



José Mazoy, PhD
Global Chief Investment Officer at SAM

Tokenomics: when the unit of AI becomes a market variable

Why the economics of tokens now matter for market returns, capital intensity and portfolio allocation

Artificial intelligence is no longer just a software story. As models move from answering questions to executing tasks, the token has become the basic unit linking usage, revenues, margins, infrastructure demand and capital allocation. For investors, tokenomics is becoming one of the clearest lenses through which to understand where value is being created — and where it may be competed away.

For much of the AI cycle, the market has focused on a simple question: which company has the best model? That remains important, but it is no longer sufficient. The more relevant investment question is becoming: who controls the scarce units that turn AI usage into revenue, margins and return on capital?

That unit is the token

A token is the basic unit of text an AI model reads or writes. It is not a financial token, nor a crypto asset. It is a unit of computation and language — roughly three to four characters, or about three-quarters of a word. Every prompt, every answer, every tool call, every coding task and every autonomous action is converted into input and output tokens. In practical terms, although imperfect and not yet standardized, tokens are emerging as the de facto metric of the AI economy.

This matters because what is measured eventually becomes priced, scaled and allocated capital. For a software company, tokens are billable usage. For a semiconductor company, they represent compute demand. For a data-centre operator, they represent utilization. For a utility, they become electricity demand. For a portfolio manager, they are increasingly the bridge between digital adoption and physical capital intensity.

The rise of AI has therefore created a new analytical variable. We cannot only ask how many users an application has, or how many models are being deployed. We need to understand how many tokens are being consumed, what kind of tokens they are, what they cost to produce, who captures the margin, and which physical chokepoints constrain their supply.

From messages to agent-hours

The first important point is volume. Token consumption is rising extremely quickly. OpenAI disclosed in March 2026 that its APIs process more than 15 billion tokens per minute, equivalent to over 20 trillion tokens per day. Google reported processing more than 3.2 quadrillion tokens per month across its own products and platforms—over 100 trillion tokens per day—illustrating the extraordinary scale at which token consumption is growing. On the surface, this looks like another technology adoption curve. In reality, the more important story is not volume alone, but mix.

Until recently, most AI usage was essentially conversational. A user asked a question, the model answered, and the interaction ended. A typical message might consume a few hundred or a few thousand tokens. That was meaningful, but still manageable within the economics of a consumer internet service.

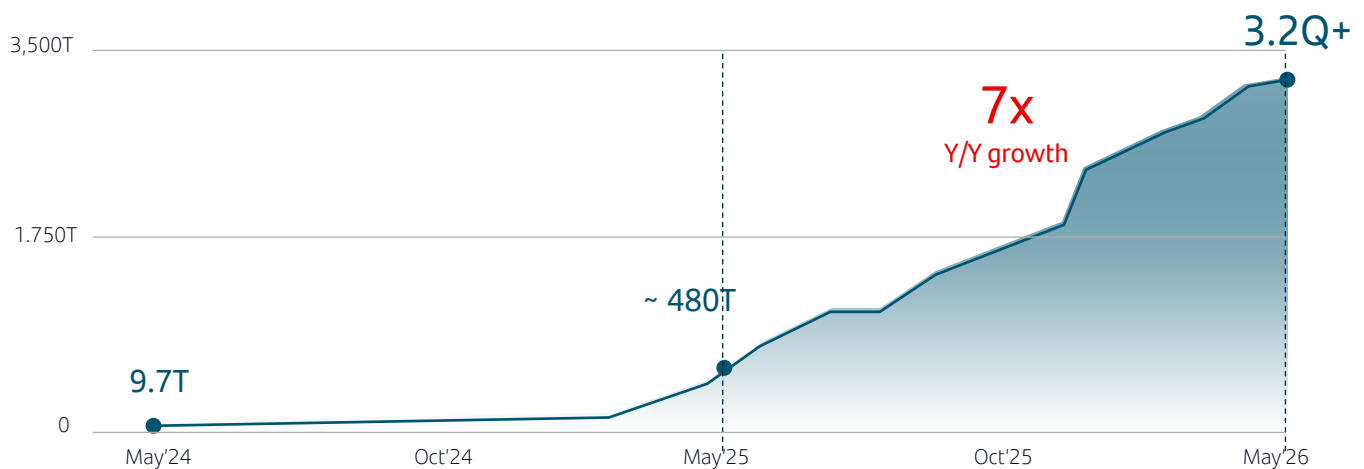


The shift toward reasoning and agentic AI changes this completely. In an agentic workflow, the model does not simply respond. It plans, searches, retrieves information, calls tools, writes code, checks its own output, revises and iterates. What used to be one question becomes a sequence of computational steps. The unit of consumption moves from the message to the workflow — or, more precisely, from the prompt to the agent-hour.

