Insight

INSS
המכון למחקרי ביטחון לאומי
THE INSTITUTE FOR NATIONAL SECURITY STUDIES
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

The Token Economy: Understanding the Infrastructure Behind the AI Revolution

Hadas Lorber and Ido Ben-David[1] | No. 2186 | August 18, 2026

Artificial intelligence is not only altering the processing capabilities of computers, but also creating a new economy based on the token as the central unit of measurement for model usage, the demand for computing power, and the economic cost of operating them. The token reflects the scope of inference work—the processing performed by the model in real time—and therefore serves as a direct metric for the consumption of computing resources, including graphics processing units (GPUs), data centers, and energy. In this sense, the token is not what drives investment in AI infrastructure but rather the metric through which demand for computing power—the key strategic resource in the age of artificial intelligence—can be understood. The article presents the concept of the "token economy" as a new analytical framework for understanding the development of the AI market and demonstrates how it explains processes such as global investment in data centers, the rise of NeoCloud companies, the race to establish national AI infrastructure, and the shaping of digital sovereignty based on control over computing power. Alongside this, it emphasizes that the token economy is also changing how organizations manage artificial intelligence systems, requiring a shift from measuring model performance alone to efficiently managing token consumption, inference costs, and the business value generated from each interaction. In light of this, the article argues that understanding the token economy is key to understanding the business model of the AI industry and the global competition over computing infrastructure and technological sovereignty.

The Token Economy: Much More Than a Unit of Pricing

Since the launch of ChatGPT in late 2022, the public debate on artificial intelligence has primarily focused on AI model capabilities: the quality of their writing, their reasoning abilities, and their capacity to generate code, images, and videos. However, behind the technological progress, a deeper economic revolution is developing, relying on a concept that most users are barely aware exists—the token.

Technically, a token is the basic unit of information through which language models process information. It is neither identical to a word nor to an individual character; a token can be an entire word, part of a word, a punctuation mark, or a sequence of characters, depending on the tokenization method

---

[1] **Ido Ben David** – Tech Partners Manager at Nebius, responsible for the company's engagement with the Israel Innovation Authority (IIA) on the development of Israel's national supercomputer.

used by the model. Non-textual information—images, audio clips, documents, or video—is also ultimately converted into representations that can be processed using similar computing units.

These differences are not merely technical. Different languages require different numbers of tokens to represent the same idea. In Hebrew, for example, the number of tokens required remains relatively high compared with English. This is due both to the language's morphology and to the fact that most models have been trained primarily on English-language texts, while Hebrew characters are represented using a larger number of bytes. The gaps between languages are gradually narrowing as tokenization technologies improve.

However, the token is not merely a unit of processing. For model providers such as Google, Anthropic, and OpenAI, it has also become the central pricing unit for artificial intelligence services. Instead of charging by the number of questions asked, billing is generally based on the number of tokens the model receives (input tokens) and generates (output tokens). Behind every query often lies a much larger volume of information: system instructions, conversation history, documents retrieved from RAG systems (systems that integrate real-time information retrieval from external knowledge bases), outputs from external tools, and additional context streamed to the model. All of these are ultimately translated into tokens.

Yet, the true significance of the token does not lie specifically in its use as a pricing unit. Even if model providers chose to charge based on processing time, the number of questions, or energy consumption, the need for computing infrastructure would remain similar. The token's importance stems from the fact that it has become the best unit for measuring demand for real-time processing (inference). The more tokens that are processed, the more computing power is required; and the more computing power is required, the greater the need for data centers, GPUs, and energy. From this perspective, the token economy is not merely a pricing model. It is a new way of understanding how demand is generated for the most important strategic resource in the age of artificial intelligence: computing power.

From Tokens to Inference: Why Artificial Intelligence Is a Different Economy

To understand the token economy, it is important to distinguish between the economy of artificial intelligence and the traditional software model (Software as a Service – SaaS). For the past two decades, the software industry has relied on a simple assumption: the main cost lies in developing the product. Once the product has been developed, the cost of each additional user is relatively low. Therefore, most cloud services are priced through a fixed monthly subscription, even when actual usage varies from user to user.

Generative artificial intelligence operates differently. Even after the model has been trained, every interaction with the user requires real-time computations. Every question, document, image, or piece of code passed to the model triggers a chain of computational operations performed on GPUs, consuming memory, bandwidth, and electricity. In other words, unlike traditional software, the cost does not end with the development of the model but continues throughout its entire lifecycle, each time it is used.

