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

Was Widespread TokenMaxxing Ever Really Here?

Token 最大化曾真的普遍存在吗?

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）

JUL 01, 2026 2026 年 7 月 1 日 · PAID · 付费内容

It’s been reported that token consumption inside of enterprises is hitting a budgeting wall after unhinged consumption earlier this year. The SemiAnalysis team talked with over 50 customers by slack, phone, and at the Databricks AI Summit to understand trends within the enterprise.

据报道，在年初无节制消耗之后，企业内部的 Token 消耗正触及预算瓶颈。SemiAnalysis 团队通过 Slack、电话以及在 Databricks AI 峰会上，与超过 50 位客户进行了交流，以了解企业内的趋势。

- Widely reported responses to Tokenmaxxing budgets from companies like Meta and Uber are overstated and stem from poor incentives and employee allocation we didn’t find present at other organizations

  像 Meta 和 Uber 等公司对 Tokenmaxxing 预算的广泛报道反应被夸大了，这源于不良的激励措施和员工分配问题，而我们并未在其他组织中发现这种情况。

- Budgets are now the new norm, but there’s no consensus number with budgets starting at $250 and going up to tens of thousands a month.

  预算已成为新常态，但并无统一标准——企业预算从每月 250 美元到数万美元不等。

- Companies are downgrading default models and turning off premium tiers while employees' game subscription M365 Copilot usage to stretch their token allowance.

  企业正在降级默认模型并关闭高级服务层级，而员工则像玩《使命召唤》那样精打细算地使用微软 365 Copilot，以延长自己的代币配额。

Rise and Fall of Tokenmaxxing

Tokenmaxxing 的兴衰

Tokenmaxxing started earlier this year when companies like Meta and Salesforce began encouraging their employees to consume as many AI tokens as possible to boost productivity. At Meta, an employee even built a "Claudeconomics" dashboard that ranked the top 250 power users in the company. The results showed that Meta employees consumed over 60T tokens over a 30-day period, with the single highest individual accounting for roughly 280B tokens. Employees started competing for rankings like "Token Legend" and "Cache Wizard" by having agents do research for hours simply to burn tokens.

代币极限利用潮始于今年早些时候，当时 Meta 和 Salesforce 等公司开始鼓励员工尽可能多地消耗 AI 代币以提升生产力。在 Meta，一名员工甚至搭建了名为"克劳德经济学"的仪表盘，对公司排名前 250 名的重度用户进行排行。数据显示，Meta 员工在 30 天内消耗了超过 60 万亿个代币，单个最高消耗用户约达 2800 亿个代币。员工们开始通过让 AI 代理花数小时做研究来"烧代币"，竞逐"代币传奇"和"缓存巫师"之类的排名。

The dashboard was shut down 2 days later after The Information reported the spend.

该仪表盘在《信息报》报道其支出后的两天内即被关闭。

That episode was just one amongst others in the enterprise tokenmaxxing trend in 1H26. Companies are now shifting focus from tokenmaxxing to token budgeting. Most recently, Uber made headlines for burning through their Claude Code and Codex annual budget in four months. In response, the company imposed a $1,500/month/employee limit, with over-limit requests allowed and approved on a case-

by-case basis. To see if the news reports on early 2026 tokenmaxxing and now tokenbudgeting were true, the SemiAnalysis team conducted on-the-ground conversations at the Databricks AI Summit and talked with large enterprises to understand the trends.

那个插曲只是 2026 年上半年企业 token 最大化浪潮中的一个案例。如今，企业正将重心从 token 最大化转向 token 预算管理。最近，优步因在四个月内烧光了 Claude Code 和 Codex 的年度预算而登上新闻头条。作为应对措施，该公司实施了每人每月 1500 美元的限制额度，超出部分需逐案审批。为验证 2026 年初关于 token 最大化及当前 token 预算管理的报道是否属实，SemiAnalysis 团队在 Databricks 人工智能峰会上进行了实地调研，与多家大型企业深入交流，以了解这一趋势的真相。

