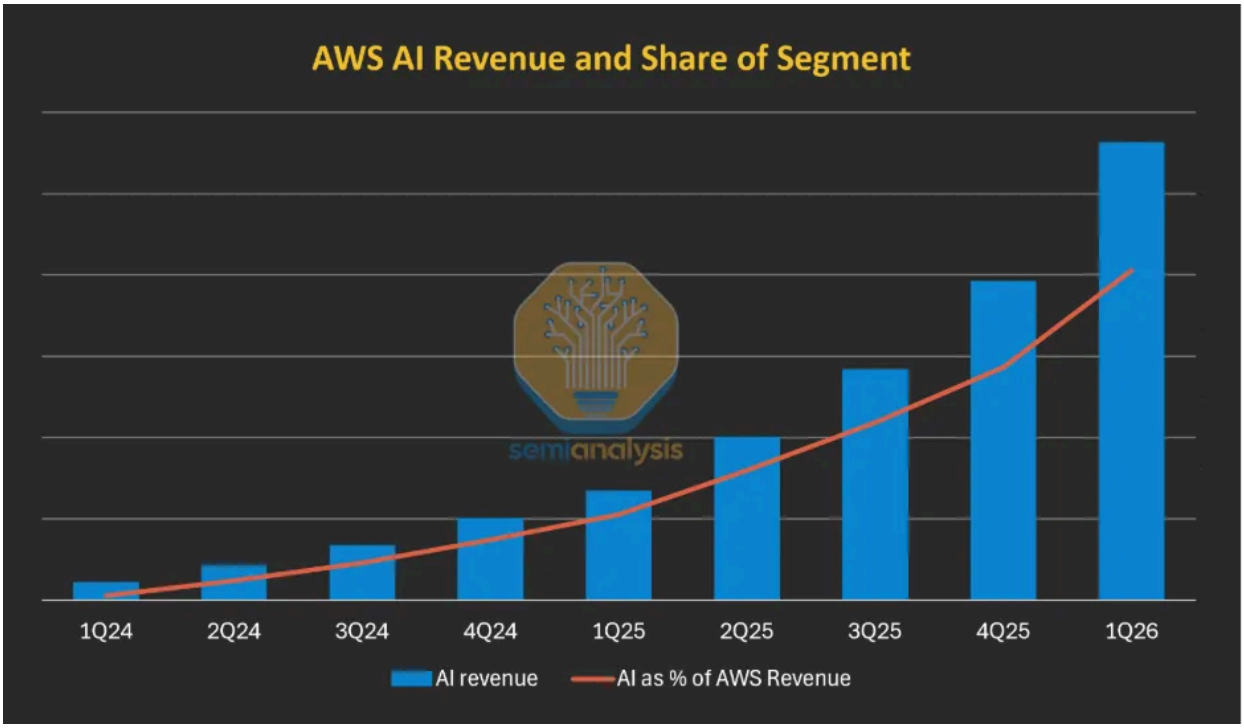
AWS Bedrock Mix: The Numbers

AWS Bedrock 组合：数据解析

Our Tokenomics Model goes into detail within each hyperscaler AI business to show the breakout of segments such as Bedrock at AWS, GitHub CoPilot within Azure, and Gemini API vs AI IaaS at Google Cloud. We have triangulated AWS' AI Business and the mix of Bedrock to better assist investors and enterprises in understanding these trends. For AWS, the mix of AI as a % of total revenue has increased from 2% in 1Q24 to 10% today in 1Q26. Our estimates going forward, available to Tokenomics subscribers, show this continuing to increase. In 1Q26, the mix of AI as a % of total GCP and Azure revenue was 36% and 27%, respectively. However, we believe the mix difference between IaaS andTaaS is a major contributor to margin differences despite

AWS' lower total AI mix.

