SemiAnalysis

TokenBudgeting: Our Conversations with Enterprises on Token Spend

2026-07-01T02:32:39.225000+08:00 | Crystal Huang, Joey Brookhart, Dylan Patel | only_paid | accessibleTextLength=12806

URL:

https://newsletter.semianalysis.com/p/tokenbudgeting-our-conversations

Summary:

Was Widespread TokenMaxxing Ever Really Here?

Accessible page text:

TokenBudgeting: Our Conversations with Enterprises on Token Spend Was Widespread TokenMaxxing Ever Really Here? Crystal Huang , Joey Brookhart , and Dylan Patel Jun 30, 2026 · Paid 144 6 6 Share 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. 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 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. Companies are downgrading default models and turning off premium tiers while employees· game subscription M365 Copilot usage to stretch their token allowance. Rise and Fall of Tokenmaxxing 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. The dashboard was shut down 2 days later after The Information reported the spend. 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. SemiAnalysis is a reader-supported publication. To receive new posts and support our work, consider becoming a free or paid subscriber. Subscribe Our View of the Data & Narrative There are many news stories out there on tokenmaxxing and resulting budget blowouts. However, in our work in the Tokenomics Model , we estimate that 90 th + 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. 99 th percentile customers spend almost $90,000/yr per employee while 90 th 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. Source: Ramp Economics Lab 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 (90 th -99 th percentile customers who are seeing their revenue accelerate too). 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. The Tokenomics Model estimates coding related spend at the AI Labs on a 1 st and 3 rd 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%). 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. · Source: SemiAnalysis Tokenomics Model , TaaS Tab 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. Budgeting: Pick a Number, Any Number 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. A subset of companies has not yet imposed any limit at all, for what seems like two possible reasons: Employees· access to AI tools is so limited that cost simply isn·t a concern; or The company has derived enough additional output from employees to justify the spend. 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. The ROI, where it exists, can be dramatic: 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. 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. 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. SemiAnalysis is a reader-supported publication. To receive new posts and support our work, consider becoming a free or paid subscriber. Subscribe The Art of Token Conservation 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. 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. 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. 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. 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. AI As Headcount Lever 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. 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. SemiAnalysis is a reader-supported publication. To receive new posts and support our work, consider becoming a free or paid subscriber. Subscribe Budgeting Here to Stay 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. To learn more about the Tokenomics Model , including our estimates on AI Labs and Hyperscaler financials, email sales@semianalysis.com Summary of Selected Conversations We·ve added below a summary of selected conversations with enterprises: This post is for paid subscribers Subscribe Already a paid subscriber? Sign in Previous Next © 2026 Dylan Patel · Privacy · Terms · Collection notice Start your Substack Get the app Substack is the home for great culture

Generated from SemiAnalysis public archive/page text available without authenticated rendering.