Monthly Tokens Processed

Across our surfaces

Source: blog.Google, I/O 2026: Welcome to the agentic Gemini era by Sundar Pichai CEO of Google and Alphabet



That distinction is critical. A research query that once consumed a relatively modest number of tokens can now require orders of magnitude more as it evolves into an agentic workflow. Coding agents can generate outputs several times longer than ordinary chat. AI video, multi-step research and enterprise automation require orders of magnitude more computation than simple text completion. The demand curve has steepened because the product itself has changed.

This is why tokenomics matters for market returns. A company exposed to AI is not automatically attractive because it has a chatbot, a model or a press release. The question is whether the AI use case generates high-value tokens, whether those tokens are priced at a premium, and whether the company can produce or consume them in a way that improves margins rather than diluting them.

The market is bifurcating

The second important point is pricing. The instinct in technology markets is to assume that prices always fall. In many areas of AI, that is correct. Commodity tokens are becoming cheaper. Lightweight models, open-weight alternatives and intense competition are pushing low-end inference pricing lower.

But this is not the whole story. At the frontier, prices are diverging rather than converging. Commodity-tier output prices have fallen toward cents per hundred thousand tokens, while frontier reasoning models charge hundreds of times more per output token than commodity inference models. The spread between low-end and frontier-premium output tokens has reached levels that would be unusual in most software markets.

This is the clearest signal that the AI market is splitting into two very different businesses.

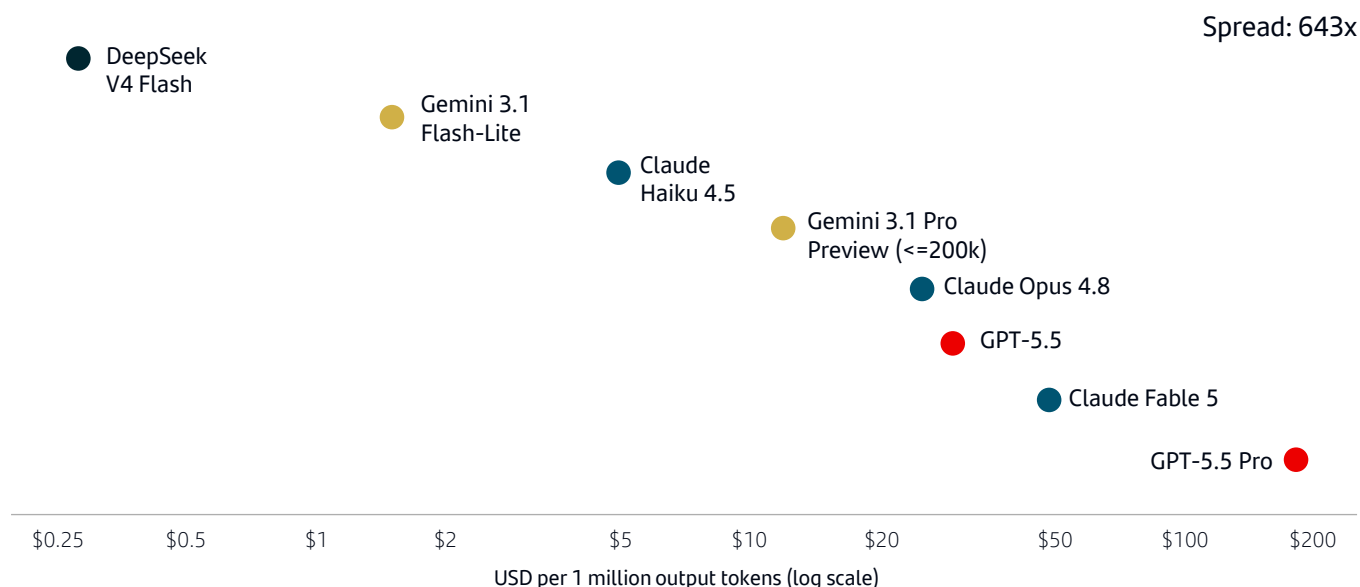
At the bottom of the stack, token production is becoming commoditized. Efficiency gains, smaller models and open-weight competition are forcing prices lower. This is good for adoption, but it is not necessarily good for margins. Businesses that only resell undifferentiated tokens may face the same problem that many infrastructure-light platforms faced in previous cycles: high usage growth, but limited pricing power.

