AI Research

The Ramp in Token Optimization

23 June 2026

Summary

The AI trade has powered through a number of hurdles over the last few years, from the DeepSeek moment to concerns about compute AI capacity and the rate of enterprise adoption. In this note, we address the latest hurdle – push-back to the sheer cost of AI, as enterprises begin to optimize AI token consumption post a surge in use or "token-maxxing". We first addressed this issue in a note two weeks ago and in this follow-up report, we update our thoughts and review the key investor debates and impacts on each layer of the AI stack. Bottom line, the spike in token or AI compute spend optimization could temporarily weigh on AI revenue growth, but the secular usage trends remain very strong and intact.

Moderate Risk to the AI Trade

Rising AI token /compute costs have now become a real concern for most organizations, and we offer anecdotes from 10 more enterprises to supplement what we offered in our initial report on the subject two weeks ago (see here). Most organizations are introducing some form of guardrail to monitor or control AI token usage, with an emerging trend to down-tier/pivot to cheaper open-source Chinese models (such as Alibaba's Qwen models) for tasks that don't require the high-end closed-source models. Token optimization and a mix shift to cheaper SLMs or open-source models raises the risk profile of the growth curves for the AI ecosystem, but for now, we remain measured about this risk and are not ringing the alarm bells. Token costs are rising because AI usage is rising – it's a healthy problem. Some measure of AI spend optimization is normal, no one is hitting the brakes on AI deployment and it is likely that we're sitting in front of new models trained on next-gen chips that might drive token costs down further.

Impact by Tech Sector

In our view, the biggest impact is at the model layer as cost-conscious enterprises down-tier to cheaper and less token-heavy frontier models, SLMs and open-source Chinese models (share gains for Alibaba, Z.ai, Moonshot and DeepSeek), especially for non-coding use cases. The AI model market is now more competitive and is bifurcating. With overall AI use still ramping materially, and hence compute demand still robust, we don't see material risk to the semi/hardware layer and enterprise token optimization efforts don't appear to have had any impact on AI capex growth. The hyperscalers are already multi-model platforms and are also facing a wall of AI demand, but we come out neutral on a view that token optimization could impact frontier model API revs growth and that AI price sensitivity could limit near-term GPU price increases. At the software layer, it's a mixed bag. Customers are likely to react to rising AI token costs by throttling other budget items (headcount growth as well as IT services and apps spend) but we wonder if software firms can gain an edge over the frontier AI labs by positioning themselves as model-neutral token optimizers and controlling internal costs by harnessing cheaper models.

Controlling AI Spend

Let's start with a discussion of token optimization trends before addressing the impacts across the tech sector. We noted in our initial report two weeks ago (see link here) that as enterprise AI use moves past simple and "token-light" chatbot tasks to more "token-heavy" AI agents for software coding and other tasks, a new source of enterprise adoption friction – the sheer cost of AI - has emerged as the AI topic du jour. We noted that based on a dozen+ conversations with enterprise IT execs over the prior several weeks, ~60% of enterprises were now in some manner throttling AI spend by putting some degree of guardrails in place. Measuring the degree of AI spending push-back – the length and magnitude of any spend optimization phase – is a critical subject for tech investors, and frankly broader equity markets, given that a material AI spending pull-back is likely not priced into AI-exposed stocks.

Initial View – Speed Bump, Secular Trend Intact

Our initial conclusion was that this was an emerging headwind but that the impact on the "AI trade" may prove to be modest. Our level of concern was tempered by several realities. First, many organizations are so early in their AI adoption journeys that "token-maxxing" was either not yet a thing or more advanced enterprises were expressing a reluctance to materially limit or throttle AI use beyond the obvious elimination of real waste – they are still focused on trying to drive AI adoption in their organizations to encourage employees to use these tools in order to drive innovation. In other words, it was simply too early-stage to be throttling use, and in many cases, these organizations were optimizing/cutting other areas of their IT spend to contain runaway AI costs. Second, many enterprises were expressing content with "token-maxxing" because it was accompanied by sufficient ROI or productivity gains, measured in metrics such as the improvement in lines of code contributed by programmers or in call deflections by customer support AI agents. In most cases, this behavior struck us as "normal" enterprise cost-containment actions at the early stage of a new technology adoption cycle, as no one – not a single customer check – was being an alarmist and talking about slamming on the AI brakes or throttling their AI initiatives. It felt like a spending speed bump.