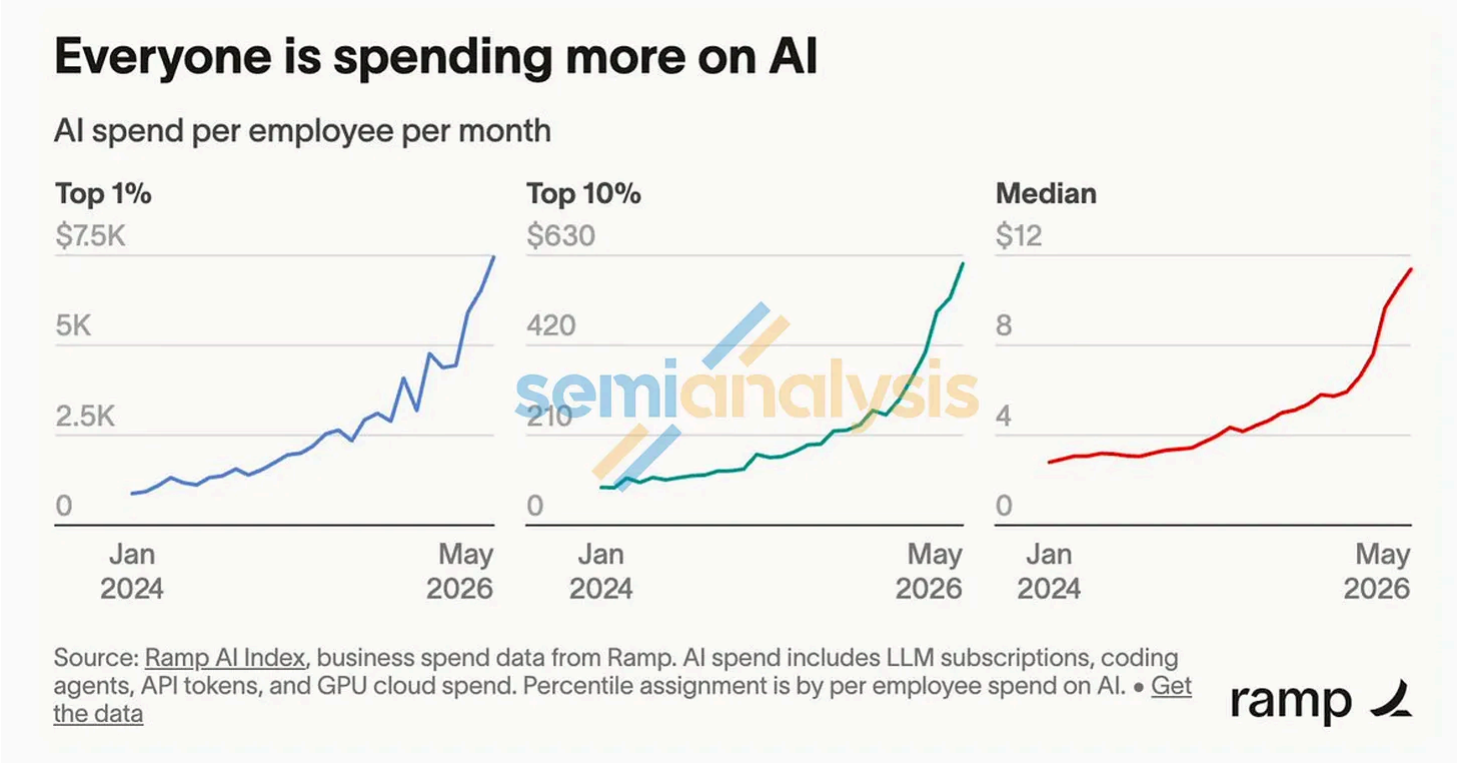
Our View of the Data & Narrative

我们对数据与叙事的看法

There are many news stories out there on tokenmaxxing and resulting budget blowouts. However, in our work in the Tokenomics Model, we estimate that 90th+ percentile customers make up most of the revenue and are at very little risk to API revenue cuts through the rest of the year. Even Meta, who was burning through 70T tokens per month in February and is spending close to at least $50,000/year per employee (at list price) is only a 3-5% customer for Anthropic per our estimates. Ramp data shows a similar trend amongst the top customers. 99th percentile customers spend almost $90,000/yr per employee while 90th percentile customers spend ~$7,300. This is a stark contrast to the median Ramp customer spending just $136. Note Ramp customers are generally way more tech forward so it's already a high spend skewed

distribution. The media fortune 500 is well below $100 per employee still.

关于令牌使用量激增及其预算超支的新闻屡见不鲜。然而，根据我们在代币经济学模型中的研究，我们估算前 90%的高端客户贡献了大部分收入，且这类客户在本年度剩余时间内基本不会对 API 收入造成削减风险。即便是今年二月消耗了 70 万亿令牌的 Meta——该公司按标价计算每年每位员工花费接近 5 万美元——据我们估算也仅占 Anthropic 客户的 3%-5%。Ramp 的数据显示顶级客户群体呈现相似趋势：前 1%的头部客户每位员工年均支出约 9 万美元，前 10%的客户则为 7300 美元。这与 Ramp 客户中位数（每位员工仅 136 美元）形成鲜明对比。值得注意的是，Ramp 客户整体技术应用水平远高于普通企业，因此其数据本身已呈现高消费的偏态分布。而《财富》500 强媒体类客户的人均支出仍远低于 100 美元。



Source: Ramp Economics Lab

数据来源：Ramp 经济学实验室

Our conversations with enterprise customers, including many Fortune 500s, followed this split. Many tech forward Fortune 500s spend well under $2,000/year per employee in AI, with the larger spend mostly in the engineering and data science departments. This suggests that the s-curve of growth in enterprise usage still has plenty of runway. Today's market is driven by the coding vertical explosion and other VC-backed AI companies whose products build on top of Anthropic or OpenAI models (90th-99th

percentile customers who are seeing their revenue accelerate too).

