Crusoe

AI tokenomics in the agentic era



Agentic AI can use 10 to 100x more tokens per task than a chatbot

This shift is pushing teams to rethink how they evaluate AI infrastructure.

AI tokenomics provides a full view of what drives efficiency and value — from the physical foundation to the software stack. Picture it like an inference iceberg: the key to lowering token cost lies beneath what's comparable above the surface.

The inference iceberg

Above the waterline

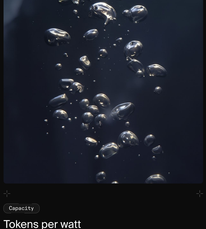
Originally, GPU per hour and FLOPS per dollar were used to compare infrastructure. These metrics were built for single turn chat workflows. Now, agents consume far more tokens to complete a task. To accurately evaluate your AI inference economics you have to dive deeper.

Cost per GPU per hour
Time reserved per dollar spent

FLOPS per dollar
Raw compute delivered per dollar spent

What is AI Tokenomics?
30s

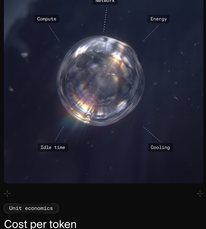
Below the surface: A more complete AI tokenomics view looks at four metrics.



Capacity

Tokens per watt

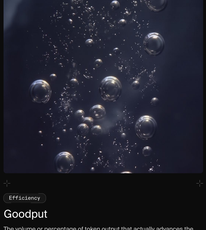
The amount of intelligence per unit of GPU power drawn. In an efficient AI factory, less power is lost to overhead, so more of your energy budget goes directly toward producing tokens.



Unit economics

Cost per token

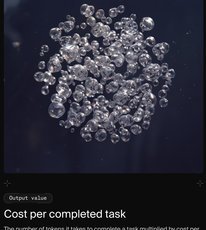
The true cost of a delivered token, not just the compute used to produce it. This tells you how much value you're receiving for every dollar spent, not just how much compute you rented.



Efficiency

Goodput

The volume or percentage of token output that actually advances the task toward completion, within acceptable latency. High throughput with low goodput still means wasted spend, no matter how cheap each token was priced.



Output value

Cost per completed task

The number of tokens it takes to complete a task multiplied by cost per token. The lower your cost per completed task, the more efficient your agent's token usage is.

Determining total output value

Two layers form the foundation of modern AI tokenomics for agentic systems: the physical foundation and the software stack.

The physical foundation

The physical layer determines the intelligence ceiling

The physical foundation determines sustained throughput, energy efficiency, and scale. Many teams don't own their AI factory, so they can't control how much intelligence every watt produces. Maximizing tokens per watt spans three key components.

- Accelerated computing
- Efficient design
- Vertical integration
- The AI factory — 30s

The software layer

The software layer maximizes total performance

The right models, optimizations, and solutions help teams get the most valuable output from agentic systems.

- Models
- Optimizations
- Solutions
- Choosing the right provider — 30s

The full picture

AI tokenomics can't be judged by any single metric

Together, tokens per watt, cost per token, cost per completed task, and goodput show the true value per dollar for inference and agentic workloads. From energy to infrastructure to software, Crusoe and NVIDIA are collaborating to build efficient and reliable AI factories.

Optimize your AI tokenomics with Crusoe

Stop leaving efficiency on the table.

Talk with our team about benchmarking your workloads, transforming your AI inference economics, and building a high-efficiency foundation for your models.

Contact us

Want to read this page offline?

Download PDF

2026 Crusoe, Inc. All rights reserved.