Palantir and Databricks Weigh In

Since our initial token optimization note on June 8, we spoke with several participants in the AI ecosystem, including Palantir and Databricks. On the margin, their views added to our concern that an enterprise AI spending push-back was gaining momentum:

- **Palantir**: A month ago, largely in response to a discernible uptick in the number of Palantir's enterprise customers that were keen to better manage their AI model/token costs, Palantir commercialized a new tool – called AIP Evolve – that serves to help customers optimize their AI agent spend, by choosing the best AI model for the task (including using open-source cheaper models), tuning prompts and pulling in data more effectively. In one example, Evolve recommended a model swap that cut token costs by 97%. Palantir told us that Evolve achieved a 90% adoption rate in the first three weeks, strong evidence of token optimization interest.
- **Databricks**: Last week we also spoke about these trends with Databricks, which, like Palantir, has seen a material recent increase in customer focus on reducing token spend, citing "widespread" customer push-back to AI costs. We asked whether our initial description of this trend "small speed bump, secular spending trend intact" aligned with their views and the CEO responded that we're mostly right – "*the AI spending trend is indeed up and to the right, yes I would call this optimization trend a speed bump, but* ***it's a big speed bump****, not a small one*".

Recent Customer/Partner Checks

In our "Part 1" note on June 8, we published all of the enterprise commentary that we'd picked up to date about token spend optimization . The initial feedback was interesting/ incremental and we reacted by doubling down over the last two weeks to speak to even more organizations about their token spend. See below excerpts from our more recent conversations since that initial report. In general, these anecdotes reaffirm our initial view that token spend optimization has become a key issue in most organizations, resulting in a big spending speed bump for some organizations, but a smaller speed bump for others that are either too early in their AI deployments or are far deeper but are unwilling to throttle users because they see the offsetting ROI or have an organizational priority in place to drive innovation and hence AI use. These organizations are likely to optimize AI spend more gently, or later. In no case did we hear of companies making a wholesale pull-back on their AI initiatives, this is about being more efficient.

- **Check 1**: *AI is a core part of our product workflow, so the usage naturally grows over time. This means that more tokens will be used. So the question isn't whether to use tokens, it's how to use them efficiently. As a result, optimization becomes an ongoing engineering discipline rather than a reaction to a budget crisis. I think the industry is moving from a phase of experimentation to a phase of cost efficiency. Enterprises still want more AI usage, but they increasingly want predictable economics and measurable returns alongside that growth, so they have to manage the efficiency and the cost because at the end of the day, the users want performance. We look at how we make the most out of tokens, we want to do things as cheap as possible. But ROI on AI is rarely measured in terms of token consumption alone, you need to look at the business outcomes and workflow efficiency . If a model helps reduce development time and improves the output quality, it accelerates the research process. Token costs are one side of the equation , but productivity gains and output quality are also important. In practice, we are usually balancing three variables, which is quality, latency and cost. The goal isn't necessarily to minimize token usage, but it's to maximize the value generated per dollar spent. So every $1 we are spending, how much are we making ? Token costs are definitely something that's monitored really closely. I wouldn't characterize our experience as hitting a wall and having to dramatically*
  curtail usage and stop people from using it because what we've seen is that as AI adoption matures, the teams and our team become more disciplined about matching the model to the task.

- **Check 2**: *At the start of the AI craze, our CTO told us that he wants everyone using every AI tool. We have 5 AI tools internally and all of the LLM products. Like others, we ran into the issue where we have already used most of our token budget for the entire year. Now we're only using 2 AI tools and being careful around usage. Internally, we're structuring it so that we use multiple LLM tools but have one that's preferred for each individual . We're using a lot of Anthropic Claude and also have Google Gemini in there . I think it'll ultimately come down to cost management for these model providers . Engineers will use any tool they can with all of the bells and whistles, so it'll ultimately come down to who's the most cost effective. You can actually get away with using Anthropic Sonnet for a lot of engineering work. You can even do some things with Haiku, so we're using Opus 4.8 only for the most complex projects. We haven't checked out Fable yet since it's too early.*

- **Check 3**: *We haven't run into the issue of token-maxxing impacting our budget yet, but that could change. Since we started rolling out AI products internally, there's been an emphasis on responsible usage. We didn't have a token leaderboard or anything along those lines. I think one of the challenges companies have been facing, and why they're running into token issues, is that everyone now has the power to code. For example, someone that's a product manager can now turn a spreadsheet into a live website or dashboard, and they become addicted to the potential things you can build with these tools. You combine that incremental usage with software engineers using coding agents, and that's how you blow through your token budgets.*

- **Check 4**: *We have not run into the issue of token-maxxing severely increasing our costs, but that's because we have seat-based pricing and also limit access to the smartest models like Opus.*

- **Check 5**: *We are seeing a spike in token costs. I see a dashboard with spend by user and some users are spending tens of thousands a month. One guy hit $35k in a single month. All this is through AWS Bedrock and is pushing up our AWS spend. I haven't heard about any pushback to this yet. Maybe the organization has said something to him but not through me. I'm trying to push my team to be conscious of how they're using these tools so they don't end up on this list, like by making sure the context window isn't filling to a million tokens every time.*