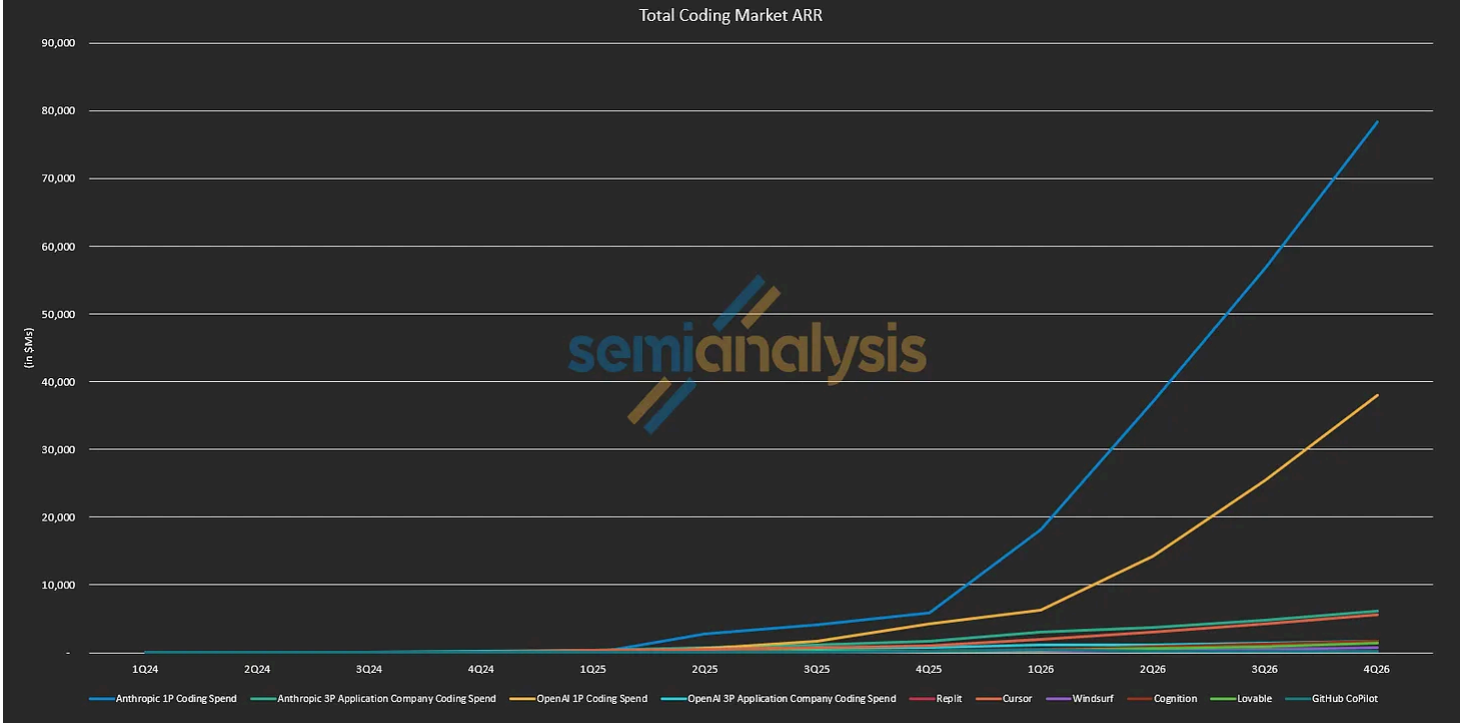
我们与包括众多《财富》500 强企业在内的企业客户进行了交流，话题围绕上述分野展开。许多技术前沿的《财富》500 强企业在 AI 领域的每位员工年均支出远低于 2000 美元，其中较大支出主要集中在工程和数据科学部门。这表明企业使用 AI 的 S 形增长曲线仍有充足的发展空间。当前市场由编程垂直领域的爆发以及依赖 Anthropic 或 OpenAI 模型构建产品的风投支持的 AI 公司（处于 90 th -99 th 百分位的客户，其收入也在加速增长）所驱动。

What the coding market has done to AI Lab ARR will be repeated with Cyber (Mythos re-release dependent) at an even faster pace than Claude Code and again with white-collar knowledge work as Cowork, CoPilot, Codex, and Computer type products penetrate the enterprise.