我们的代币经济学模型深入分析了各大超大规模云服务商的 AI 业务，详细拆分了 AWS 的 Bedrock、Azure 的 GitHub Copilot 以及 Google Cloud 的 Gemini API 与 AI IaaS 等细分板块。我们通过三角验证了 AWS 的 AI 业务及其 Bedrock 组合，以更好地帮助投资者和企业理解这些趋势。对于 AWS 而言，AI 收入占总收入的比例已从 2024 年第一季度的 2%增长至 2026 年第一季度的 10%。我们面向代币经济学订阅用户提供的未来预测显示，这一比例将持续上升。在 2026 年第一季度，AI 收入占 GCP 和 Azure 总收入的比例分别为 36%和 27%。然而，我们认为，尽管 AWS 的 AI 总占比相对较低，但 IaaS 与 TaaS 之间的组合差异是导致利润率差异的主要因素。



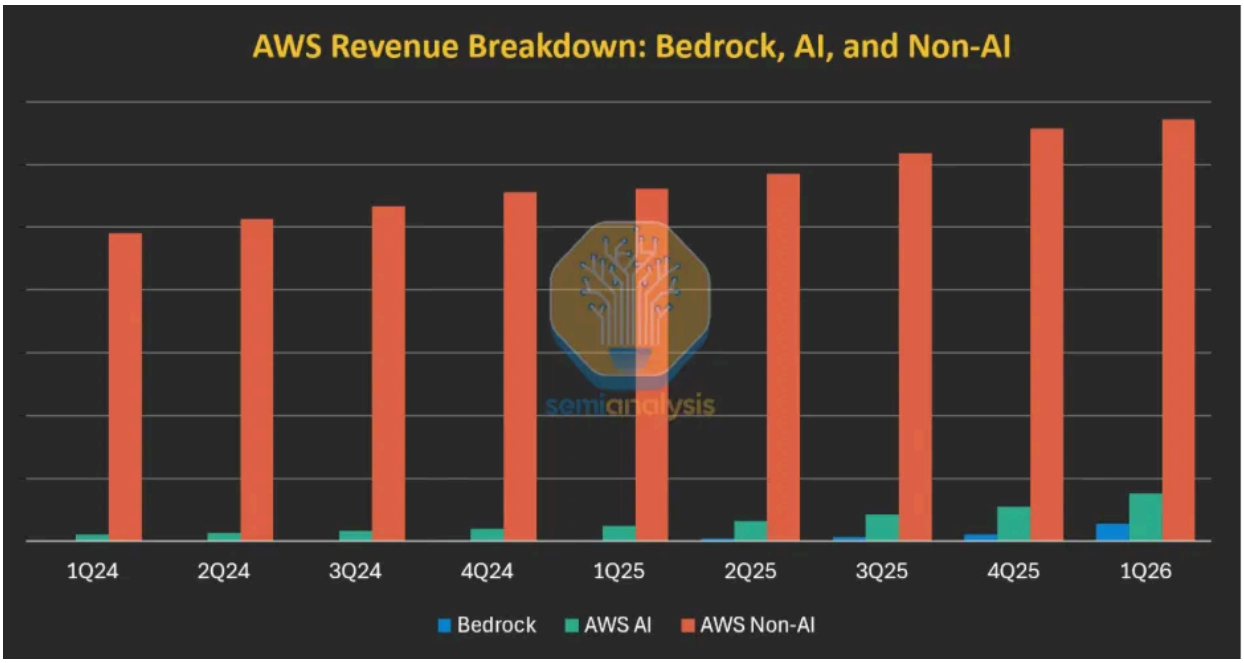
Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济学模型

As discussed above and as we dive deeper into the deal economics below, AWS' Bedrock mix vs Azure Foundry and GCP Gemini API is important. Bedrock has steadily grown as a % of total AWS AI revenue to 37% today, up from 9% in 1Q25 when IaaS dominated the AI business. In comparison, AI IaaS is still 80%+ of the AI business mix at Azure and GCP. Per Amazon's 1Q26 and 4Q25 earnings call, Bedrock revenue grew 170% and 60% Q/Q, respectively. SemiAnalysis believes Bedrock is a $5.5B run rate business today with the vast majority of customers (80-90%+) using

Anthropic models.