- **Check 6**: *I'm on the DevOps and infra team. We're heavy users of AI by now. I use it every day. We have both Microsoft GitHub Copilot and OpenAI ChatGPT. I'm routinely using 100-200% of my weekly token allotment and haven't heard any concern yet.*

- **Check 7**: *It feels as though customers are really going all-in on AI now. Last year, no one was talking about doing things with AI but it's the only thing to talk about now. Tokenization is becoming more prominent. I'm not yet seeing concern on token costs. If it's there, it's quite small today, but I imagine it will come. Customers are concerned about getting the ROI. They have big budgets to spend. As long as they are getting ROI, they aren't focused on the dollar amount. Some customers are even giving perhaps $1,000 in tokens to employees and seeing how much they use. That's then factored into performance reviews to get employees to use it.*

- **Check 8**: *I have heard some customers consciously thinking about how they can scale AI effectively while managing token costs. In many cases, this is not in response to having overspent but rather that they've heard about other customers doing so and wanted to make sure they didn't wind up there. Customers have lofty ambitions about AI. They're trying to scale these tools while managing headcount thoughtfully and controlling AI cost.*

- **Check 9**: *In terms of token spend, there is some tension between me as the CFO and our CEO. I want to more aggressively control AI model and compute spend, he wants to accelerate it. He set a goal at the beginning of the year to accelerate innovation, and the CEO went so far as to set employee compensation targets tied to AI use. So what do you think our employees did in reaction? They leaned heavily into AI tools, and token consumption has taken off. I am reacting by putting token usage caps in place and if employees need more they have to submit a request. At least this way I have some visibility and control. But given the CEO prioritization on AI and innovation, we are not throttling our employees in any meaningful way yet.*

- **Check 10**: *The most acute area of rising token use is in the AI coding arena, where we could see massive increase in token use. I am personally burning through millions of tokens a day in my spare time and evenings, so I think agentic software development in general will inevitably drive up usage of models massively. You also burn through a lot of tokens by pointing AI agents at large databases, this may be why large Databricks customers like us are seeing huge use of tokens. We'll have to manage this, by putting more loads on-premise, by moving to cheaper models or by cutting our IT spend elsewhere.*

Model Routing in Practice

Let's now discuss HOW companies are becoming more efficient consumers of AI, and what the key levers are. Many of the organizations have reacted like Check 9 above (or like Uber as we described in our previous note) by putting token usage caps on employees and forcing users to request additional token access, offering these firms better visibility and control. Other guardrails include pooling tokens across teams or departments, sending alerts ("be conscious of waste") to heavy token consumers, optimizing prompts (unnecessarily large context windows can invite huge and resource-intensive prompts) and caching/storing model output where appropriate (eliminating redundancy). In some cases, organizations have indicated that they are willing to allow developers to spend beyond their usage caps because the productivity gains from AI coding agents are relatively clear, and to throttle token use in other areas where the ROI is less obvious.

The AI token optimization technique that interests us (and investors) the most is model routing – dropping down to cheaper more token-efficient AI models where appropriate. To better understand the mechanics of model routing, we spoke last week to an AI-native company that actively routes requests to lower-cost AI models in an effort to control token consumption.

- *Our efforts to control AI costs include routing different tasks for different models, making sure that you are using premium models only when the additional performance justifies the cost. If it doesn't, then there's no need. Speed is a factor, but the main thing is cost control. Model selection has become much more dynamic. A couple of months ago, many teams were trying to standardize on a single provider. This is the best, let's go with this. Earlier on, a lot of our tasks would simply run everything through the most capable model without caring about other factors because we just wanted to perform to perform to perform. Over time, we learned that not every workload requires the most expensive model or the largest context window. But today, it's much more common to use different models for different jobs or tasks because the performance gap has been varying depending on the task that you perform. As you learn that, you start to economize. The focus shifts from maximizing token consumption to maximizing value per token.*

- *When it comes to routing, the general principle is that different models have different strengths. Some are stronger at complex reasoning and some are better at coding tasks or long context analysis, some are faster. Routing matters a lot because most sophisticated AI users aren't choosing a single model anymore. They are matching workloads to the most appropriate model in that environment. And they are checking the cost efficiency, which isn't just a property of the model itself. It's also a property of how intelligently you allocate workloads across models.*

- *When you evaluate the cost, we don't look at price per token in isolation. We look at what I would call an effective cost per successful outcome. So a model that is slightly more expensive but produces higher quality results on the first attempt can actually be more cost effective than a cheaper model that requires multiple iterations because this will mean that we use a lot more tokens. In practice, we evaluate a combination of output quality, latency, reliability and of course, the total cost. So if a model delivers comparable results at a lower cost, that's obviously attractive to us because we are also trying to reduce the cost that we are spending. And if it delivers better results, but at a modest premium, then that can also be attractive depending on the workflow that we are processing.*