This is why the concept of inference has become one of the most important economic variables in the artificial intelligence market. While the training phase is a one-time investment, the inference phase is an ongoing activity whose scope grows as the number of users and applications increases. In fact, as

artificial intelligence becomes an integral part of business processes, the bulk of the economic cost shifts from the model development phase to its operational phase.

The token is the metric through which this activity can be measured. The number of tokens processed reflects, approximately, the scope of inference work the model is required to perform. Therefore, for model providers, data center operators, and even investors, the token has become a key metric for understanding the future demand for computing power. Ostensibly, one might have expected technological progress to reduce the importance of computing infrastructure. Indeed, the cost of processing a token is steadily declining. New generations of graphics accelerators, improvements in model architecture, quantization techniques, smart routing mechanisms, and various optimization methods make it possible to perform the same task at a lower cost than just a year ago.

However, it is precisely this drop in price that is expected to increase overall demand for processing. This phenomenon is known in economics as the Jevons Paradox: when the use of a resource becomes more efficient and less expensive, its overall consumption tends to increase rather than decrease. This occurred in the past with coal, electricity, and bandwidth, and it may well happen with artificial intelligence.

Microsoft CEO Satya Nadella recently referenced this principle in the context of AI. According to Nadella, as the cost of running models drops, organizations will integrate them into a greater number of processes, expand their scope of usage, and generate even higher demand for computing power. In other words, the decrease in cost per token is not expected to reduce the need for infrastructure, but rather to increase it.

"Thinking" Models Are Changing the Cost Structure

Alongside the decrease in cost per token, another trend with profound economic significance is developing: the transition to reasoning models. Unlike traditional models, whose goal is to produce an answer as quickly as possible, these models perform a more complex inference process involving intermediate steps, self-checks, and sometimes even the generation of a chain of thought (CoT).

The result is a significant improvement in the quality of responses, particularly for complex tasks such as coding, research, document analysis, or decision-making. However, this improvement also entails an increase in the number of tokens processed and the amount of computing time required. As more organizations adopt such models, overall demand for inference is also expected to grow, even if the number of users remains similar.

At the same time, AI agent-based systems are expected to amplify this phenomenon even further. Unlike a human user, an AI agent may execute dozens or even hundreds of calls to a model in the course of performing a single task—searching for information, writing code, operating tools, verifying results, and formulating a response. Therefore, in the near future, the number of tokens consumed by a single "user" could grow significantly, even if the total user count does not change.

These trends have given rise to a new and expanding field—Token FinOps. Similar to Cloud FinOps, which aims to manage cloud costs intelligently, Token FinOps deals with the economic management of artificial intelligence usage, establishing a link between computing resource consumption and the business value generated from them.

Such management is not limited to measuring token counts. It includes selecting the right model for each task, shortening context windows, using caching mechanisms that enable the reuse of previously computed results, and routing simple tasks to smaller, less expensive models. In certain models, processing information retrieved from a cache is priced at one-tenth or even less of the cost of processing new information; thus, proper data flow design can significantly reduce costs. In other words, just as organizations learned over the past decade to manage their cloud consumption, in the coming years they will also need to learn how to manage their token economy.

From Tokens to Data Centers: How the Inference Economy Is Shaping the Infrastructure Market

Understanding the token economy helps explain why the market's primary focus today is not directed at the models themselves, but at the infrastructure that enables them to operate. If the token is the unit of measurement for the volume of inference work, then the growth in the number of processed tokens translates almost directly into a growing demand for computing power.

From this perspective, the global race to establish data centers is not the result of a passing fad or a specific pricing model, but of a structural shift in the economics of artificial intelligence. As more organizations integrate AI into their workflows, and as models become more complex and agent systems execute a growing number of calls to models, the demand for inference increases. This demand requires a continuous expansion of computing infrastructure, including GPUs, communication networks, cooling systems, and power supplies on scales never before required.

This trend is reflected in the unprecedented wave of investments currently characterizing the sector. OpenAI completed a capital raise this year at a valuation of hundreds of billions of dollars, while the Stargate project is expected to invest up to $500 billion in establishing AI infrastructure in the United States over the coming decade. At the same time, tech giants continue to accelerate the expansion of their data centers, investing tens of billions of additional dollars in purchasing GPUs, constructing new computing facilities, and securing energy supplies.

The meteoric rise in demand for NVIDIA processors also stems not merely from the company's success in developing advanced hardware, but from the fact that the GPU has become the central production resource of the artificial intelligence economy. Just as oil was the raw material of the industrial economy, computing power is gradually becoming the foundational resource of the new digital economy.