如上所述，随着我们深入分析交易经济性，AWS 的 Bedrock 与 Azure Foundry 及 GCP Gemini API 的混合模式至关重要。Bedrock 在 AWS AI 总收入中的占比已从 2025 年第一季度（当时 IaaS 主导 AI 业务）的 9%稳步增长至如今的 37%。相比之下，Azure 和 GCP 的 AI 业务中，AI IaaS 仍占 80%以上。根据亚马逊 2026 年第一季度和 2025 年第四季度的财报电话会议，Bedrock 收入分别环比增长 170%和 60%。SemiAnalysis 认为，Bedrock 目前年化收入已达 55 亿美元，绝大多数客户（80-90%以上）使用 Anthropic 模型。



Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济学模型

Clearly, Amazon has found a winner in terms of revenue growth and mix with Bedrock and the Anthropic deal. Microsoft, in contrast, is heavily IaaS driven, and other AI efforts like CoPilot in 365 and GitHub have failed to create traction. Google has done well with Gemini API within Gemini Enterprise Agent Platform, but has not benefited from the same coding market trends that have helped drive Anthropic's API revenue up ~13x y/y.

显然，亚马逊凭借 Bedrock 和 Anthropic 交易在收入增长和业务结构上取得了成功。相比之下，微软严重依赖 IaaS，而其在 365 Copilot 和 GitHub 等其他 AI 领域的努力未能取得进展。谷歌在 Gemini Enterprise Agent 平台上的 Gemini API 表现良好，但未能受益于推动 Anthropic API 收入同比增长~13 倍的相同编程市场趋势。

Bedrock/Anthropic Deal Bedrock/Anthropic 交易

Our understanding of Anthropic's Bedrock deal includes a flat IaaS fee, revenue share, and performance hurdles for outperformance above certain levels of token/spend throughput.

我们对 Anthropic 与 Bedrock 合作的理解包括：固定的 IaaS 费用、收入分成，以及在代币/吞吐量超过特定水平时的超额业绩门槛。

The Tokenomics team models Anthropic's and OpenAI's businesses extensively by segment (Consumer, Business/Team, Enterprise) and plan (Free, Pro, Max20, API, etc) at an ARR, revenue, margin, token volume, and token pricing level to help investors and corporate strategy departments understand this flow of dollars and tokens between customers, labs, and CSPs. This work includes modeling the Anthropic and Bedrock deal, which we have previewed below to illustrate the Gross Margins of Anthropic and EBIT Margins of Bedrock at differing levels of Anthropic ARR per MW on Bedrock Compute. We illustrate that margins for Bedrock, or any Token as a Service product, have superior margins than the typical IaaS fee business. The takeaway for hyperscalers is that these TaaS arrangements are excellent margin accretive additions when structured beneficially for both the hyperscaler and lab to drive revenue higher.

Tokenomics 团队按细分市场（消费者、商业/团队、企业）和套餐（免费版、Pro 版、Max20 版、API 等）对 Anthropic 和 OpenAI 的业务进行了广泛建模，涵盖 ARR、收入、利润率、代币量和代币定价等维度，以帮助投资者和企业战略部门理解客户、实验室和云服务商之间的资金与代币流动。这项工作包括对 Anthropic 与 Bedrock 合作的建模，我们已在下方预览，以展示在 Bedrock 计算资源上每兆瓦 Anthropic ARR 不同水平下，Anthropic 的毛利率和 Bedrock 的息税前利润率。我们表明，Bedrock 或任何"代币即服务"产品的利润率均优于典型的 IaaS 费用业务。对超大规模云服务商而言，关键在于：当这些 TaaS 安排对云服务商和实验室双方均有利时，它们能成为极佳的利润率增长点，推动收入提升。

AWS Bedrock Deal Illustrative Economics:

AWS/Anthroipic Revenue Share	%	25%
Bedrock IaaS Fee per MW per Year	$M	7.5
AMZN OpEx per MW per Year	$M	6.5