Down-Tiering to Cheaper Models

As we note in Figure 1 below, an effort to route API model requests to the lower end of model suites can save materially on token costs, although they may not of course offer the same performance. In many cases, as noted below, premium-end models are 5x the cost (on a millions of output tokens basis) compared to lower-end models. The AI model providers each offer a suite of models, precisely because not all workloads require the highest-tier model and because organizations have very different budget considerations. This desire for cheaper, but good-enough models, is the opportunity that Microsoft is now addressing via its recently-launched suite of MAI "small language models" or SLMs. Microsoft described its general-purpose MAI "Thinking" model as being in the "medium-sized weight class" (35 billion parameters) and its Code-1 coding model as being more like a lower-end frontier model.

Claude Model	Input ($/M Tok)	Output ($/M Tok)
Fable 5	$10.00	$50.00
Mythos 5	$10.00	$50.00
Opus 4.8	$5.00	$25.00
Opus 4.7	$5.00	$25.00
Opus 4.6	$5.00	$25.00
Opus 4.5	$5.00	$25.00
Sonnet 4.6	$3.00	$15.00
Sonnet 4.5	$3.00	$15.00
Haiku 4.5	$1.00	$5.00

Source: Company Disclosures

Mix Shift to Open Chinese Models

The desire to harness less-expensive models is now going beyond down-tiering within the frontier model suites. In our conversation with Databricks last week, the company said something that surprised us – that over the last few months they had begun to see a marked uptick in interest among its enterprise customers in open-source Chinese models. We had been under the impression that this was a phenomenon among AI-native tech firms (such as Cursor developing a 1P model by utilizing an open-source Kimi model) or smaller, more price-sensitive organizations but not the type of F500 enterprises that Databricks serves (except a few examples such as BMW that have disclosed a partnership with Alibaba to use open-source Qwen models). We immediately reached out to the CTO of a large global bank, which we assumed would be about the
last type of organization that would go down this route. Instead, we heard:

- *To manage our AI token spend, we're now hosting models like Qwen [UBS: a family of open-weight AI models developed by Alibaba] on-prem to balance our use of premium models like Claude. This is not something we have to pay token-based charges for, and we just have to size our on-prem hardware to meet demand. We won't go as far as to use the cheaper Chinese externally-hosted options, but we will use respectable internally-deployable models like Qwen. These are definitely a good option, as they are very close in terms of results to the recent Claude models. We will have to spend a bit more on GPUs and make lots of use of on-prem Qwen to control our AI costs.*

Microsoft and AWS Use of Chinese Models

Importantly, AWS Bedrock already offers a drop-down menu of open-source Chinese models (see Figure 2 below), but we had assumed this was not a popular option for most large organizations. Readers may have also seen a media story just last week (see link here) about Microsoft considering offering customers the option of using DeepSeek models in its new Copilot Cowork product (fine-tuned, with additional safeguards). We reached out to Microsoft to confirm, and they responded by saying "*fair to assume we'll continue to pursue a multi-model strategy, leveraging the strengths of different models to optimize both performance and cost. We've already made DeepSeek available through Azure AI Foundry, and we'll continue evaluating a range of models as we bring the best AI innovation into our first-party experiences* ". That sounds like a yes to us. That's important, as Microsoft and AWS would never offer Chinese models to US-based customers unless they considered them to be safe to deploy in a controlled (Azure or AWS-hosted, or in the case of the bank above, on-premise) environment. As such, any material mix shift to cheaper open-source Chinese models could have the effect of lifting demand for hyperscaler services such as AWS Bedrock and Azure Foundry.

Figure 2: Models Available on AWS Bedrock

Company	Models
Anthropic	Opus 4.8, Opus 4.7, Sonnet 4.6, Haiku 4.5
OpenAI	GPT 5.5, GPT 5.4
Google	Gemini 3
xAI	Grok 4.3
Nvidia	Nemotron Nano 3, Nemotron Super
Mistral	Devstral 2123B, Large 3, Ministral 3, Magistral
Amazon	Nova 2 Lite, Nova Micro, Nova Pro, Nova Premier, Nova Sonic
Meta	Llama 4 Scout, Llama 4 Maverick
Cohere	Embed v4, Rerank 3.5
MiniMax	M2, M2.1, M2.5
Kimi (Moonshot)	K2 Thinking, K2.5
Qwen (Alibaba)	Qwen3 32B, Qwen3 Coder, Qwen3 Next
DeepSeek	R-1, V3.1, V3.2
GLM (Zhipu AI)	GLM 4.7, Flash, GLM 5

