



WHAT I READ THIS WEEK

Open Models, Closed Economics: Who Actually Pays for “Free” AI?

How open weights change model margins, infrastructure demand, cloud economics and American AI strategy

OTTO ANALYTICS · THE WEEKLY READ · NO. 06 · JULY 24, 2026

Open weights may commoditize parts of the model layer without commoditizing the AI economy. The weights can be downloaded. A useful production system still needs chips, memory, networking, power, software, security and skilled operators. That shifts pricing pressure toward models while expanding the number of organizations that can buy everything around them.

On Friday, a broad coalition of American technology companies published [a joint letter on open weights and American AI leadership](#). The signatories span model developers, cloud and hardware vendors, enterprise software companies, investors and developer platforms. Their argument is that downloadable models can broaden access, increase competition and give organizations more control over where AI runs.

That is the strategic case. The economic case is more interesting.

If the model weights are available at little or no upfront cost, who funds their development? Who pays to run them? Which companies lose pricing power, and which sell more of the physical and software stack underneath?

The answer begins with a distinction the industry often blurs.

Three models of access

CLOSED, OPEN-WEIGHT AND OPEN-SOURCE ARE NOT SYNONYMS

A model is a trained mathematical system. Its *weights* are the learned numerical parameters that shape what it produces. Access to those weights determines how much control a user has, but weights alone do not reveal the complete training recipe.

A PLAIN-ENGLISH COMPARISON

What does the customer actually control?

Closed model: rent the intelligence

The provider keeps the weights and usually runs the system. The customer pays through an API, subscription or enterprise contract. Deployment is simple, but control and portability are limited.

Open-weight model: take the trained engine

The customer can download the weights, modify them and run the model on chosen infrastructure, subject to its license. The training data and complete training code may remain undisclosed.

Open-source AI: study and change the full system

Under the Open Source Initiative's definition, users need the freedom and preferred form needed to use, study, modify and share the system. That includes weights, code and detailed information about the training data.

That is why we use *open-weight* unless a system meets the stronger open-source standard. The terminology matters because the commercial rights, transparency and operating burden can differ substantially.

A free model is not a free system

DOWNLOADING THE WEIGHTS IS THE BEGINNING OF THE BILL

Imagine a manufacturer is given a free industrial engine. The engine is valuable, but it is not a working factory. The company still needs a building, electricity, machinery, controls, technicians, safety systems and maintenance.

Open-weight AI follows the same logic. Production inference requires accelerators to perform calculations, memory to keep data near those chips, networking to move information, storage for the model and its context, and power and cooling for the facility. Then come inference software, monitoring, cybersecurity, evaluation, fine-tuning and the engineers who keep the service reliable.

A company can avoid paying a frontier lab for every token and still spend heavily to own the outcome. The cost moves from a visible API invoice into cloud consumption, equipment, software and operations.

WHO PAYS FOR “FREE” AI?

Follow the capital through the stack

1. Who can subsidize development?

Advertising and platform cash flow, hardware and cloud profits, venture capital, governments, sovereign funds and enterprise funding.

2. The weights become broadly available

More developers and enterprises can build without training a foundation model from scratch.

3. Spending reappears around the model

Accelerators and memory, cloud and neocloud inference, networking, storage, power, cooling, security, support, fine-tuning, deployment tools, proprietary applications and distribution.

The weights can be free because the ecosystem around them is not.

Why would anyone give the weights away?

NOT EVERY SPONSOR NEEDS TO MONETIZE THE MODEL DIRECTLY

Closed-model providers monetize access. They can sell API calls, subscriptions, enterprise licenses, premium reasoning, agentic tools and applications. Their advantage is a direct tollbooth around the model.

An open-weight sponsor can monetize indirectly. A platform company may want a common model ecosystem that improves its products and reduces dependence on another provider. Meta has said that selling model access is not its business model and that an open ecosystem can reduce lock-in around its own technology.

For a chip company, every additional deployer is a potential buyer of accelerated compute and production software. For a cloud, a portable model can still generate paid storage, networking and inference consumption. NVIDIA's NIM offerings illustrate the pattern: some inference microservices are

free, while its enterprise-certified tier adds lifecycle, security and support guarantees and requires NVIDIA AI Enterprise.

These businesses do not need the model layer to keep all the margin. They need it to create more workloads for the layer where they already earn money.

Who gains, and who feels pressure?

THE MODEL LAYER GETS MORE COMPETITIVE AS DEPLOYMENT SPREADS

Frontier model labs

More price competition for routine tasks, while the hardest problems can preserve premium pricing.

Accelerators, memory and networking

A larger deployment universe, tempered by model efficiency, custom silicon and price competition.