At the top of the stack, frontier reasoning remains scarce. Premium models are expensive because demand exceeds available supply, especially for tasks where reliability, context length, reasoning quality and tool use matter. In these areas, customers are not simply buying text output. They are buying accuracy, productivity and time saved. That creates pricing power — at least while the supply chain remains constrained.

Token pricing is bifurcating, not converging

Official API output-token prices (USD per 1 million output tokens)

Source: Official API pricing pages of OpenAI, Anthropic, Google AI for Developers and DeepSeek. Data retrieved 21 July 2026. Note: Prices are official public API rates for 1 million output tokens. The x-axis is logarithmic. Gemini 3.1 Pro Preview is shown at the standard rate for prompts $\leq 200k$ tokens.



The investment conclusion is straightforward: the thesis is not “AI volume.” The thesis is where in the token value chain the economics accrue. The most attractive businesses are likely to be those that either control scarce token supply, own the infrastructure required to produce it, or apply tokens to high-value workflows where the customer return on investment is clear.

The physical economy behind the digital economy

The third point is capital intensity. Previous software cycles scaled with relatively low marginal cost once the platform was built. AI is different. Every token has a physical cost. It requires GPU cycles, high-bandwidth memory, advanced packaging, data-centre space, cooling and electricity. The more capable the model and the more agentic the workflow, the more visible this physical footprint becomes.

This makes AI a hybrid cycle: part software, part semiconductor, part infrastructure, part energy. That is why the market reaction has been so concentrated. Value has not only accrued to companies with the most visible applications. It has accrued to the providers of scarce inputs: GPUs, advanced packaging, HBM (High-Bandwidth Memory), leading-edge foundry capacity, power equipment and data-centre infrastructure.

The chokepoints are well known, but their market relevance has increased. Advanced packaging capacity, particularly CoWoS (Chip-on-Wafer-on-Substrate), remains tight. HBM supply is constrained. Grid availability is becoming a binding factor for data-centre construction. Power and permitting delays are no longer local operational issues; they are part of the AI capital-allocation equation.

This explains the scale of hyperscaler capex. Microsoft, Alphabet, Meta and Amazon have committed hundreds of billions of dollars to infrastructure, with a substantial majority tied to AI. This is not discretionary technology spending in the old sense. It is an attempt to secure the physical capacity required to produce future tokens.

For markets, that creates both opportunity and risk. The opportunity is clear: companies providing scarce infrastructure can capture strong revenues and margins while bottlenecks persist. The risk is equally clear: investors will increasingly ask whether this capex generates adequate returns. If token demand continues to compound, the spending can be justified. If application-layer revenue fails to scale, or if pricing falls faster than utilization rises, the same capex will look more vulnerable.

Why tokenomics changes capital allocation

Tokenomics matters because it forces investors to connect three variables that were often analysed separately: demand, cost and capital intensity.

On the demand side, the relevant question is no longer whether AI usage is growing. It is what kind of usage is growing. A million low-value tokens generated by a commodity chatbot do not have the same economic significance as a million tokens used in a high-value legal, financial, coding or industrial workflow. Token volume needs to be adjusted for quality, pricing and customer willingness to pay.

On the cost side, the key question is whether efficiency gains accrue to customers, model providers or infrastructure owners. Lower cost per token can expand the market, but it can also compress margins if the benefit is passed through too quickly. Conversely, when frontier capability improves faster than prices decline, providers may retain or even increase pricing power.

On the capital side, the question is whether the next dollar of AI investment produces more revenue, more margin, or simply more depreciation. The AI cycle is becoming increasingly asset-heavy. This raises the hurdle for capital allocation. It also broadens the opportunity set beyond the obvious technology leaders.

The beneficiaries are not limited to model providers. Data-centre operators, electrical equipment companies, grid infrastructure, cooling systems, power generators, semiconductor equipment, foundries and memory suppliers are all part of the token supply chain. In some cases, the most durable economics may sit further upstream than investors initially expected.

This has important implications for portfolio construction. AI exposure should not be defined only by ownership of the largest technology companies. It should include a broader assessment of where physical constraints, pricing power and capital discipline intersect. The digital economy is becoming more dependent on the physical economy, not less.