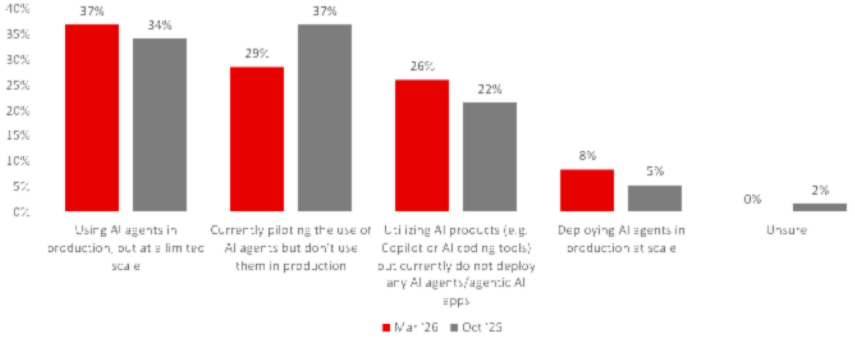
Source: Company Disclosures

The Secular AI Trend is Intact

Yes, enterprises are absolutely moving down the model cost curve and routing model requests to cheaper closed-source and increasingly open-source Chinese models. Token optimization is now prevalent, especially in mainstream enterprises, such that there is certainly SOME pressure on token growth and hence the "AI trade." But we still want to stay measured here and are reluctant to sound the alarm bells, for a few reasons. First, many organizations are NOT yet materially throttling token use (in an effort to drive innovation and adoption) and/or are still very early in their AI deployments. As we show in Figure 3 below, our most recent survey of ~130 enterprises (see link here to our full enterprise AI survey) revealed that only 8% of enterprise respondents are deploying AI agents "in production at scale". It's still very early-stage. More heavy token consumers, such as AI-native tech firms, continue to rapidly scale their token consumption. Harvey, for instance, revealed in a recent podcast (see link to story here) that its token consumption has gone from 1 trillion in the month of January to 12-13 trillion in May.

Moreover, it still seems as if we're sitting in front of a continued innovation curve in terms of AI model releases. The AI ecosystem has only over the last few quarters stood up material Nvidia GB200/GB300 capacity and hence the pending models trained on these next-gen chips should have even better price-performance and hence even lower unit token costs, helping to mitigate some of the AI sticker shock that enterprises are experiencing. Bottom line, this optimization wave looks a lot different than the post-COVID optimization wave that impacted the hyperscalers and usage-based software firms in 2022-2024, largely because the AI compute cycle is still so early-stage.

Figure 3: Which one of the following statements best describes the way your company is currently using AI agents?



Source: UBS Evidence Lab (> Access Dataset)

Token-Related Investor Debates

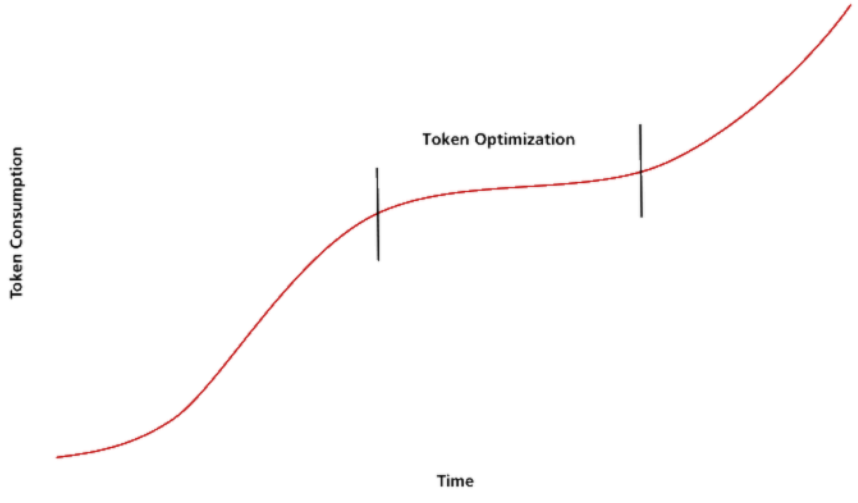
Our token optimization note from a few weeks ago kicked-off a number of investor conversations about the subject and in this section, we'll run through these debates and share our (still-preliminary) views.

Quantifying the Token Growth Deceleration

It seems very reasonable to conclude that the collective impact of many token optimization actions across enterprises, as we've highlighted in this note and in the one from two weeks ago, will at least temporarily weigh on the overall token consumption growth curve. But by how much and for how long? This is the tough question, as there is not a lot of reliable data out there on token consumption and frankly, many organizations we speak to don't even know what their token spend is (this lack of visibility will surely be solved by better token monitoring software). We've spoken to several investors that have referenced the "Tokenomics" paper published by an investment firm, which discusses the overall token optimization trend, calls out (as we have) the mix shift to cheaper models and attempts to measure the deceleration in token consumption growth.

