

The Future of Tokens: Demand at Machine Speed

Shay O'Kelly · August 2026

The macro essay argued that human labor stops being the binding constraint on output once capital can do cognitive work. This one is about what that does to the meter. Every argument in this collection eventually reduces to a single number, which is how many tokens the world is going to need, and nearly every forecast of that number is built on a picture of a person typing. That picture is already obsolete. Human-scale usage, one prompt and one answer, was never the market. It was the demo.

Start with what is actually being counted, because the unit is easy to wave past. A token is a piece of a word. Models read and write in these pieces, and every one of them is metered and billed, so a token is both the atom of machine thinking and the line item on the invoice. The Declaration of Independence is about 1,600 of them¹. A long conversation with a model might run a few thousand. Those numbers are why token counts sound small and harmless, and they are also why the market keeps underestimating them, because the workloads that matter no longer look anything like a document or a conversation.

Agentic work runs in loops. Plan, act, observe, verify, retry. A model asked a question answers once. An agent given a job writes a plan, calls tools, reads what came back, checks its own work, finds the error, and goes again, sometimes for hours, and every pass through that loop is inference. Gartner puts the gap at five to thirty times the tokens of a comparable chat request, and Stanford researchers measuring coding agents against simple code chat found it running to a thousand times². Priced out, one completed agentic task runs between ten cents and a dollar against a fraction of a cent for the chat exchange it replaces². To anyone who does not work with these systems every day, asking a model a question and handing an agent a job can sound like the same request. It is not the same work. The agent is doing far more, and it is consuming orders of magnitude more to do it.

Then there is the clock, and this is the part I think is most underpriced. A human-driven system consumes tokens at the speed of human attention, because someone has to read the output, decide, and prompt again. That caps demand at waking hours and the pace of one person's judgment. The third essay argued that human cognitive labor stops being the constraint on what the economy can produce. Read that claim off the meter instead of the ledger and it says something sharper. Take the person out of the middle of the loop and the ceiling on consumption leaves with them. An agent that verifies its own work does not stop at five, does not sleep, and does not wait for a colleague to approve step four before it starts step five. Every hour the review step used to take becomes an hour of inference instead. The economy has spent its entire history metering cognitive work by the availability of the people doing it. That meter is coming off.

The second-order effect is larger than the first. As agents begin transacting with other companies' agents, scoping work, negotiating terms, dispatching jobs, confirming completion, every commercial interaction that used to be a phone call or a flat file becomes a multi-turn exchange between two reasoning systems,

each burning inference on both sides of the trade. This is not speculative. The plumbing is being laid right now. The protocol that lets agents call tools has been downloaded 97 million times and adopted by every major lab, the protocol that lets agents discover and negotiate with each other has more than 150 participating organizations and sits at the Linux Foundation, and there is now a payments standard whose entire purpose is proving that a human authorized what an agent just bought³. There is also precedent for how large machine-initiated volume gets once rails like that exist. Visa has found that only about a tenth of the transactions crossing its network are initiated by a person touching something⁴. The rest is already software talking to software. Commerce made that transition once, quietly, on rails built for something else. It is about to make it again, except this time both ends of the conversation are thinking.

And the loop feeds itself. Agents produce outcomes, and outcomes are training data of a kind the internet never contained, a record of what was actually tried and what actually worked. The more work the economy routes through agents, the more valuable the next generation of models becomes and the more compute it takes to build. Inference demand recycles into training demand. Notice how completely that inverts the old arrangement. For fifty years we trained people to use software. Now the software is the thing being trained, and every hour it runs is another hour of instruction. That turns the usage itself into an asset rather than a cost. A machine wears out as you run it. This one gets better, because every job an agent finishes is another example the next model learns from, and a better model gets handed more work, which produces more examples again. It is also why the familiar technology curve does not apply. A maturing product normally needs less compute per unit of output every year, and this one takes every efficiency it earns and spends it on a bigger job.