编程市场为 AI 实验室带来的 ARR 增长，将随着 Cyber（取决于神话重制版）以比 Claude Code 更快的速度重演，并随着 Cowork、CoPilot、Codex 和 Computer 类产品渗透企业，在白领知识工作领域再次上演。

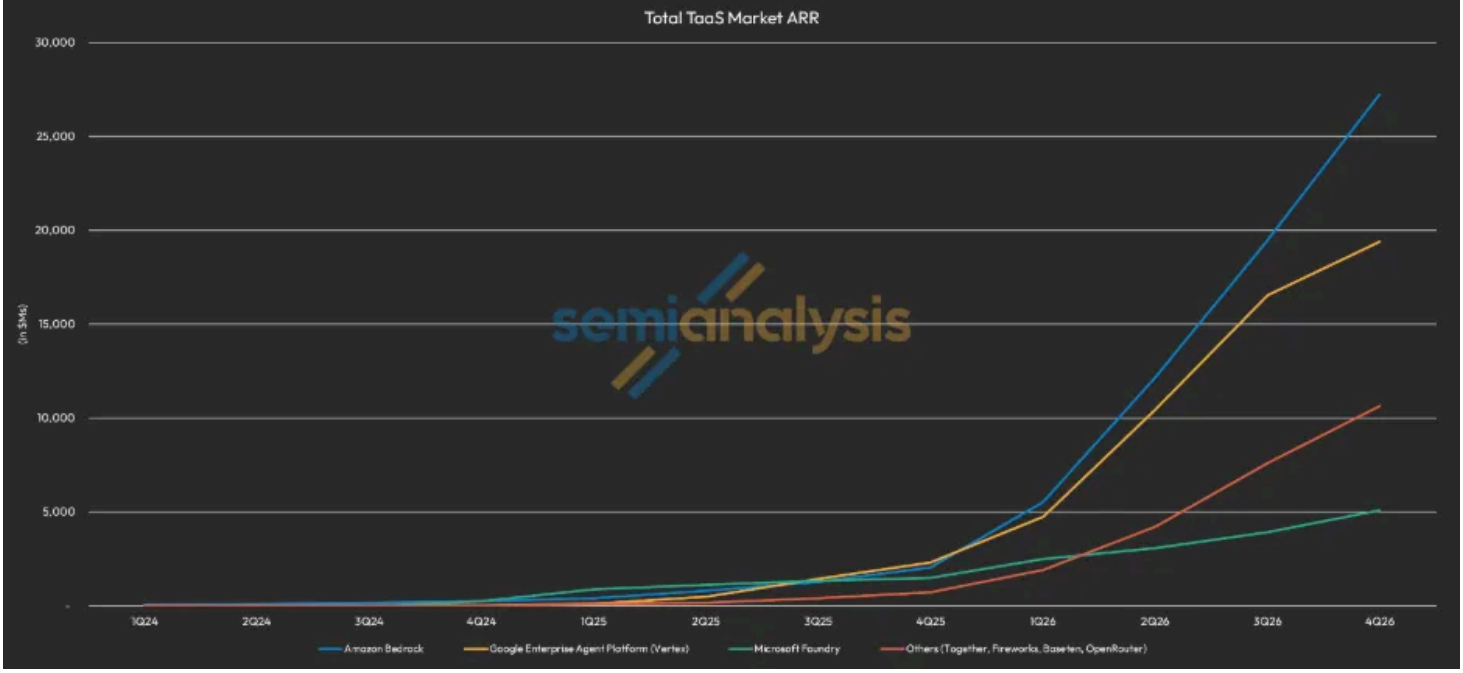
The Tokenomics Model estimates coding related spend at the AI Labs on a 1st and 3rd party basis and ARR and margins at the application layers (companies like Cursor, Loveable, GitHub CoPilot, and more) to help investors and corporates track growth of and within the vertical. We believe that over 70% of ARR today across OpenAI and Anthropic can be attributed to coding use cases with Anthropic spend higher than OpenAI's given the difference in B2B vs B2C mix (Anthropic 90%+ B2B vs OpenAI 60%).

代币经济模型以第一方和第三方为基础，估算了 AI 实验室在编码相关的支出，同时评估了应用层（如 Cursor、Lovable、GitHub Copilot 等公司）的 ARR 和利润率，以帮助投资者和企业追踪该垂直领域及其内部增长。我们认为，当前 OpenAI 和 Anthropic 超过 70%的 ARR 可归因于编码用例，而 Anthropic 的支出高于 OpenAI，原因是两者在 B2B 与 B2C 构成上的差异（Anthropic 的 B2B 占比超过 90%，而 OpenAI 为 60%）。



Still, cheap tokens are in widespread demand. We see massive growth in spend across both the for the Token-as-a-Service (TaaS)/API Endpoint Market for both frontier models and open-source models. Our estimates for AWS Bedrock this quarter drive our total AWS growth rate number well above street. The Tokenomics Model also forecasts massive demand for TaaS providers like Together, Fireworks, Baseten, and others who make up over $4B of ARR today.