We don't yet have an accurate methodology to track and measure AI token consumption, and hence are relying for now on the qualitative feedback from our customer checks as well as vendor discussions ("a big speed bump" as per Databricks). Our best guess is that if most token-minimizing organizations were spending $100 on AI tokens before and this was on a path to being $150 in a few months, the qualitative commentary about token optimization suggests that they might land at $120-130 instead. We simply did not hear any "slam on the brakes" commentary to make us think that $100 will end up at $80. A directional token consumption chart for these token-minimizers might look like this:

Figure 4: Hypothetical Token Consumption Growth



Source: Estimates

What Might Get De-Prioritized?

Enterprises don't have unlimited budgets, and unexpectedly-high AI model and compute costs typically need to be funded by pulling back on something else. What? A few weeks ago we published an update on IT Spending trends (see link to report here) and in that report we highlighted efforts of ~25 organizations to offset rising AI costs via a variety of levers, including tempering headcount growth, pulling back on the use of external IT services firms and limiting growth in their SaaS/application software spend. The Uber CEO, for instance, has been clear that Uber remains full steam ahead in terms of its planned AI use, but that they're looking to offset rising AI token costs by hitting the brakes on internal headcount growth. We're hearing such anecdotes more these days, and believe that the uninspiring results of late from IT Services firms (Accenture through the India-based firms) as well as SaaS companies (Salesforce, Workday) reflects this spending de-prioritization. In our judgement, many organizations are early-stage in responding to rising AI costs and are, as a first step, trying to become more efficient by utilizing cheaper models and other methods described above. We suspect that many will pull those efficiency levers first BEFORE cutting elsewhere, suggesting that the demand pressure on external IT services and SaaS/application software spend might get worse in 2H26/1H27.

Push-Back to AI Pricing?

Needless to say, periods in which customers are absorbing sticker shock are not the best time to be hitting these enterprise customers with price increases on AI models, compute and/or applications. How might this manifest itself?

- **AI Models**: Investors are now asking whether the AI cost push-back and the growing pivot to cheaper (even Chinese open-source) models could motivate the AI firms to limit or even cut per-token model pricing (a risk highlighted by a WSJ story, see here).
- **Infra Providers**: Some investors are now asking if the enterprise push-back to token costs and/or price cuts by the frontier labs might weigh on GPU cloud spot pricing, dissuading providers of GPU or other AI infrastructure (the hyperscalers, the neo-Clouds) from raising per-hour cloud GPU costs as many investors have been expecting (and as the stocks of the neo-Clouds increasingly reflect). Fair question, and a negative outcome if this curtails cloud infra pricing power, but in our view the pricing push-back is more at the model layer and GPU pricing power remains high given the wall of compute demand and how early-stage AI penetration still is.
- **App Software**: Just as many enterprises are putting in controls to better manage their AI token spend, and considering ways to cut non-AI spend to offset token- maxxing, most application software firms are trying to push seat-to -usage pricing model transitions that are intended to be revenue accretive for the software firms, meaning they result in a net spending increase by customers. We wonder whether this spike in enterprise token cost optimization efforts, not to mention the higher level of overall uncertainty about AI technology trends, has created a tough environment for software firms to be pushing a new SaaS pricing model.

Who Wins/Loses from Token Optimization?

Let's address the impacts of this surge in token optimization efforts across the AI and tech ecosystem:

Chinese Model Providers

This emerging mix shift and openness to cheaper Chinese models could obviously favor the likes of Alibaba (Qwen models) and Z.ai (GLM models) as well as MiniMax, DeepSeek and Moonshot (Kimi models). Their models are typically open-source and free, limiting the monetization upside, but we could see more deals akin to the BMW-Alibaba agreement for access to Qwen models. Some investors have asked whether a rotation to Chinese AI models could invite US government push-back or restrictions, noting the recent news (see link here) about Airbnb's use of Alibaba's Qwen models (to power its customer support agents) motivating the US government to launch an inquiry. This may be a non-zero risk in light of Airbnb's experience, but in our view the use of Chinese AI models may already be too prevalent among US-based organizations to think that the US government can realistically restrict their use, especially if US-based firms are deploying them in controlled (on-premise, or AWS/Azure-hosted) environments.

