

Enterprises Embrace Open Weights but Infrastructure Sellers Profit Most

Futurum Data Show Enterprises Moving Routine AI Work to Open Weights – Platform Vendors Capturing the Gains and Frontier Labs the Spending

Analyst(s): Nick Patience

Publication Date: October 26, 2026

Document #: AIONP202610

What You Need to Know

- **Open Weights Reach Production:** Open-weight models now run in production at 42% of enterprises in Futurum's September 2026 drill-down (N=200), up from 31% in July. Production users expect open weights to carry 41% of their AI tokens within a year, up from 34% today.
- **Platform Vendors Take the Share:** Google's Gemma runs in 55% of production accounts, Microsoft's Phi debuted at 39%, and NVIDIA's Nemotron at 25%, while Meta's Llama fell to 35% from 55%. Meta's new Muse Spark Open will reach enterprises mostly through the same hyperscaler catalogs.
- **Origin Matters More:** Country of origin is now a major or deciding factor for 47% of production users, up from 35%, and export-control exposure matters to 30%, up from 10%, as Chinese families lose ground in enterprise production.
- **Developers Move Faster, Toward Different Models:** Open weights carry most tokens on the Vercel and OpenRouter gateways, led by Chinese models, yet take only a small share of the spending.
- **Frontier Labs Keep the Money:** Open-weight models account for just 5% of enterprises' metered AI consumption in Futurum's October 2026 AI Tools Pulse, against 36% for Microsoft 365 Copilot and 29% for Claude, and in-house open-weight builds report the weakest ROI of any approach.

The Futurum View

Open-weight models have won the argument over whether enterprises will run them, so the more important question now is who profits. Our data suggest the answer is the companies that sell the infrastructure underneath the models. Google, Microsoft, and NVIDIA give their open models away to sell cloud capacity and chips, and enterprises are taking up the offer, while the frontier labs lose token share but keep most of the money. Meta's return with Muse Spark Open in October 2026 is a serious attempt to break into that position, but the model

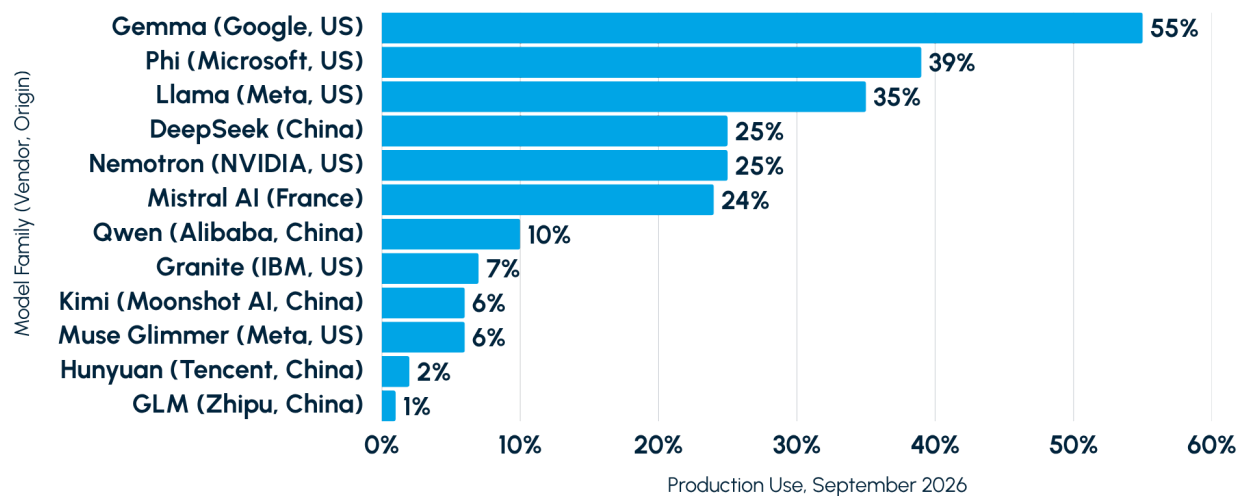
reaches enterprises through the same hyperscaler catalogs, under a more restrictive license, with Meta's best model still held back.