Illustrative Sensitivity Table of Amazon Bedrock EBIT Margins

Anthropic Gross Revenue per MW	10	15	20	25	30	35	40
AMZN Net Revenue per MW	10	11	13	14	15	16	18
AMZN EBIT per MW	4	5	6	7	9	10	11
Amazon EBIT Margins	35%	42%	48%	53%	57%	60%	63%

Illustrative Sensitivity Table of Anthropic Gross Margins

Anthropic Gross Revenue per MW	10	15	20	25	30	35	40
Anthropic Gross Margin $ per MW	0	4	8	11	15	19	23
Anthropic Gross Margins	0%	25%	38%	45%	50%	54%	56%

Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis Tokenomics 模型

We believe that Anthropic's recent revenue outperformance (discussed below) puts Anthropic Bedrock revenue around $26M per MW in 1Q26, implying Bedrock EBIT margins of ~55% in the above analysis. For 1Q26 vs 1Q25 at AWS, we believe that reported incremental EBIT margins were primarily driven by Bedrock, with Bedrock accounting for 30% of the step-up in Gross Profit Dollars Y/Y despite accounting for only 4% of total AWS revenue. For 2Q we see Anthropic ARR per MW of Total Compute around $42M and Bedrock Revenue mix increasing to 53% of AWS AI Revenue contributing an incremental 9 points to total AWS revenue growth. We note that given Anthropic's ARR recognition accounting on a gross basis, that margins on Bedrock are slightly negative mix vs their current blended low 60% inference gross margins.

我们认为，Anthropic 近期收入表现超预期（下文详述）使其在 2026 年第一季度的 Bedrock 收入达到每兆瓦约 2600 万美元，这意味着在上述分析中 Bedrock 的 EBIT 利润率约为~55%。对比 AWS 2026 年第一季度与 2025 年第一季度，我们认为报告中的增量 EBIT 利润率主要受 Bedrock 驱动——尽管 Bedrock 仅占 AWS 总收入的 4%，却贡献了毛利润同比增量的 30%。展望第二季度，我们预计 Anthropic 每兆瓦总计算能力的 ARR 约为 4200 万美元，Bedrock 收入占 AWS AI 收入的比重将升至 53%，为 AWS 总收入增长额外贡献 9 个百分点。需注意，由于 Anthropic 按总额确认 ARR 的会计处理方式，Bedrock 的利润率相对于其当前约 60%的推理毛利率（低端水平）呈现轻微负向混合效应。

Anthropic's Blowout Q1 Anthropic 第一季度业绩爆发

Anthropic's 1Q26 was characterized by rapid growth which helped propel Bedrock revenue and margins higher. The company added $21B in net new ARR in the quarter to reach $30B of ARR. Most of Anthropic's revenue is API and Enterprise based whereas OpenAI skews consumer and has a relatively heavy inference cost load from free users. This revenue explosion is driven by Claude Code, which has taken enterprises by storm. Consumers, too, have begun to flock to Claude. In a recent note to subscribers, we noted Anthropic's increasing (and now majority) market share of net new customers vs OpenAI within card panels as well as higher average transaction values. We continue to see strong demand for Anthropic - primarily driven by enterprise API spending - through the end of the year. Our latest Tokenomics model

estimates see potential for well over $100B in ARR by year.

Anthropic 2026 年第一季度的快速增长推动了 Bedrock 收入和利润率的提升。该公司当季新增 210 亿美元净新增年度经常性收入（ARR），总 ARR 达到 300 亿美元。Anthropic 的收入主要来自 API 和企业端，而 OpenAI 则偏向消费者领域，且因免费用户承担着相对较高的推理成本负担。这一收入激增主要得益于 Claude Code，该产品已在企业市场掀起热潮。消费者也开始纷纷涌向 Claude。在最近给订阅用户的一份报告中，我们注意到在信用卡面板数据中，Anthropic 在净新增客户中的市场份额（现已占多数）以及平均交易价值均持续超越 OpenAI。我们预计到年底前，主要由企业 API 支出驱动的 Anthropic 需求仍将保持强劲。根据我们最新的代币经济学模型估算，其年度经常性收入有望在年底前突破 1000 亿美元大关。

