10-16-25

WSJ Print Edition

# AI Growth Comes at Staggering Price Tag

BY STEVEN ROSENBUSH



AI companies are losing money at an epic pace, and the reasons go deeper than profligacy. The economics of artificial intelligence have turned sharply against them, at least for now, and for reasons that weren't widely anticipated.

There are increasingly urgent concerns about massive capital spending, soaring valuations, high debt levels and the circular nature of AI firms pouring money into other AI firms. -Note

And even the most likely eventual winners in AI are losing billions of dollars now.

It is hard to predict how this will play out in the financial markets, but here is a clue. Demand for AI, measured in units of data processed, is soaring. The entire AI bet may turn on how far and fast it ramps up. -Note

A key mistake occurred a year ago when AI leaders focused solely on AI cost curves for a unit of computing and took their eye off the number of units needed to get the work done, said Heath Terry, the global sector lead for technology and communications research at Citi. AI computing costs had been declining around 90% every seven months, a dynamic akin to Moore's Law for microchips, giving AI companies reason to believe their price-performance ratio will improve.

Then usage, which everybody expected to increase as the price fell, went absolutely parabolic.

"This time last year, we were processing 9.7 trillion tokens a month across our products and APIs [application programming interfaces]," **Alphabet's** Google said in a May blog post, referring to units of AI usage. "Now, we're processing over 480 trillion— 50 times more."

Google was just getting warmed up. "Since then we have doubled that number, now processing over 980 trillion monthly tokens, a remarkable increase," the tech giant said in July during its second-quarter earnings call.

This month, Google said the figure had reached 1.3 quadrillion.

(News Corp, owner of The Wall Street Journal, has a commercial agreement to supply content on Google platforms.)

The demand for tokens has been driven in part by the fight against the "hallucinations," or incorrect or nonsensical outputs stated as if they are true, notoriously produced by generative AI models. Current cutting- edge models are designed to counter the problem often with the help of a kind of reinforcement learning in which they answer the same question multiple times internally be-

fore issuing a response. They also might query a mixture of experts, meaning specialized models. That has helped AI produce better answers—and exacerbated AI's resource needs.

So the unit price of a token has fallen, but the overall consumption of tokens has exploded. - Note

That demand will only continue to soar. **OpenAI 's** new Sora short-form AI video app powered by the latest version of its Sora video generator is a hit. (News Corp has a contentlicensing partnership with OpenAI.) The company is looking to ramp up efforts in AI commerce as well. And Anthropic's new large language model, Claude Sonnet 4.5, can code for 30 hours straight.

Such longer-thinking models are an emerging form of AI that will put the industry's ballooning data-center capacity to use.

AI startups are losing tens of billions of dollars more than they would have partly because of this dynamic. Startups and venture-capital investors so far are underwriting such losses, much as Uber Technologies once subsidized ride prices for users to build a market, Terry said.

Three forces are on track to come together and push the economics of AI into the black, in Terry's view.

First, models will continue to improve. While the evolution of the transformer architecture underlying generative AI such as large language models may have slowed down, developers are creating new forms of hybrid models such as neurosymbolic AI. Some AI developers say they are closer to the point where AI systems are beginning to assist in their own improvement, a concept that Terry calls AI squared.

"Over the last few months, we've begun to see glimpses of our AI systems improving themselves," **Meta Platforms** Chief Executive Mark Zuckerberg said in July on the socialmedia giant's second-quarter earnings call.

All of that should make AI models more efficient, which could have a beneficial effect on the economics of AI.

Second, data centers under construction around the world will soon begin operations, increasing the supply of computing power and energy and driving down the cost of training and using AI. **Microsoft** last month said it is in the final stages of construction on a $3.3 billion data center in Wisconsin and announced plans for a second, $4 billion facility in the area.

And third, increasing demand for tokens could finally become an economic benefit as models improve and the supply of computing infrastructure expands.

Another massive increase in demand for AI is likely in the coming months, just as those other forces take hold, according to Terry. That is because the most promising corporate AI trials are likely to move into broader deployment.

AI computing will increase "by a billion X," **Nvidia** CEO Jensen Huang said last month on the " BG2" podcast. "That's the part that most people haven't completely internalized," Huang said. "…This is the industrial revolution."

Should all of this go exactly according to plan, the AI boom will keep booming. Citi forecasts that AI revenue will grow by nearly 80% annually over the next five years, reaching $780 billion in 2030, compared with $43 billion this year.

But even then, returns in AI may be very highly concentrated, eventually leaving many contenders out in the cold.

Not

"In venture capital, 6% of investments result in 60% of returns," venture capitalist Vinod Khosla told me, alluding to a 2015 blog post from "a16z" that cited three decades worth of data from investor Horsley Bridge. "In AI, I think it will be half that percentage resulting in more than 60% of returns." *Steven Rosenbush writes for WSJ Pro's CIO Journal.*





 