The move into production has happened faster than most enterprise buyers expected. In our September 2026 drill-down, 42% of 200 enterprises run open-weight models in production, up from 31% just two months earlier, and only 3% have never evaluated one. Our respondents are predominantly North American, so these figures describe the US market most closely. Among the 83 production users, open weights carry an average of 34% of AI tokens, against 23% a year ago, and those users expect 41% within 12 months. That works out to roughly half of each new token they consume. Most of that work is routine. Low-complexity workloads rose to 51% of production users' open-weight use from 39%, and they are the high-volume tasks where agent loops consume the most tokens. We expect a split between cheaper workhorse models for routine work and premium frontier models for the hardest reason to become the standard enterprise architecture over the next year.

The Platform Vendors Take the Share

Google's Gemma now runs in 55% of production accounts, up from 35%, while Microsoft's Phi entered our list at 39% and NVIDIA's Nemotron at 25% in their first survey cycle. Three of the five most-used families in production now come from companies that sell the cloud capacity or the chips those models run on, and each of them can afford to give a model away to win the infrastructure spending behind it. Meta's Llama has paid the price. It led both the production and pilot groups in July, but now runs in 35% of production accounts, down from 55%. In our broader AI Product Series (N=511), more respondents have tried and dropped Llama (15%) than plan to keep using it (9%). Non-US families lost ground over the same period. DeepSeek fell to 25% of production accounts from 32%, Mistral to 24% from 32%, Alibaba's Qwen to 10% from 13%, and Zhipu's GLM to 1% from 10%, while Moonshot's Kimi held at 6%.

Figure 1: Open-Weight Model Families in Production Use



Source: Futurum Research, *Open-Weight AI Models Drill Down, production users, N=83 (September 2026), of which 56 in North America and N=31 (July 2026); multiple responses allowed.*

Meta is now trying to regain that ground with a different model. Muse Spark Open is an 816-billion-parameter mixture-of-experts model built for coding and agentic work, derived from the Muse Spark model Meta kept proprietary at its April launch, and Meta describes it as the most capable open-weight model yet from a US lab. Our data predates the launch, but three details suggest that the platform vendors keep the advantage in the enterprise. Meta is distributing the weights through Amazon SageMaker, Google Cloud Model Garden, and Microsoft Foundry, so enterprise adoption will run through the hyperscalers' catalogs. The Muse Model License carries limited commercial terms and a required product label, which will slow approval in enterprises where incomplete security and compliance reviews already hold up 62% of pilots. Meta has also kept its most capable model, Muse Wonder, proprietary. Like the frontier labs, Meta is opening its second tier and keeping its top tier closed.

The Frontier Labs Keep the Money

Developer traffic is moving faster still and toward different models. Two independent gateways show the same pattern. Vercel reported open-weight models handling 56% of token volume on its AI Gateway in August 2026, up from 13% in April, while taking only 14% of estimated gateway spending. OpenRouter reported that open-weight models carried 72% of its tokens in mid-September 2026, with Chinese labs alone accounting for 69%. Both platforms mainly serve developers and startups, so their data describes developer behavior rather than market share, and the volume is led by Chinese families such as DeepSeek, which hold a smaller share of enterprise production, where DeepSeek runs in 25% of accounts, down from 32% in July. Developers choose models largely on cost per token, while enterprises also weigh origin, compliance, and support, so the enterprise shift runs slower and favors different vendors.

Table 1: Open-Weight Share of AI Tokens and Spending

Population	Open-Weight Share of Tokens	Open-Weight Share of Spending
Enterprise production users (Futurum, September 2026)	34% (average)	Not measured
Enterprise-metered AI consumption (Futurum AI Tools Pulse, October 2026)	Not measured	5% (average share of metered consumption)
Vercel AI Gateway (August 2026)	56%	14%
OpenRouter (mid-September 2026)	72%	Not shown

Source: Futurum Research – Open-Weight AI Models Drill Down, September 2026; Futurum AI Tools Pulse, October 2026 (N=100); Vercel AI Gateway Production Index, September 2026; OpenRouter, September 2026

Inside enterprises, the frontier labs face a smaller share of a larger pool, which is a more comfortable position than the headline shift suggests. Some 93% of production users say their proprietary usage is still growing. Asked how the two relate, 53% describe partial substitution, 40% call them complementary, and only 7% say proprietary usage has slowed or declined. The arithmetic supports the labs, too. If open weights rise from 34% to 41% of tokens, proprietary volume stays flat only if total usage grows by less than about 13%, and at 50% growth,

proprietary volume still rises by roughly a third. The pressure shows up in price rather than volume. At performance parity, 66% of production users would move a workload for a discount of 25% or less, so the labs have every reason to cut prices before buyers move work elsewhere. Anthropic's Opus 5.5, launched in September as a cheaper alternative to its Fable model, is an early example.