Anthropic Model	No Error	0											
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
ARR Build:														
ARR (Annualized Run Rate), end of period	$M		264	509	755	1,000	2,000	3,500	6,000	9,000	30,000	57,999	88,454	121,340
y/y %	%					1,049%	859%	588%	695%	800%	1,400%	1,557%	1,374%	1,248%
q/q %	%		203%	93%	48%	33%	100%	75%	71%	50%	233%	93%	53%	37%
y/y $	$M					913	1,736	2,991	5,245	8,000	28,000	54,499	82,454	112,340
q/q $	$M		177	245	245	245	1,000	1,500	2,500	3,000	21,000	27,999	30,455	32,886
Net New ARR	$M		177	245	245	245	1,000	1,500	2,500	3,000	21,000	27,999	30,455	32,886
y/y %	%						466%	511%	919%	1,122%	2,000%	1,767%	1,118%	996%
q/q %	%			39%	-	0%	307%	50%	67%	20%	600%	33%	9%	8%
Total Revenue	$M		44	97	158	219	375	688	1,188	1,875	4,875	11,000	18,307	26,224
y/y %	%						756%	612%	652%	755%	1,200%	1,500%	1,442%	1,299%
q/q %	%			120%	64%	39%	71%	83%	73%	58%	160%	126%	68%	43%
Monthly Build:														
ARR, end of period	$M		264	509	755	1,000	2,000	3,500	6,000	9,000	30,000	57,999	88,454	121,340
Month 1	$M		100	345	591	836	1,000	2,500	4,000	7,000	12,000	44,000		
Month 2	$M		182	427	673	918	1,400	3,000	5,000	8,000	19,000			
Month 3	$M		264	509	755	1,000	2,000	3,500	6,000	9,000	30,000			
Net New ARR $ P/P	$M		177	245	245	245	1,000	1,500	2,500	3,000	21,000			
Month 1	$M		13	82	82	82	-	500	500	1,000	3,000	14,000		
Month 2	$M		82	82	82	82	400	500	1,000	1,000	7,000			
Month 3	$M		82	82	82	82	600	500	1,000	1,000	11,000			
ARR Growth P/P	%		203%	93%	48%	33%	100%	75%	71%	50%	233%	93%		
Month 1	%		15%	31%	16%	11%	-	25%	14%	17%	33%	47%		
Month 2	%		82%	24%	14%	10%	40%	20%	25%	14%	58%			
Month 3	%		45%	18%	12%	9%	43%	17%	20%	13%	58%			
ARR Growth Y/Y	%						559%	588%	695%	800%	1,400%			
Month 1	%						900%	624%	577%	737%	1,100%			
Month 2	%						570%	602%	643%	771%	1,257%			
Month 3	%						859%	588%	695%	800%	1,400%			
Segment ARR Build:														
Overall ARR by Segment:														
Consumer Subscription ARR	$M		30	35	65	95	170	200	325	400	1,550	2,388	3,067	3,718
Team Subscription ARR	$M		-	21	26	20	40	70	120	180	600	883	1,417	1,872
Enterprise Subscription ARR	$M		-	-	-	10	50	185	395	680	2,400	4,200	6,000	7,900
API Consumption ARR	$M		235	453	664	880	1,740	3,045	5,160	7,740	25,450	50,450	77,950	107,950
Overall ARR Mix Details:														
Consumer Subscription ARR	%		11%	7%	9%	10%	9%	6%	5%	4%	5%	4%	3%	3%
Team Subscription ARR	%		-	4%	3%	2%	2%	2%	2%	2%	2%	2%	2%	2%
Enterprise Subscription ARR	%		-	-	-	1%	3%	5%	7%	8%	8%	7%	7%	6%
API Consumption ARR	%		89%	89%	86%	86%	87%	87%	86%	86%	85%	87%	88%	89%
Overall ARR Mix Growth Y/Y:	%		-	-	-	-	-	-	-	-	-	-	-	-
Consumer Subscription ARR Growth Y/Y	%						467%	471%	400%	321%	812%	1,083%	850%	830%
Team Subscription ARR Growth Y/Y	%							233%	370%	800%	1,400%	1,304%	1,081%	940%
Enterprise Subscription ARR Growth Y/Y	%									5,700%	4,700%	2,170%	1,419%	1,047%
API Consumption ARR Growth Y/Y	%						642%	572%	677%	780%	1,383%	1,557%	1,411%	1,295%