Hyperscalers and neoclouds

More hosting demand and portability, alongside sharper competition for the same workload.

Enterprise software and applications

More choice for embedded and customized AI. Proprietary data, workflow and distribution become more valuable.

Sovereign and edge deployments

More control over data location, latency and infrastructure, especially where dependence on a remote API is unacceptable.

Enterprise customers

Lower switching costs and more control, but greater responsibility for security, reliability and operations.

The joint letter makes this diffusion part of American AI strategy. Its signatories argue that broad access can spread AI through factories, hospitals, farms, classrooms and small businesses while increasing competition across chips, clouds, applications and services. It also acknowledges a real tradeoff: once weights are released, the original developer cannot fully control or recall modified versions.

Openness expands capability and responsibility at the same time.

The closed-model bull case remains strong

CONVENIENCE, ACCOUNTABILITY AND FRONTIER PERFORMANCE STILL MATTER

Training frontier models remains enormously expensive. Stanford's AI Index estimates several recent frontier training runs in the tens or hundreds of millions of dollars. The strongest closed providers can spread that investment across many customers and improve the service centrally.

Many enterprises will prefer a provider that handles reliability, updates, security, support and contractual accountability. The best closed models can retain pricing power on difficult tasks where capability matters more than unit cost. Intelligence delivered as a managed service is also much easier than assembling and operating the stack yourself.

Open weights do not eliminate closed models. They force closed providers to prove where the premium is earned.

The open-weight bull case is broader deployment

LOWER MODEL RENTS CAN ENLARGE THE INFRASTRUCTURE MARKET

Open weights reduce switching costs, permit deeper customization and support private, on-premises and sovereign deployments. A specialized company can adapt a smaller model to a narrow task. A government can keep sensitive workloads inside its own boundary. A device maker can move inference closer to the user for lower latency or better privacy.

Each deployment may be smaller than a frontier training cluster. The number of deployments can be vastly larger. That is the infrastructure opportunity: not one model consuming everything, but thousands of organizations operating many models across clouds, private data centers and edge devices.

The investment question underneath the argument

DOES USAGE GROW FASTER THAN EFFICIENCY?

Open models can reduce compute required for a fixed task by encouraging smaller models, better software, quantization and custom chips. They can also make AI cheap and controllable enough to enter many more workflows. The decisive variable is not model price alone. It is aggregate compute consumed after lower prices, new applications and more efficient systems interact.

The infrastructure thesis weakens if efficiency improves faster than new usage grows, if more work moves to low-power edge devices, or if custom silicon sharply reduces spending captured by today's suppliers. It strengthens if lower model costs unlock enough new inference, customization and agentic work to overwhelm those savings.

That is the Jevons question for AI infrastructure. More efficient intelligence does not guarantee more physical demand. It lowers the economic barrier that new demand must clear.

THAT IS WHAT WE READ THIS WEEK.

We expect a hybrid market. Closed models should retain the hardest frontier work and customers who value convenience and accountability. Open weights should spread across countries, enterprises, devices and applications that value control and customization.

As the model layer becomes more competitive, economic power can migrate toward infrastructure, distribution and proprietary applications. NVIDIA does not need every open model to win. The larger opportunity is that openness multiplies the number of organizations building AI, and every production deployment still has to run somewhere.

Otto Analytics

DOWNLOAD THIS ISSUE (PDF)

THE WEEKLY READ RETURNS NEXT WEEK.

PAST ISSUES

No. 05 · July 19, 2026 · *The Cheaper Intelligence Gets, the More Infrastructure It May Need* **Read →**

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

[Open Weights and American AI Leadership](#), July 24, 2026. Joint industry letter on access, competition, customer control, ecosystem diffusion and risks.

[Open Source Initiative, Open Source AI Definition 1.0](#). Definition of open-source AI and the required access to data information, code and parameters.

[Meta, Open Source AI Is the Path Forward](#), July 23, 2024. Company explanation of its model strategy, ecosystem incentives and position on monetization.

[NVIDIA NIM for Large Language Models, NIM Offerings](#). Product tiers, production availability and enterprise licensing requirements.

[Stanford HAI, 2025 AI Index Report](#). Estimates of training costs for selected frontier models and broader industry data.

Otto Analytics publishes this column for informational and educational purposes only. This commentary is impersonal analysis offered uniformly to all subscribers and the general public. It is not investment advice, not an offer or solicitation, and does not account for your individual circumstances. Any ratings or conviction scores reflect our editorial assessment and are not recommendations to buy, sell, or hold any security. The author or Company may hold positions in securities discussed in this column. All investing involves risk, including loss of principal.

© 2026 Otto Analytics, LLC. All rights reserved.