A SPECIAL SUNDAY ISSUE

The Cheaper Intelligence Gets, the More Infrastructure It May Need

Kimi K3 reopened the debate over model economics. The more important question is whether cheaper intelligence increases total compute consumption.

OTTO ANALYTICS · THE WEEKLY READ · NO. 05 · JULY 19, 2026

> *Kimi K3 may pressure model pricing. We think it strengthens the infrastructure thesis. The cost of completing a fixed task can fall while total compute consumption rises, because lower prices make more tasks worth doing and agents consume intelligence continuously. That outcome is not guaranteed. It is the question investors should underwrite.*

This week's news requires a special issue.

Moonshot AI introduced Kimi K3, bringing another increasingly capable model into the frontier conversation. The company says K3 has 2.8 trillion total parameters, a one-million-token context window and native vision capabilities. It uses a mixture-of-experts design, which means each token is routed through 16 specialized blocks out of 896 rather than activating the entire model at once.

That design improves efficiency. It does not make the infrastructure disappear. Moonshot recommends deployments with at least 64 accelerators, a reminder that an enormous model can be efficient relative to its size and still require an enormous physical system.

Axios framed the development as a split in the AI race: premium closed models on one side, cheaper and increasingly capable open-weight alternatives on the other. Moonshot says Kimi's full weights will be released by July 27. As of this Sunday, they are not yet public. Independent researchers cannot fully

inspect, modify or deploy those weights today, and several benchmark results remain company-reported or API-based.

The immediate debate is about which model companies win. We think that is the less durable question. Kimi is the catalyst for this issue, not proof of its conclusion.

The price of a token is not total compute

THE DISTINCTION UNDERNEATH THE DEBATE

A token is a small unit of information that a model reads or produces. When the price per token falls, an existing workload becomes cheaper. That is real efficiency. It is not the same thing as saying the economy will consume less compute.

The historical evidence is already striking. Stanford's AI Index found that the cost of querying a model at roughly GPT-3.5 quality fell from $20 per million tokens in November 2022 to seven cents by October 2024, a decline of more than 280 times. Yet the physical buildout accelerated over the same period.

Why? Lower prices do two things. They reduce the cost of work that already exists. More importantly, they make work possible that did not previously make economic sense.

A SIMPLE THOUGHT EXPERIMENT

Efficiency can rise while demand rises faster

Today: 1 million tasks at 100 compute units each equals 100 million units.

Later: efficiency improves tenfold, reducing each task to 10 units.

Lower costs expand usage to 20 million tasks.

Total consumption becomes 200 million units.

Compute per task falls 90 percent while aggregate compute demand doubles.

Economists call this pattern the Jevons paradox: making a resource more efficient can increase total consumption if the lower effective price expands usage by even more. The label is less important than the arithmetic.

What cheaper intelligence makes possible

FROM OCCASIONAL PROMPTS TO CONTINUOUS WORK

A chatbot waits for a person to ask a question. An agent can read a repository, write code, run a test, inspect the failure, revise the code and try again. A customer-service system can monitor thousands of conversations. A multimodal system can process text, images and video together. An autonomous machine can keep sensing and deciding as long as it operates.

These are not one-prompt workloads. They are loops. Gartner estimates that agentic models can use five to thirty times more tokens per task than a standard chatbot. Anthropic's own usage research finds that more valuable and more autonomous outputs tend to consume more tokens. Building an application uses more compute than producing a simple explanation because the model reasons longer, takes more turns and creates a more complex artifact.

Efficiency is often reinvested rather than pocketed. Give a developer a cheaper model and the savings can fund a longer context window, more reasoning, several independent attempts and a verification pass. The cost of one attempt falls, but the product becomes reliable enough to use. Reliability creates the workload.

Cheaper intelligence does not remove travelers from the road. It may create more travelers, taking longer trips, with more cargo.

Who gets capacity when the road is full

PRICE PER TOKEN IS NOT THE ALLOCATION RULE

If compute supply tightens, capacity will not necessarily flow to whichever model charges the highest API price. Suppliers care about who can commit, pay and deploy.