The obvious objection is price. Inference costs have fallen roughly a thousandfold in three years⁵, so a skeptic can argue that all this volume arrives at a collapsing unit price and the revenue never shows up. The macro essay named why that argument fails, because when the price of a valuable input collapses, demand does not politely stay where it was. There is a sharper version here. Falling token prices are not a headwind to this thesis, they are the mechanism of it. Agentic loops are only economic because tokens got cheap. At 2023 prices, spending a thousand steps of verification to finish one task would have been absurd, and at today's prices it is a rounding error against the salary of whoever used to do that task. Every price cut converts another category of work from too expensive to automate into obviously worth automating, and the volume that walks through the door is larger than the price decline that opened it.

And the largest source of that volume is not the work agents take from people. It is the work agents make that nobody could afford to do before. Replacing existing work comes first, because it is the easiest kind of spending to approve. A company already knows what that work costs it today, so the savings are simple to point at. The volume that follows comes from everything that was never done at all. The analysis a small business could never buy. The second opinion nobody had time to seek. The thousand molecules never screened because screening them cost more than the answer was worth. Every one of those is a job that did not exist as demand last year and is a metered workload this year. When something scarce becomes abundant, people do not buy the same amount of it at a lower price. They start buying things they could never afford to buy at all.

Put all of that together and you get a decoupling I think is the most underpriced fact in this collection. Token consumption stops scaling with the number of humans typing and starts scaling with the volume of economic activity itself. Those are different curves with different ceilings. The first is bounded by population and attention and grows a few percent a year at best. The second is bounded by GDP. Which means compute, and the power underneath it, stops being a line item in an information technology budget and becomes a direct input cost to output, closer to electricity or freight than to software licensing. Consensus is still modeling the first curve. Google was already processing more than three quadrillion tokens a month by last spring, seven times the prior year, before agents were anything more than a pilot in most companies⁶, and the forecasts that try to price the agentic era have consumption multiplying more than twenty times again by 2030⁷.

So this is not a software cycle. Software cycles are gated by budgets and seat counts, they saturate, and the marginal copy is free. This is an infrastructure buildout with a demand curve tied to the whole economy, and every unit of it is paid for in silicon, electricity, and cooling. It is capital constrained at every layer, generation, interconnection, silicon, and the research that keeps the loop compounding. Which leaves one question, and it is the one the final essay has to answer. If the demand is this visible and the constraints this physical, why is almost none of it in the price?

SOURCES

1. Token counts for reference documents (Declaration of Independence ~1,627 tokens): Ribbit Capital, "Token Letter," June 2025; OpenAI tokenizer documentation.
2. Agentic token multipliers (Gartner: 5-30x tokens per task vs. chatbots, March 2026; Stanford coding-agent measurements up to 1,000x vs. simple code chat; \$0.10-\$1.00 per completed agentic task vs. fractions of a cent per chat call; 50,000 tokens per code review, 200,000 per browser procurement task): Gartner via Spheron; AgentMarketCap; Zylos Research.
3. Agent protocol adoption (Model Context Protocol ~97M downloads, adopted by Anthropic, OpenAI, Google, and Microsoft; Agent2Agent protocol with 150+ participating organizations, donated to the Linux Foundation June 2025, v1.0 with signed agent cards in early 2026; Agent Payments Protocol for cryptographically verifiable purchase authorization): Linux Foundation; Google; Cipher Projects; Digital Applied.
4. Visa analysis finding roughly 10% of network transaction volume is human-initiated: Visa, "Making Sense of Stablecoins," April 2024.
5. Inference cost decline of roughly 1,000x over three years: Ribbit Capital, "Token Letter," June 2025; Epoch AI.
6. Google processing 3.2 quadrillion tokens per month, seven times year over year: Google I/O 2026.
7. Forecast of token consumption multiplying ~24x to ~120 quadrillion tokens per month between 2026 and 2030: Zylos Research; industry forecasts compiled by AgentMarketCap.

No Offer or Solicitation. The only purpose of this site is to provide general background information on RAXIA. This is not an offer to sell or the solicitation of any offer to buy any securities. Any such offer would be made only to qualified investors after distribution and return of the Private Placement Memorandum and governing agreement for the applicable Fund (the "Offering Materials"). Nothing on

this site is investment advice or a recommendation of any security. The author may hold positions in companies discussed. Past performance is not indicative of future results.