尽管如此，廉价代币的需求依然广泛。无论是在前沿模型还是开源模型的 Token 即服务（TaaS）/API 端点市场中，我们都能看到支出的大幅增长。本季度我们对 AWS Bedrock 的估算数据，使得 AWS 的整体增长率远超市场预期。Tokenomics 模型还预测，像 Together、Fireworks、Baseten 等 TaaS 提供商将面临巨大需求，这些企业目前的年经常性收入（ARR）已超过 40 亿美元。



Source: SemiAnalysis Tokenomics Model, TaaS Tab

来源：SemiAnalysis 代币经济模型，TaaS 标签页

Thus, to say, our work suggests that headlines are overblown, enterprises continue to spend, and new use case/verticals for demand and token consumption are keeping the AI train moving forward at an aggressive pace.

因此，我们的研究表明，媒体的夸大报道并不属实——企业仍在持续投入，而新的需求场景和垂直领域带来的代币消耗，正推动着 AI 列车以迅猛的速度继续前行。

Budgeting: Pick a Number, Any Number

预算：选个数字就行

Most companies we talked with (n>50), also impose a hard cap on AI usage - though there doesn't seem to be a token number that companies are converging towards. At the lower end, we spoke with the head of AI at a top 3 US aerospace and defense manufacturer and one of the world's largest pharmaceutical companies which cap employees at $250 and $500 a month, respectively. At the higher end, our conversations with companies like Workday and Stripe revealed that their employees'

budgets are about $2000 a month.

在我们走访的 50 多家企业中，大多数都对 AI 使用设置了硬性上限——尽管各公司并未趋同于某个特定的代币数值。在预算较低的一端，美国排名前三的航空航天与国防制造商及全球最大制药公司之一的 AI 负责人告诉我们，他们为员工设定的月度上限分别为 250 美元和 500 美元。而在预算较高的另一端，与 Workday 和 Stripe 等公司的交流显示，员工月度预算约为 2000 美元。

A subset of companies has not yet imposed any limit at all, for what seems like two possible reasons:

一部分公司尚未施加任何限制，原因可能有两个：

1. Employees' access to AI tools is so limited that cost simply isn't a concern; or

   员工对 AI 工具的使用权限非常有限，以至于成本根本不构成问题；或者

2. The company has derived enough additional output from employees to justify the spend.

   公司从员工身上获得了足够的额外产出，足以证明这笔支出的合理性。

The financial industry is a clear example of the former. Much of it has been slow to adopt AI, and those that have moved have done so minimally. In conversations with various data scientists, analysts, and AI leads across a range of asset managers, regional banks, and auto-finance firms, the same pattern kept surfacing: employees boxed into the tools provided through Microsoft's platform.

金融行业是前者的典型例子。该行业大部分领域对 AI 的采用进展缓慢，即便有所行动的企业也仅停留在极小规模。在与多家资产管理公司、地区性银行及汽车金融公司的数据科学家、分析师和 AI 负责人交流时，同样的模式反复出现：员工被局限在微软平台提供的工具中。

The ROI, where it exists, can be dramatic:

投资回报率（ROI）在可行的情况下效果显著：

- A recruiter at Amazon responsible for scouting and placing principal engineers within the company noted that the process from initial screening call to team

placement used to take 6-9 months, but with AI tools to take interview notes and create reports, that timeline been cut in half.

亚马逊的一位招聘专员负责为公司物色并安置首席工程师，他表示从初步筛选面试到团队安置的流程过去需要 6-9 个月，但借助 AI 工具记录面试要点并生成报告后，这一周期缩短了一半。

- An employee at a data and analytics provider serving 85% of the Fortune 500 said what used to take them a week to do, can be done in just a few hours now.

一家为 85%的《财富》500 强企业提供数据与分析服务的员工透露，过去需要一周完成的工作，如今只需几小时即可完成。

The most mature companies are implementing a soft limit, which should be viewed by employees as guidelines rather than a hard rule. At a public cybersecurity company, the Director of Analytics, who oversees all developers and data scientists, said they set a “limit” of $800 a month for juniors and anywhere from $1,600-$4,000 a month for more senior staff. Data scientists are given the largest budget, as they tend to work with larger datasets requiring more tokens, a pattern that recurs across companies with flexible budgets. Should employees exceed their allowance, managers are alerted to have a conversation rather than cutting them off until the count resets.