Meta's Open-Source Llama Models

As noted, the UBS Tech team publishes an Enterprise AI survey every six months or so (see here for our most recent May 2026 version, based on March 2026 responses). In this survey we ask ~130 enterprises which AI model they are building on. As noted in Figure 5 below, the answers skew heavily to the Big 3 frontier model labs. Among the open-source alternatives, it is Meta's open-source Llama models that rank highest, with a 16% share compared to a combined 8% share for DeepSeek and other Chinese models (we expect this share to climb further in the next survey). What is interesting is that in our recent checks, customers reacting to unexpectedly high AI costs and pivoting to cheaper open-source AI models are increasingly turning to Qwen or DeepSeek models, far more so than to open-source Llama.

Ostensibly, this creates an opportunity for Meta to improve and externally commercialize its Llama suite, but Meta has clearly decided to move in a different direction via the shift away from open-source Llama to investments in its proprietary (closed-source) Muse Spark AI models to power Meta AI internally, "the first product of a ground-up overhaul of our AI efforts" according to Meta. This pivot has left a void in the AI model market that is being filled by open-source Chinese models as well as cheaper SLMs, including those announced just weeks ago by Microsoft.

Figure 5: Which if any of the following LLMs (large language models) are your company currently using? (select all that apply)

LLMs	Mar '26	Oct '25
OpenAI ChatGPT 5.0 – 5.3	54%	46%
Anthropic Claude 4.5-4.6	46%	
Google Gemini	38%	46%
OpenAI ChatGPT 4.0	34%	50%
Anthropic Claude 3.5	28%	34%
OpenAI ChatGPT 3.5	20%	22%
Open Source Llama	16%	16%
OpenAI o3	12%	11%
Custom-built model (e.g., we're building our own model)	10%	10%
Amazon Nova	9%	9%
Nvidia (NeMo, BioNemo, etc.)	9%	11%
Open Source Mistral	6%	8%
xAI Grok	5%	6%
DeepSeek	4%	3%
China-based open source models	4%	0%
Cohere	3%	4%
Stability AI	2%	4%
Other	5%	3%
Not applicable	2%	4%
Don't know	1%	0%

Source: UBS Evidence Lab (> Access Dataset)

AWS, Azure and Google Cloud

The hyperscalers are already "multi-model", offering a plethora of choice (from closed-source to open-source) to enterprises. They are already acting as "AI platforms". As such, a mix shift down the model pricing curve to SLMs and open-source models (whether Chinese or Llama) may not have a material impact on AWS and Azure, which are not solely tied to the fortunes of the frontier labs. That said, as we recently described in a report on the hyperscalers (see link here to our "The Hyperscaler Bull Case" report), AWS, Microsoft Azure and Google Cloud are now materially monetizing 1P and 3P model API access, as well as IP product demand, and any loss of share by the frontier model firms could skim off the growth rate of this new revenue source.

However, regardless of what models an enterprise customer elects to use, they still need

cloud-hosted inference compute. One might retort with the argument that a wholesale mix shift to cheaper and less token-heavy AI models for the same task results in LESS token and therefore compute demand for the hyperscalers. True, but this may be completely dwarfed by the fact that the hyperscalers (and neo-Clouds) are facing an acute demand-supply imbalance and have near-unlimited demand for AI compute resources. Moreover, it seems plausible that if the premium-priced frontier labs are now facing greater competition from cheaper models, they may react (as they did to the DeepSeek moment) with an even greater investment in training infrastructure - next-gen chips and improved algorithms - to drive unit token costs (and hence pricing) lower. That would obviously be a good thing for their compute partners.

A related and important investor debate has emerged about potential share shifts among the AI cloud infrastructure providers in light of the spike in AI cost sensitivity among customers. Greater price or token cost sensitivity should, all else equal, favor the lower token cost provider, which most investors believe is Google Cloud and AWS by virtue of their proprietary chips (TPUs, Trainium chips) and models (Google Gemini). It's tough to push back on this thesis, apart from countering that a) neither Google Cloud nor AWS are likely to take advantage of this edge by lowering AI compute pricing at a time of near-unlimited demand and b) Microsoft appears to be hustling to close this gap
by itself launching proprietary chips and AI models.

Baseten and Fireworks at the Inference Layer

A number of private AI-native firms have sprung up over the last few years to address the growing need for inference compute. Firms such as Baseten and Fireworks offer "production inference" layers on top of the AI infrastructure providers for mostly AI-native customers but increasingly enterprise customers that are running open-source and custom-built AI models. According to Baseten, 90%+ of inference compute spend is from the use of frontier models, with 5-10% from custom or post-trained open-source models. Baseten also noted that open-source models are about 90-days behind the closed-source frontier models in terms of performance (general view is they're 6-9 months behind), but can run 70-90% cheaper.

As such, the growth of firms such as Baseten and Fireworks can offer insights into the popularity of custom or open-source models, at least among AI-natives. According to Baseten, its inference layer now handles 30 trillion tokens per day, which the company said on a recent podcast may be greater than that of Google Gemini. Based on various sources, the combined revenues of Baseten and Fireworks are now over $1 billion, and we think this AI category will garner a lot more investor attention going forward. In mid-July at the annual UBS Private AI and Software conference in Menlo Park, we'll be hosting the CEO of Fireworks and will explore the impact of token optimization.