One of the most interesting developments in recent years has been the emergence of NeoCloud companies—cloud providers focused almost exclusively on AI infrastructure. Unlike traditional cloud services, which were designed to serve a broad variety of workloads, these companies build their operations around providing dedicated computing power for training and running models.

Companies such as Lambda, Crusoe, CoreWeave, and Nebius compete primarily on their ability to provide fast and efficient access to GPUs, high-speed networks, and software systems optimized for artificial intelligence workloads, rather than on storage capacity or traditional IT services. In this sense, they reflect the transition from a "general" cloud economy to a compute economy—one in which competitive advantage is measured first and foremost by the availability of computing power.

From Infrastructure to Compute Sovereignty

The implications of the infrastructure race extend beyond the world of computing. As data centers expand, so does demand for electricity, water for cooling, and suitable land on which to build new facilities.

It is no coincidence that big tech companies are currently investing in power plants, solar farms, small modular reactors (SMRs), and long-term energy supply agreements. Unlike the previous decade, when the bottleneck was primarily chip manufacturing, the main challenge of the coming years is expected to be the ability to provide data centers with a stable and readily available supply of energy.

This trend is also reshaping the global investment map. Countries and regions capable of providing land, electricity, regulatory connectivity, and appropriate infrastructure are becoming sought-after destinations for establishing data centers, while countries unable to provide these conditions risk finding themselves outside the artificial intelligence value chain.

This development is also transforming the concept of national security. For years, digital sovereignty was perceived primarily as a question of data control, cybersecurity, or regulation of digital platforms. Today, a broader perception is taking root, according to which access to computing power is also a strategic asset.

Countries understand that without stable access to GPUs, data centers, and advanced AI models, they may become dependent on the decisions of a small number of foreign companies and technologies. Recent developments surrounding the restriction of access to advanced models for national security reasons in the United States have demonstrated that access to AI capabilities themselves is beginning to serve as a policy tool, much like export controls on advanced chips.

For this reason, the United States, the European Union, the Gulf states, and other countries are currently investing vast sums in expanding their national computing capabilities. The objective is not merely to encourage innovation, but to ensure that countries have independent access to the infrastructure on which the economy will rely in the coming decades.

In other words, while the previous decade focused on data sovereignty, the coming decade is likely to center on compute sovereignty—a country's ability to ensure continuous, reliable, and secure access to AI resources. The token economy provides the lens through which to understand this shift: not because the token itself drives investment, but because it reflects the scope of demand for computing power, which is gradually becoming the most important strategic resource of the artificial intelligence era.

Implications for Israel

Israel is not immune to the shifts taking place in the artificial intelligence economy. For years, the country’s comparative advantage has rested on its human capital, strong academic system, and well-developed startup ecosystem. However, as artificial intelligence evolves from software into infrastructure, it becomes clear that these advantages, important as they may be, are no longer enough. Innovative countries must also secure continuous access to advanced computing power.

This understanding is reflected in Government Resolution 4255 from June 2026, which defines artificial intelligence as a strategic asset with implications for economic growth, national security, and social resilience. One of the resolution's key innovations is the recognition that the challenge is no longer

limited to encouraging research or increasing the number of startups, but also involves ensuring long-term access to advanced computing resources.

Accordingly, the government set a target to guarantee that, within five years, the Israeli economy has access to processing power equivalent to the performance of approximately 100,000 AI accelerators (GPUs), serving academia, the public sector, and industry. This is not merely a technological objective; it is a new strategic concept in which computing becomes part of the nation's core infrastructure—similar to electricity, water, or telecommunications.

The establishment of the national AI supercomputer (2025) is an initial expression of this concept. The project's importance stems not only from increasing the supply of computing power, but from laying the foundation for a national ecosystem that will enable academia, startups, industry, and the government to develop and deploy AI applications without relying entirely on access to resources outside of Israel.

However, limitations must also be acknowledged. Israel is not expected to compete with the United States, China, or the Gulf states in the scale of data center investments or the sheer number of GPUs. Its advantage will continue to rest primarily on its ability to develop innovative applications, algorithms, cybersecurity solutions, digital health, and defense technologies. Israel's strategic challenge, therefore, is not to become a global computing power, but to ensure that its researchers, companies, and state institutions have reliable and continuous access to advanced AI resources.

In this sense, compute sovereignty does not require complete ownership over the entire value chain. It requires, first and foremost, access security—the ability to ensure that, even during periods of geopolitical crisis, global shortages, or export restrictions, sufficient computing resources remain available to the Israeli economy.

Editors of the series: Anat Kurz, Rinat Harash, Eldad Shavit and Keri Rosenbluh