最成熟的企业正在推行弹性限额制度，员工应将其视为指导性建议而非硬性规定。一家上市网络安全公司的分析总监（负责管理所有开发人员与数据科学家）表示，他们为初级员工设定了每月 800 美元的"限额"，高级员工则为每月 1600 至 4000 美元不等。数据科学家拥有最大预算额度，因为他们通常需要处理更大的数据集、消耗更多 token——这一模式在实行弹性预算的公司中反复出现。若员工超出预算额度，管理人员会收到通知并与其沟通，而非直接切断权限直至额度重置。

The Art of Token Conservation

Token 节约的艺术

With rising token spend and increasing executive consciousness towards that spend, employees are starting to learn to adapt. When that same aerospace and defense manufacturer first introduced its $250/month limit, some of the power users burned through it in four days. Employees can now request a higher budget, but that was not always an option. Employees now must get creative to conserve tokens. The company has disclosed that although employees have access to Claude, it has “turned-off” Opus 4.8 and Fast-Mode deeming it unnecessary. Management believes that handing employees larger token budgets would push them to automate tasks that shouldn’t be automated at all, like writing emails. We believe this anti-automation view by management teams is naive. Email, much like Slack with connectors from the AI Labs, will become more automated and AI native over time enabling greater productivity, visibility, and collaboration throughout the organization.

随着代币支出的增加以及管理层对此意识的提升，员工们开始学着适应。当那家航空航天与国防制造商首次推出每月 250 美元的使用上限时，部分重度用户在四天内就耗尽了额度。如今员工可以申请更高预算——但并非一直如此。现在员工必须想方设法节约代币。该公司透露，虽然员工可以使用 Claude，但已"关闭"了 Opus 4.8 和快速模式，认为这些功能并无必要。管理层认为，给员工更大的代币预算会促使他们自动化那些根本不该自动化的任务，比如撰写邮件。我们认为管理层的这种反自动化观点过于天真。与接入 AI 实验室连接器的 Slack 类似，邮件功能将逐步实现自动化和原生 AI 化，从而提升整个组织的生产效率、可见性和协作能力。

One global travel-tech company took a slightly lighter-touch approach. With 800 engineers out of 1,500 total employees who collectively spend a little under $10M a year on AI, it recently switched the default Claude model for all staff from Opus to Sonnet. Opus remains available but using it now requires a conscious choice. Most employees have a $200/month budget by default, though it can be increased to tens of thousands of dollars depending on seniority and role, with expected budget increases

coming soon.

一家全球旅游科技公司采取了较为宽松的策略。该公司 1500 名员工中有 800 名工程师，团队每年在 AI 上的总支出略低于 1000 万美元。近期，它将所有员工使用的 Claude 默认模型从 Opus 切换为 Sonnet。Opus 仍可使用，但需主动选择。大多数员工的默认月预算为 200 美元，根据职位和资历可增至数万美元，且预算即将上调。

As more companies impose limits, many employees are unbothered as they don't come close to the cap. Our conversations with customers at the Databricks AI Summit and in research calls showed that at many organizations most employees do not come close to the limit.

随着越来越多的公司设定限制，多数员工并不在意——因为他们远未触及上限。我们在 Databricks AI 峰会及研究访谈中与客户的交流表明，许多组织中大部分员工的用量远低于限额。

This mirrors the power usage we see here at SemiAnalysis where a handful of employees spend 4 to 5 figures per day, and others spend close to 0. Those who approach or exceed it have found ways to stretch their tokens. We have not impossed limits as the top performers are the ones utilizing a lot of tokens.

这反映出 SemiAnalysis 内部的算力使用模式：少数员工每日消耗四到五位数的 token，而其他人几乎零消耗。那些接近或超出限额的员工已找到扩展 token 使用的方法。我们并未设置硬性限制，因为顶级表现者正是高消耗 token 的核心用户。

Employees of companies on the Microsoft 365 Enterprise subscription receive free, unlimited access to the standard Copilot chatbot. Because that usage isn't tracked against the monthly budget, employees can game the system by using Copilot's 365 chat to draft and synthesize ideas first, before spending metered tokens on Claude or Codex.

微软 365 企业版订阅用户可免费无限使用标准 Copilot 聊天机器人。由于这部分使用量不计入月度预算，员工能通过先使用 Copilot 的 365 聊天功能草拟和整合创意，再消耗计量令牌调用 Claude 或 Codex 的方式钻系统空子。

AI As Headcount Lever AI 作为人力杠杆

The cost of AI tools sits with the company, but so does the benefit. A big 3 US airline stands apart from the others we met with in how it budgets. Its token allocation is tied to the specific project and the revenue that project is expected to bring in. When a project comes in, the financing team decides what percentage of revenue is set aside for expenses like travel and contractor fees, and now that same budget must cover token usage as well.