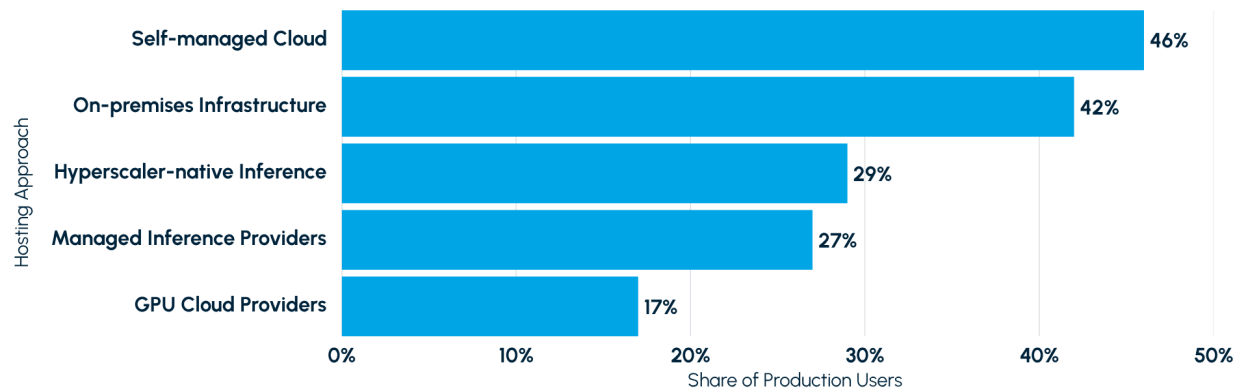
Our October 2026 AI Tools Pulse (N=100) shows how little of the money has moved. Two-thirds of enterprises now use open-weight models in some form, up from 46% in August, yet open weights account for an average of only 5% of companies' metered AI consumption, against 36% for Microsoft 365 Copilot, 29% for Claude, 15% for ChatGPT, and 10% for Gemini. Part of the gap is measurement, because self-hosted models generate no vendor bill, but that supports our argument. The token volume moving to open weights leaves the AI budget and reappears as infrastructure spending, which is where the platform vendors make their money. Substitution also remains shallow. On average, 47% of open-weight users report reduced use of each proprietary tool, a figure that has not moved in three months, and only 8% describe the reduction as significant. Hard limits are proving difficult as well, with only 22% of enterprises imposing them against the 41% that expected to by October, and shifting to cheaper models is the most-cited lever among those planning limits, at 58%.

The frontier labs also hold a capability lead at the top, although its size moves with each release cycle. On the Artificial Analysis Intelligence Index, which does not yet include Muse Spark Open, the best open-weight model trailed the leaders by about seven points six months ago and by about four points in July, according to Mozilla's analysis, but the autumn frontier launches have pushed the gap back to 12 points, with 10 proprietary models now ahead of the best open one. The strongest open models are mostly Chinese, while the open families that enterprises run in production score far lower, with NVIDIA's Nemotron 3 Ultra at 23 against 58 for the leading proprietary model. Meta's new model narrows the distance for US open weights, although early comparisons show it trailing Kimi K3 and GLM 5.3 on agentic coding benchmarks. The hardest reasoning stays with frontier models, and 22% of production users would not switch at any price when an open model falls well short.

The Savings Come With Hidden Costs

Enterprises should be realistic about where the savings go. Self-managed cloud (46%) and on-premises infrastructure (42%) lead production hosting, well ahead of hyperscaler-native inference (29%) and managed inference providers (27%). The cost leaves the token bill and reappears as servers, cloud compute, and engineering time. Self-hosting costs now hold up 50% of pilots, up from 40%. Our Macro Views survey adds a sharper warning. In-house open-weight builds report sustained ROI at scale in only 10% of cases, with 25% reporting no traction, against 17% at scale for external solutions, and 52% of non-users say they lack the self-hosting and fine-tuning skills required. For most enterprises, buying open weights through a platform will make more sense than building around them.