Software Firms as Token Optimizers

As noted, customers are likely (perhaps with a lag) to react to rising AI token costs by cutting elsewhere, reducing headcount growth, optimizing seats or trimming the growth in their application software budgets. None of that is good for the larger and well-penetrated seat-based SaaS firms such as Salesforce, ServiceNow and Workday. Moreover, as we flagged above, their transition to a seat-plus-usage model may now be met with even greater resistance. Software/infra firms such as Datadog and Oracle have outsized exposure to the success of the frontier labs and anything that weighs on the growth curve of the frontier labs could obviously dent investor confidence in these names. On the other hand, while token optimization concerns mount, we wonder if investor interest in "safer" corners of the tech space "not in the token path" (including software) might even ramp.

There is an angle to this optimization theme that in our view could be a fundamental positive – could software firms gain by playing an independent "token optimizer" role, akin to Palantir's launch of its new Evolve product to help customers minimize token consumption? In general, software firms don't own 1P AI models and can therefore be truly model-agnostic, similar to the advantage secured by the likes of Snowflake and MongoDB that were "multi-cloud" while AWS and Microsoft were pushing their native databases. Indeed, Snowflake and others have already begun pitching this advantage (the CEO of Snowflake has said *"cost is an issue, you can have sub-agents work with smaller models, and if there's an open-source model that's perfectly great at some job that also happens to be hosted by Snowflake, we can use one of those models, you don't always need a fancy LLM").* We expect almost every software firm to begin playing this card more vocally going forward.

In the looming battle with the frontier AI labs that are targeting the enterprise application software market with 1P agents built on their suite of proprietary models, being multi-model and utilizing performant open-source models or SLMs could also give the application software firms a cost advantage as well as reducing the pressure on gross margins, especially for non-coding use cases.

Nvidia and Other Hardware Firms

While it is clear that token budgets among enterprises are certainly not unlimited, we remain bullish on the infrastructure and semis demand picture as the bulk of capex is being ultimately driven by the frontier model providers and we have yet to see the wave of Blackwell trained (and inferenced) models – this should present much better token economics. Additionally, there are emerging demand frontiers related to the development of models able to interact with the world in multiple modalities beyond text and images (e.g. audio, video, physical AI data streams, etc..) which we think continues to add upside bias for infrastructure demand. On top of this, we are seeing more open models, down-tiering of models, and demand for specialized infrastructure capable of meeting demand for certain portions of the market. One visible example of this is the demand for real time tokens in applications like coding where speed is the main consideration rather than the ability to serve multiple users at once. We wrote about this extensively in our recent CBRS initiation (Cerebras Systems, Inc. "A Fast Inference Pure Play With Backlog Building;...").

Outside of the chip front, most server, storage and networking hardware firms have limited exposure to the hyperscalers and are more tethered to the pace of on-premise hardware investments. Could the token optimization trend motivate more enterprises to
scale their on-premise investments as a cost control measure? Frankly, we do not pick up many anecdotes to support this view, although we noted above an anecdote of a large global bank deploying Alibaba 's Qwen models on-premise (although this appeared to stem from security concerns more so than a desire to cut infrastructure costs). Broadcom guided to a near-term acceleration in VMware growth due to the impact of AI agents on demand for VMware cores, but we remain unclear about this on-premise AI pull-through and are not hearing of this from other hardware firms or enterprise customers.

***UBS Evidence Lab** is a sell-side team of experts that work across numerous specialized labs creating insight-ready datasets. The experts turn data into evidence by applying a combination of tools and techniques to harvest, cleanse, and connect billions of data items each month. Since 2014, UBS Research analysts have utilized the expertise of UBS Evidence Lab for insight-ready datasets on companies, sectors, and themes, resulting in the production of thousands of differentiated UBS Research reports. UBS Evidence Lab does not provide investment recommendations or advice but provides insight-ready datasets for further analysis by UBS Research and by clients. All published UBS Evidence Lab content is available via UBS Neo. The amount and type of content available may vary. Please contact your UBS sales representative if you wish to discuss access.*

US AI Business Survey (>Access Dataset)

This report leverages the following UBS Evidence Lab asset: US AI Business Survey. This survey of IT executives/ data engineers involved in making decisions related to the evaluation and use of Generative Artificial Intelligence (AI) provides a view on companies' expected adoption and usage of AI.

Valuation Method and Risk Statement

Investing in the technology sector entails above-average risk given low sales visibility, rapid pace of innovation and technological change, intense competition, frequent M&A, and low barriers to entry in many markets.