AI 工具的成本由企业承担，但收益也同样归属于企业。一家美国三大航空公司在预算管理上与我们接触的其他企业截然不同：其令牌分配与具体项目及其预期营收直接挂钩。项目启动时，财务团队会确定将营收的百分之多少用于差旅费、承包商费用等开支，如今这项预算也需涵盖令牌使用成本。

This reframes AI spend entirely. Companies that hand out generous allowances do so on the expectation that employees will work faster. Output expectations rise to match spend, and many workers have found themselves putting in even longer hours than before. Across over 50 conversations with medium to large enterprises, one thing became clear: the headline Uber, Meta, and other Fortune 500 tokenmaxxing stories were a result of poor incentives and lax oversight rather than an absence of high ROI activities/projects to spend those tokens on. The clearest expression of the ROI dynamic is at Amazon. Despite its widely reported layoffs, the company is hiring at a much faster pace due to efficiencies unlocked by AI tools.

这完全改变了企业对 AI 支出的看法。那些慷慨发放津贴的公司，是期望员工能因此提高工作效率。产出预期与支出同步增长，许多员工发现自己甚至比以前工作时间更长。在超过 50 次与中大型企业的对话中，有一点变得清晰：Uber、Meta 及其他《财富》500 强企业那些关于大量消耗 token 的新闻头条，根本原因在于激励措施不当和监管松懈，而非缺乏高投资回报率的项目来消耗这些 token。亚马逊最清晰地体现了这种投资回报率动态。尽管其大规模裁员的消息广为人知，但得益于 AI 工具带来的效率提升，该公司正在以更快的速度招聘员工。

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。

Budgeting Here to Stay 预算管理将长期存在

Caps, soft limits, and conservation tricks are all ways of managing the bill at the end of the month. In contrast to the high numbers of token spend in the news flow, Anthropic's own documentation says the average Claude Code usage per developer is between $150-$250 per month, and only 10% of users spend over $30 per day. One thing is clear after our on-the-ground work talking with customers: there is not a material risk present to 2H26 AI budgets and we expect Anthropic and OpenAI's API business to continue to grow at their current net new rates m/m for the foreseeable future.

上限、软限制和节约技巧都是管理月末账单的方式。与新闻报道中高额的代币支出数据形成对比的是，Anthropic 自身的文档显示，每位开发者每月使用 Claude Code 的平均支出在 150 至 250 美元之间，仅 10%的用户每日花费超过 30 美元。通过实地与客户的交流，我们明确了一件事：2026 年下半年的 AI 预算并未面临实质性风险，我们预计 Anthropic 和 OpenAI 的 API 业务将在可预见的未来保持当前每月环比增长率。

To learn more about the Tokenomics Model, including our estimates on AI Labs and Hyperscaler financials, email sales@semianalysis.com

如需了解更多关于代币经济模型的信息（包括我们对 AI 实验室和超大规模企业财务的估算），请发送邮件至 sales@semianalysis.com

Summary of Selected Conversations 精选对话摘要

We've added below a summary of selected conversations with enterprises:

我们在下方附上了精选对话摘要。

Large Cap HR Software Company

大型企业级 HR 软件公司

- Cursor Budget of $75 per day

  Cursor 预算为每天 75 美元

- Using Gemini Enterprise Plan but have not exceeded budget there

  使用 Gemini 企业版但尚未超出其预算

- Don't use Anthropic models for work, only for product functions

  工作中不使用 Anthropic 模型，仅用于产品功能

$300B AUM Vancouver Based Asset Manager

管理着 3000 亿美元资产的温哥华资产管理公司

- Use Claude everyday 每天使用 Claude
- Company tracks usage as a % of days in a month which went from 50% to 90/100% in the past 6 months

  公司以每月使用天数占比来衡量使用情况，过去 6 个月该比例从 50%上升到 90%至 100

- Mostly uses sonnet as it's sufficient but will use opus for coding tasks

  主要使用 Sonnet，因为它足够用，但在编程任务时会使用 Opus

Megacap Online Retail and Cloud Provider – Recruiting Org

大型在线零售与云服务提供商——招聘部门

- Uses mostly for interview notes, creating reports, etc.

  主要用于面试记录、生成报告等

- Entire recruitment process used to take 6-9 months but now only takes 3-4 months

  整个招聘流程过去需要 6-9 个月，现在缩短至只需 3-4 个月

- Hiring people at a much faster rate now

  目前以更快的速度招聘人才

- Optional ai sessions every 1-2 weeks hosted by company to teach people how to use things like cursor

  公司每 1-2 周举办可选的 AI 课程，教授员工如何使用 Cursor 等工具

Big 3 US Wireless Carrier – Database Developer

美国三大无线运营商之一——数据库开发人员

- Still exploring Claude Code for database development

  仍在探索将 Claude Code 用于数据库开发

- Mostly use ChatGPT for creating reports and making dashboards

  主要使用 ChatGPT 来创建报告和制作仪表盘

Large Dutch CPG and Health Technology Company

大型荷兰消费品与健康科技公司

- Budgeting to be increased in these upcoming months because of increased model pricing