Source: SemiAnalyis Tokenomics Model

来源：SemiAnalysis 代币经济学模型

In addition, Anthropic's margins on inference compute have shot up. We believe inference gross margins are now in the mid 60s, up from 38% in 2025 and -94% in 2024. Per a Wall Street Journal article on 5/20/26, Anthropic is expected to be Operating Income profitable in 2Q after adjusting for stock-based compensation. Our Tokenomics model for Anthropic allocates costs to Free and Paid User Inference,

Training & Research Spend, and OpEx. Our latest estimates published to clients last week were extremely aligned across revenue, cost bucket mix and dollar amounts as well as overall GAAP (unadjusted for stock-based compensation) EBIT profitability.

此外，Anthropic 在推理计算方面的利润率大幅提升。我们认为其推理毛利率已从 2025 年的 38%和 2024 年的-94%上升至 60%左右的中段。根据《华尔街日报》2026 年 5 月 20 日的报道，Anthropic 预计在调整股权激励后，将于第二季度实现营业利润盈利。我们的 Anthropic 代币经济学模型将成本分配至免费与付费用户推理、训练与研究支出以及运营支出。上周向客户发布的最新估算在收入、成本结构组合与金额，以及整体 GAAP（未调整股权激励）息税前利润方面均高度吻合。

Hyperscaler Takeaways: Amazon Met Demand With Capacity

超大规模云服务商要点：亚马逊以产能满足需求

Key to deploying a token-as-a-service platform at much larger scale than rivals was simply having more compute capacity, given the high compute-intensity of AI inference. And as shown below, Amazon is adding significantly more capacity than rivals, with only Microsoft a close tie in 2024-26 but largely dwarfed in 2027. Our Datacenter Industry Model provides quarter-by-quarter capacity forecast by hyperscaler.

在远超竞争对手的规模上部署代币即服务平台，关键在于拥有更多计算能力——尤其是考虑到 AI 推理的高计算密集度。如下图所示，亚马逊正在大幅增加产能，远超竞争对手：2024-2026 年间仅微软能与之接近，但到 2027 年亚马逊将完全占据压倒性优势。我们的数据中心行业模型按季度提供了各超大规模云服务商的产能预测。

Source: SemiAnalysis Datacenter Model

来源：SemiAnalysis 数据中心模型

However, total capacity is not the only relevant item. We need to break it down by end-user, which is what our Datacenter Industry Model can provide quarter by quarter. Our model quantifies the capacity going to Microsoft's internal AI efforts, which is higher than that of Amazon, reducing compute available to the broader customer base. In addition, the vast majority of Microsoft's AI compute goes to OpenAI via long-term compute contracts, as demonstrated by a remarkably large share of Microsoft's backlog, with OpenAI's backlog alone being 2.5x that of the total Azure annual

revenue.

然而，总容量并非唯一相关因素。我们需要按最终用户进行细分，这正是我们的数据中心行业模型每季度所能提供的。该模型量化了微软内部 AI 项目所占用的容量，其规模高于亚马逊，从而减少了可供更广泛客户群使用的算力。此外，微软绝大部分 AI 算力通过长期算力合同流向 OpenAI，这在其积压订单中占比极高——仅 OpenAI 一家的积压订单就达到 Azure 全年营收的 2.5 倍。

To deploy more capacity than peers, Amazon has been remarkably aggressive in growing its power pipeline, and signing multibillion PPAs with IPPs like Talen, Vistra, and NiSource.