The strongest customer offers a creditworthy multiyear contract, high utilization, a facility ready to receive equipment, strategic supplier relationships and the balance sheet to absorb premium pricing. That generally favors hyperscalers and well-funded frontier labs operating through them. Rapid revenue growth alone is not enough. Training and inference expenses can consume cash as quickly as a lab earns it.

Epoch AI estimates that global inference capacity is expanding rapidly, but its imperfect demand proxies are growing faster, especially for long-context agentic work. That is evidence worth watching, not a

settled forecast. It tells us where scarcity could emerge and which customer attributes would matter if it does.

Translate the model story into physical demand

THE BILL OF MATERIALS BEHIND CHEAPER INTELLIGENCE

More inference reaches beyond accelerators. Large models need custom silicon and high-bandwidth memory to calculate, conventional memory to hold data, optical and electrical interconnects to move it, and switches to keep clusters working as one machine. The facilities around those clusters need generation, transmission, cooling, security and construction.

Optimization belongs in the stack too. Better software, lower precision, caching and specialized chips can deliver more tokens from the same equipment. That can reduce hardware required for a fixed workload. It can also lower the cost enough to create a much larger workload.

None of this means every infrastructure company wins equally. Architecture, market share, pricing, utilization, customer concentration and capital discipline determine who captures the economics. A rising compute tide can still strand the wrong product or the wrong balance sheet.

The case against

WHAT WOULD WEAKEN OR BREAK THE THESIS

Our conclusion is an informed thesis: aggregate compute consumption will grow faster than efficiency improves. It fails if the usage response is too small.

> ### The applications do not pay
> Enterprise AI may fail to produce enough return, or agents may remain too unreliable for broad deployment. Cheap tokens do not create demand for work nobody values.

> ### Efficiency wins the race
> Smaller models, better software or faster on-device migration could reduce centralized demand faster than new usage expands.

> ### The buildout outruns useful demand
> Infrastructure construction can exceed economically useful workloads. Model providers may absorb falling prices without generating enough incremental consumption to fill what gets built.

What to watch next

EVIDENCE BEFORE CONVICTION

Watch token volumes, not token prices alone. Watch whether agentic products move from trials into recurring work, whether context lengths and reasoning budgets keep expanding, and whether enterprises report measurable returns. Watch accelerator utilization, HBM and networking lead times, power commitments, data-center occupancy and the terms suppliers demand from customers.

And after July 27, watch what independent researchers find when Kimi's weights become available. The release should clarify real deployment requirements, achievable throughput and how much of the reported performance survives outside Moonshot's own stack.

THAT IS WHAT WE READ THIS WEEK.

Kimi may pressure the economics of premium model providers. It may also expand the number of developers and companies able to buy useful intelligence. Those outcomes can happen together.

The model race asks which traveler wins. Our question is how many travelers arrive, how far they go and who owns the physical road they all have to use.

Otto Analytics

DOWNLOAD THIS ISSUE (PDF)

THE WEEKLY READ RETURNS NEXT WEEK.

PAST ISSUES

No. 04 · July 16, 2026 · *The Machine Is Only the First Sale* **Read →**

No. 03 · July 10, 2026 · *The Second Door* **Read →**

No. 02 · July 3, 2026 · *The Purchase Order* **Read →**

No. 01 · July 2, 2026 · *The Robot Is Not the Product* **Read →**

SOURCES

Axios, AI race splits in two as China wages open-weight insurgency, July 18, 2026. Accessible news framing on model competition, pricing and open-weight adoption.

Moonshot AI, Kimi K3 technical announcement, July 2026. Company-reported specifications, architecture, availability, benchmark methodology and planned weight release.

Stanford HAI, AI Index 2025: State of AI in 10 Charts. Historical decline in quality-adjusted inference prices.

Gartner, inference cost and token consumption forecast, March 25, 2026. Estimates for falling unit costs and higher agentic token demand.

Epoch AI, Is a compute crunch coming?, 2026. Estimates of inference capacity, token demand and uncertainty.

Anthropic Economic Index, Cadences, June 2026. Observed relationship between token consumption, task value and autonomy.