  由于模型定价上涨，未来数月内将增加预算额度

- Current limit is 20/30k tokens a day

  当前每日限额为 20-30 千个 tokens

- CoPilot has unlimited chatting because of 365 user group so use that to chat and synthesis before using Claude or Codex to save token count

  CoPilot 因 365 用户组支持而具备无限对话功能，建议在调用 Claude 或 Codex 前，先使用 CoPilot 进行对话与内容整合以节约 tokens 消耗

Large Private Data Warehouse Vendor – Sales Org

大型私有数据仓库供应商 - 销售部门

- Personal spend around $300-400/month with no hard caps

  个人每月支出约为 300-400 美元，无硬性上限

- Mostly use for building decks, reports, and simulation

  主要用于制作演示文稿、报告和模拟工作

Large Legal Data and Risk Solutions Provider

大型法律数据及风险解决方案提供商

- Her role specifically gets $2000 a month for tokens but some other roles (operations) get $200

  她的岗位每月专门获得 2000 美元的代币预算，但其他岗位（如运营）每月只能拿到 200 美元。

- Turned a week of work into a few more hours but company expecting a lot more work to be done as a result and she's working more than before

  原本一周的工作量被压缩到几小时就能完成，但公司却因此期待她完成更多工作，结果她比之前更忙碌了。

- Mostly relies on enterprise chat bot using Claude to build reports

  主要依赖企业聊天机器人使用 Claude 来构建报告。

Public Non-Profit Focused Software Company

一家专注于非营利领域的公共软件公司

- No widespread token budget caps; use Codex, GitHub Copilot, and Claude in dev org

  没有普遍的 Token 预算上限；在开发组织中使用 Codex、GitHub Copilot 和 Claude

Large Network Security Company

大型网络安全公司

- $200/week budget for juniors, $400-$1,000/week for full-time roles

  初级人员预算为每周 200 美元，全职岗位预算为每周 400 至 1000 美元

- Spending limits are guidelines rather than hard caps

  支出限额更像是指导方针而非硬性上限

Large Travel-Tech Company

大型旅游科技公司

- Limit of $200 per person by default but can go up to 10s of thousands depending on the role

  每人默认 200 美元的限额，但根据角色不同，最高可达数万美元

- Codex used companywide but Claude for engineers only

  全公司使用 Codex，但 Claude 仅供工程师使用

- Spend is less than $10m a year but close to it

  年度开支不到 1000 万美元，但已接近这个数字

- Budget has increased a lot in past 6 months and may need to increase even more now

  过去六个月预算大幅增长，现在可能还需要进一步增加

- Default model used to be Opus, but they switched it to Sonnet now

  默认模型原本是 Opus，但现在已切换为 Sonnet

- Opus still available but need to make conscious decision to switch

  Opus 仍然可用，但需要主动决定切换

Large Global Oil Company 大型全球石油公司

- Buy AI through Databricks and Microsoft

  通过 Databricks 和 Microsoft 购买 AI 服务

- No Claude or Codex use due to sensitive data policies

  因敏感数据政策，未使用 Claude 或 Codex

- GitHub Copilot available for employees but requires training which most people are too lazy to do

  GitHub Copilot 已对员工开放，但需要进行培训，而大多数人都懒得去做

- AI budgeting is done at the company level, not team level

  AI 预算管理是在公司层面而非团队层面进行的

Large Pharmaceutical Company – Data Analytics

大型制药公司 – 数据分析

- $500 Monthly Limit throughout org

  整个组织每月 500 美元限额

- Can get $1,000 approved on a case-by-case basis

  可根据具体情况申请批准 1000 美元

- Mostly uses Claude; has access to ChatGPT but no one uses

  主要使用 Claude；虽可使用 ChatGPT 但无人使用

Big 3 US Airline 美国三大航空公司之一

- $1000 Github Copilot for devs, $300 for analysts

  开发人员使用的 GitHub Copilot 定价为 1000 美元，分析师为 300 美元

- Used to be unlimited but need to start budgeting soon

  过去不限制使用，但很快需要开始预算管理

- Budgeting varies by team and project

  预算分配因团队和项目而异

  - ex. $10m rev project and financing team approves $1m for expenses and then team allocates a % of that to tokens at their discretion but that budget is for everything including other tools

    例如：一个收入 1000 万美元的项目，财务团队批准了 100 万美元的支出预算，随后由团队自行决定将其中一定比例分配给代币使用，但这笔预算需覆盖所有开销，包括其他工具

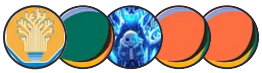
Recommend SemiAnalysis to your readers

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



96 Likes 96 个赞 · 2 Restacks 2 次转发

Previous 上一篇

Discussion about this post

关于此帖的讨论

Comments **Restacks**

Write a comment...

V3 4h

$PLTR Evolve

LIKE REPLY 回复 SHARE