为了部署比竞争对手更多的容量，亚马逊在扩大电力管线方面采取了极为激进的策略，并与 Talen、Vistra、NiSource 等独立电力生产商签署了数十亿美元的购电协议。

On the other hand, Microsoft had a year-long datacenter pause (see our Datacenter Freeze article), which significantly lowered their 2027 capacity forecast. In addition, as we've covered in our Microsoft's AI Strategy piece, Microsoft has been remarkably slow at building large-scale AI clusters in Wisconsin, the opposite of AWS' lighting speed to build close to 2GW in Indiana and Mississippi. The only way for Microsoft to catch up is to contract significant amounts of capacity from Neoclouds, which is much more expensive and will reduce the margin advantage. This is happening all while Amazon keeps innovating to accelerate capacity growth. The company is rolling out a new datacenter design at a very large scale, with increased modularity and prefabrication as they continue to prosecute against the AI opportunity.

另一方面，微软经历了长达一年的数据中心建设暂停（详见我们的《数据中心冻结》一文），这显著降低了其 2027 年的容量预期。此外，正如我们在《微软 AI 战略》分析中所指出的，微软在威斯康星州建设大规模 AI 集群的速度异常缓慢，与亚马逊在印第安纳州和密西西比州以闪电速度建设近 2 吉瓦容量的做法形成鲜明对比。微软追赶的唯一途径是向 Neocloud 大量租用容量，但这成本高昂且会削弱其利润率优势。而与此同时，亚马逊持续创新以加速容量增长——该公司正大规模推出新型数据中心设计，通过提升模块化和预制化程度，全力抓住 AI 机遇。

Amazon's AI Resurgence: AWS & Anthropic's Multi-Gigawatt Trainium Expansion

亚马逊的 AI 复兴：AWS 与 Anthropic 的多吉瓦 Trainium 扩张计划

JEREMIE ELIAHOU ONTIVEROS, DYLAN PATEL, AND 2 OTHERS

杰雷米·埃利亚胡·昂蒂维罗斯、迪伦·帕特尔等 2 人

• 2025年9月4日 2025 年 9 月 4 日

Read full story **阅读完整报道** →

The Google Pushback 谷歌的反击

The most natural pushback to our AWS outperformance thesis is that Google Cloud is also seeing increased margins, boasts the same vertical integration as AWS (if not more), and has been an even bigger outperformer, with revenue growth skyrocketing to >60% YoY in the latest quarter and margins a record high for GCP. We've been well ahead of the market in calling out Google Cloud's acceleration, with our first reports coming out in August 2025 in Accelerator, Datacenter and Core Research. Our TPU

Deep Dive explained in great detail why Google Cloud is outperforming.

对我们关于 AWS 表现优于同行的观点，最自然的反驳是：谷歌云同样在实现利润率提升，拥有与 AWS（甚至更胜一筹）的垂直整合能力，且表现更为亮眼——最新季度营收同比增速飙升至 60%以上，利润率创下 GCP 历史新高。我们早在市场之前就预判了谷歌云的加速增长，首份相关报告于 2025 年 8 月发布在《加速器》《数据中心》和《核心研究》中。我们的《TPU 深度解析》已详细阐释了谷歌云表现卓越的原因。

However, we think the margin rise is an illusion and is more akin to an "EBTIT" margin, i.e. Earnings Before Training, Interest and Taxes. Alphabet is seeing increased costs due to DeepMind/Gemini training expenses, but these costs are accounted for in "Alphabet-Level Activities," not GCP. Per Google's latest 10-Q, Alphabet-Level Activities, "primarily reflect expenses related to our shared AI research and development" and were $5.4B in 1Q26 up from $3.0B in 1Q25. Thus to say, all the Gemini API revenue is flowing to Google Cloud at higher-than-average margins, which boosts the margin profile of GCP while $10B+ of run rate costs are bucketed elsewhere.