Market returns: from narrative to earnings

The recent earnings cycle has helped close the loop. Token demand has translated into physical constraints. Physical constraints have supported pricing power. Pricing power has justified capex commitments. Capex commitments have driven revenues for infrastructure providers. And those revenues have become a central driver of market returns.

This is why AI has remained so important for equity markets. It is not simply a story about future productivity. It is already showing up in revenues, margins, order books and capex guidance. Companies positioned as enablers of AI infrastructure have been rewarded where spending has translated into current revenue. Companies spending aggressively without a clear path to monetization have faced more scrutiny.

The market is therefore becoming more selective. In the early phase of a structural theme, investors often reward exposure. In the next phase, they reward conversion. The question moves from “Who is exposed to AI?” to “Who converts AI exposure into cash flow after capex?”

This creates dispersion. Infrastructure providers with tight capacity and high utilization can continue to benefit. Model providers with frontier capabilities may retain pricing power. Application companies with clear productivity benefits can justify token consumption. But companies with vague AI strategies, weak monetization or rising compute costs may find the market less forgiving.

The same logic applies across asset classes. In credit, investors will need to distinguish between capex that strengthens competitive position and capex that weakens free cash flow. In listed infrastructure and utilities, AI-related demand can improve growth visibility, but regulatory frameworks and financing costs still matter. In private markets, the opportunity set may expand across data centres, power, cooling, grid connection and digital infrastructure, but valuation discipline will be essential.

What can break the loop

The first is capex digestion. Investor patience is not unlimited. If application-layer revenues do not scale fast enough, the market will begin to question whether infrastructure spending has moved ahead of monetization. The first signs of this are already visible when companies raise capex guidance and receive a negative share-price reaction.

The second is chokepoint relief. Scarcity supports pricing power today, but scarcity is being attacked by investment. Advanced packaging capacity is expanding. HBM supply is increasing. New accelerator platforms aim to reduce cost per token. As bottlenecks ease, some of today's excess returns may normalize.

The third is open-weight competition. Lower-cost models are pushing commodity token prices under structural downward pressure. This is positive for adoption and productivity, but it reduces the ability of weaker model providers to maintain margins. The impact may be most visible at the application layer, where customers can switch between models more easily than they can switch power grids or foundry capacity.

The fourth is power. Chips may not remain the binding constraint. Electricity, grid connection and permitting could become the limiting factors for data-centre growth. This shifts the investment debate from semiconductors alone to energy systems and infrastructure.

The fifth is concentration. A small number of hyperscalers account for a large share of AI capex. That creates operating leverage for suppliers, but also customer concentration risk. Any meaningful revision to hyperscaler spending plans would have implications across the supply chain.

What we are monitoring

The practical response is not to abandon the AI theme, neither to embrace it indiscriminately. It is to monitor it with the right variables.

The first variable is token volume, but volume must be separated by use case. Reasoning, coding, research and agentic workflows matter more than simple chat growth.

The second is token price. Commodity tokens and frontier reasoning tokens are moving in different directions. The spread between them is a key indicator of where pricing power sits.

The third is cost per token. Efficiency gains are positive, but the distribution of those gains matters. Investors should ask whether lower costs are retained as margin, passed to customers or reinvested into even larger workloads.

The fourth is infrastructure capacity. CoWoS (Chip-on-Wafer-on-Substrate), HBM (High-Bandwidth Memory), power availability and data-centre construction timelines are no longer technical details. They are market variables.

The fifth is capex productivity. The relevant question is whether AI spending converts into revenue growth, margin expansion and durable competitive advantage.

The sixth is breadth. A healthy market should see benefits move beyond a narrow group of infrastructure winners into productivity gains across sectors. If returns remain too concentrated, portfolio risk increases.

Conclusion

Tokenomics is not a new acronym for the AI debate. It is the accounting system of the AI economy.

It connects model usage to infrastructure demand, infrastructure demand to capex, capex to revenues, and revenues to market returns. It also helps explain why the AI cycle has become so capital intensive. The marginal unit of AI is not free. It consumes chips, memory, packaging, electricity and balance-sheet capacity.

This does not weaken the structural AI case. In many ways, it strengthens it. The move from chat to agents suggests that AI is becoming more embedded in real workflows, not less. But it does change the investment discipline required. Exposure is no longer enough. Investors need to understand where the token is produced, who pays for it, who captures the margin, and whether the return on invested capital is sustainable.