Figure 2: Where Enterprises Host Open-Weight Models in Production



Source: Futurum Research – Open-Weight AI Models Drill Down, 83 production users, September 2026 (multiple responses allowed)

The country-of-origin screen is also tightening. Origin is now a major or deciding factor for 47% of production users, up from 35%, and export-control exposure matters to 30%, up from 10%. Pilot use of Moonshot's Kimi fell to 6% from 13%. Mistral, the most widely used European family, runs in 24% of production accounts, down from 32% in July, so it has yet to gain from the tighter origin screen. Buyers increasingly want open models from a jurisdiction they trust. That favors US vendors with US buyers, now including Meta, and could leave room for European providers with European buyers, provided they can match the platforms on hosting and support, although our largely North American sample cannot yet measure that.

And it should be noted that these numbers describe the leading edge rather than the whole market. Across the 511 enterprises in our AI Product Series, only 14% call open weights core to their AI strategy, and 26% rate open-weight availability as an important factor in model selection, up from 20% in March. The direction is clear, however. Open weights will become the default for routine, high-volume enterprise work, the frontier labs and Meta will keep their most capable models closed, and the value in open weights will accrue mainly to the vendors that sell the infrastructure underneath them.

What to Watch

- **Muse Spark Open's Enterprise Uptake:** Whether enterprises adopt Meta's new model despite its restricted license, and whether it lifts Meta's open-weight position from Llama's 18% Net Score, down 28 points year over year.
- **Frontier Repricing:** Further price cuts, usage-commitment discounts, or seat-plus-consumption bundles from Anthropic, OpenAI, and Microsoft as they defend workloads within buyers' 25% switching band.
- **The Capability Gap:** Where Muse Spark Open lands on independent benchmarks, and whether US open models close any of the distance to the frontier or remain lower-cost models for routine work.

- **Developer and Enterprise Convergence:** Whether enterprise open-weight token share moves toward the levels seen in developer gateways, and whether Chinese models gain any enterprise ground as they have with developers.
- **European Sovereign Providers:** Whether the tightening origin screen creates room for European open-weight providers in EMEA, or whether US vendors' hosting reach wins there too.
- **Self-Hosting Economics:** Whether managed and hyperscaler-native inference options reduce the cost barrier holding up half of pilots, shifting open-weight work away from enterprise-owned infrastructure.

Other Insights from Futurum

[Why Enterprise AI Budgets Keep Growing Despite AI Safety Pledges](#)

[NVIDIA Nears \\$12.9B Deal for Hugging Face, Escalating AI Ecosystem Strategy](#)

[Open Source vs. Proprietary AI: Revolution or Just Another Market Split? – Report Summary](#)

About Us

About the Authors

Nick Patience is VP and Practice Lead for AI Platforms at The Futurum Group, covering AI infrastructure software, hyperscaler AI platforms, foundation model providers, and AI-native tools. He co-founded 451 Research (acquired by S&P Global in 2019) and has covered the technology industry for approximately 27 years. He can be reached at npatience@futurumgroup.com.

About The Futurum Group

Every day, The Futurum Group's analysts, researchers, and advisors help business leaders worldwide anticipate tectonic shifts in their industries and leverage disruptive innovation. Unlike traditional analysts, The Futurum Group works not only in analysis and research but also takes that insight and knowledge even further, engaging all the way through the go-to-market process.

Futurum Research provides in-depth research and insights on global technology markets using advisory services, custom research reports, strategic consulting engagements, digital events, go-to-market planning, and message testing. It also creates, distributes, and amplifies rich media content that all stakeholders read, watch, and listen to.

See more details on The Futurum Group at futurumgroup.com.

Copyright & Use License

Copyright Notice

Copyright ©2026 by The Futurum Group, LLC. All rights, including that of translation into other languages, are specifically reserved. No part of this publication may be reproduced in any form, stored in a retrieval system, or transmitted by any method or means, electrical, mechanical, photographic, or otherwise, without the express written permission of The Futurum Group futurumgroup.com. United States copyright laws and international treaties protect this publication. Unauthorized distribution or reproduction of this publication, or any portion of it, may result in severe civil and criminal penalties and will be prosecuted to the maximum extent necessary to protect the publisher's rights.

License Notice

This document may be distributed within the licensed organization only.

The following acts are prohibited:

Transmittal to others outside your immediate organization, including partners, resellers, external consultants, etc., in any media format

Posting on a website that is accessible to others outside your immediate organization

The possession or use within an unlicensed organization

Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. The Futurum Group shall have no liability for errors, omissions, or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve their intended results. The opinions expressed herein are subject to change without notice.