然而，我们认为利润率上升是一种假象，更类似于"EBTIT"利润率，即扣除训练成本、利息和税金前的利润。Alphabet 因 DeepMind/Gemini 的训练支出而面临成本上升，但这些成本被计入"Alphabet 层面活动"，而非 GCP。根据谷歌最新 10-Q 文件，Alphabet 层面活动"主要反映与共享 AI 研发相关的支出"，在 2026 年第一季度达到 54 亿美元，高于 2025 年第一季度的 30 亿美元。因此可以说，所有 Gemini API 收入都以高于平均水平的利润率流入谷歌云，从而提升了 GCP 的利润率表现，而超过 100 亿美元的经常性成本却被归入其他类别。

In addition, Google Cloud might have benefited from one-off royalty payments due to the sale of TPUs to Anthropic, through Broadcom, with Google acting as an IP vendor.

此外，谷歌云可能还因通过博通向 Anthropic 销售 TPU 而获得一次性专利费收入，谷歌在此过程中扮演了 IP 供应商的角色。

Cloud vs AI Lab vs Hardware: How Can Google Satisfy All Demand?

云服务 vs AI 实验室 vs 硬件：谷歌如何满足所有需求?

Google is the ultimate supply-constrained business. This is a single company which attempts to compete simultaneously with AWS on Cloud, Nvidia on hardware, Anthropic & OpenAI on models, Meta on ads, Tesla on autonomous driving, and more.

谷歌是一家终极供应受限的企业。这家公司同时试图在云服务领域与 AWS 竞争，在硬件领域与英伟达竞争，在模型领域与 Anthropic 和 OpenAI 竞争，在广告领域与 Meta 竞争，在自动驾驶领域与特斯拉竞争，等等。

When analyzing Google's capacity growth, we simply do not see a large enough buildout to serve all demand. Internal capacity is enough to build a successful AI Lab, but leaves little room for the Cloud business (excluding hardware) to grow to the same extent as AWS. In particular, Gemini Enterprise Agent Platform (previously Vertex) as a distribution platform for Claude (and maybe OpenAI in the future) **is seeing a significantly lower capacity additions than Bedrock**.

在分析谷歌的产能增长时，我们并未看到足够大规模的基础设施建设来满足所有需求。其内部产能足以支撑一个成功的 AI 实验室，但留给云业务（不含硬件）的发展空间却十分有限，难以达到与 AWS 同等的增长规模。尤其值得关注的是，作为 Claude（未来可能还包括 OpenAI）分发平台的 Gemini 企业智能体平台（原 Vertex AI），其产能增量远低于 Bedrock。

For Google, GCP appears to be a platform to upsell additional services. Meta is the best example of this: a large GPU deal that subsequently led to large-scale Gemini adoption, and then a massive TPU hardware deal.

对谷歌而言，GCP 更像是用于交叉销售附加服务的平台。Meta 就是最佳例证：先是大额 GPU 订单，随后大规模采用 Gemini，最终达成巨额 TPU 硬件交易。

Implications for Hyperscalers and Labs
对超大规模云服务商与 AI 实验室的启示

We believe that margins at leading IaaS CSPs will only get better over the next 2-3 years. But, AWS' ability to show rising margins in a period of huge capacity ramp and CapEx inflation is remarkable and driven by strategic decisions their team made. Longer term, we expect AI Labs to increasingly verticalize their inference stack and pressure margins on pure IaaS vendors. We also expect CSPs to vertically integrate through custom silicon and strategic partnerships where they can add value through

their large, diversified customer bases.

我们认为，领先的 IaaS 云服务提供商在未来 2-3 年内的利润率只会持续改善。但 AWS 在产能大幅扩张和资本支出通胀的背景下，仍能实现利润率上升，这令人瞩目，且得益于其团队做出的战略决策。长期来看，我们预计 AI 实验室将日益垂直化其推理栈，从而对纯 IaaS 供应商的利润率形成压力。同时，我们预期云服务提供商将通过定制芯片和战略合作伙伴关系实现垂直整合，利用其庞大且多元化的客户基础创造价值。

To have access to the full data, forward estimates, and talk with our Tokenomics team, reach out to sales@semianalysis.com.

如需获取完整数据、前瞻性预测及与我们的 Tokenomics 团队交流，请联系 sales@semianalysis.com。

Previous 上一篇

A guest post by 一位特邀撰稿人

Joey Brookhart 乔伊•布鲁克哈特