Our central view remains that AI will continue to be one of the defining structural drivers of markets. However, the next phase is likely to be more selective than the first. The winners will not simply be those with the best narrative. They will be those that convert token demand into durable cash flow, while managing the capital intensity required to serve it.

The decisive variable is therefore not the number of tokens alone. It is return per token — and where that return accrues across the value chain.

Sources

OpenAI Help Center and API pricing
OpenAI company announcements
Google I/O 2026
Alphabet Q3 2025 and Q1 2026 earnings materials
Anthropic API pricing and company announcements
NVIDIA FY26 Q4 earnings release
Microsoft FY26 Q3 earnings materials
Amazon Q1 2026 earnings materials
Meta Q1 2026 earnings release
TSMC Q1 2026 earnings materials
DeepSeek API pricing
xAI release notes
OpenRouter/a16z State of AI 2025: An Empirical 100 Trillion Token Study
SAM calculations

Data note: Daily token-throughput figures are derived from disclosed monthly or per-minute figures where applicable. Public API prices refer to standard listed output-token prices per 1M tokens and exclude cached-token discounts, batch pricing, committed-use discounts, regional uplifts and private enterprise terms. Reasoning-share and coding-workload figures come from OpenRouter/a16z's primary dataset, not from model-provider official disclosures.

Important Legal Notice

This report has been prepared by Santander Asset Management (hereinafter “SAM”). SAM is the functional name of the asset management business conducted by the legal entity SAM Investment Holdings S.L. and its branches, subsidiaries and representative offices.

This document contains economic forecasts and information gathered from several sources. The information contained in this document may have also been gathered from third parties. All these sources are believed to be reliable, although the accuracy, completeness or update of this information is not guaranteed, either implicitly or explicitly, and is subject to change without notice. Any opinions included in this document may not be considered as irrefutable and could differ or be, in any way, inconsistent or contrary to opinions expressed, either verbally or in writing, advices, or investment decisions taken by other areas of SAM.

This report is not intended to be and should not be construed in relation to a specific investment objective. This report is published solely for informational purposes. This report does not constitute an investment advice, an offer or solicitation to purchase or sell assets, services, financial contracts or other type of contracts, or other investment products of any type (collectively, the “Financial Assets”), and should not be relied upon as the sole basis for evaluating or assessing Financial Assets. Likewise, the distribution of this report to a client, or to a third party, should not be regarded as a provision or an offer of investment advisory services.

SAM makes no warranty in connection with any market forecasts or opinions, or with the Financial Assets mentioned in this report, including with regard to their current or future performance. The past or present performance of any markets or Financial Assets may not be an indicator of such markets or Financial Assets future performance. The Financial Assets described in this report may not be eligible for sale or distribution in certain jurisdictions or to certain categories or types of investors.

Except as otherwise expressly provided for in the legal documents of a specific Financial Assets, the Investment Products are not, and will not be, insured or guaranteed by any governmental entity, including the Federal Deposit Insurance Corporation. They are not an obligation of, or guaranteed by, Santander, and may be subject to investment risks including, but not limited to, market and currency exchange risks, credit risk, issuer and counterparty risk, liquidity risk, and possible loss of the principal invested. In connection with the Financial Advisors, investors are recommended to consult their financial, legal, tax and other advisers as such investors deem necessary to determine whether the Financial Assets are suitable based on such investors particular circumstances and financial situation. Santander, their respective directors, officers, attorneys, employees or agents assume no liability of any type for any loss or damage relating to or arising out of the use or reliance of all or any part of this report.

Past performance does not predict future returns. The returns may increase or decrease as a result of currency fluctuations relative to the respective investors’ domestic currency. Any reference to taxation should be understood as depending on the personal circumstances of each investor and which may change in the future. Costs incurred for purchasing, holding or selling Financial Assets may reduce returns and are not reflected in this report.

This report may not be reproduced in whole or in part, or further distributed, published or referred to in any manner whatsoever to any person, nor may the information or opinions contained therein be referred to without, in each case, the prior written consent of SAM.

Any third-party material (including logos, and trademarks), whether literary (articles/ studies/ reports, etc. or excerpts thereof) or artistic (photos/graphs/drawings, etc.), included in this report is registered in the name of its respective owner and only reproduced in accordance with honest industry and commercial practices.