

Run Open-Weight AI for Up to 90% Less

How Enterprises Can Dramatically Reduce the Cost of Self-Hosted Models

A technical brief for engineering leaders · Rival · 2026

The hidden cost of open-weight AI

Open-weight models are effectively free to license. Yet the total cost of running them in production has become one of the fastest-growing line items in enterprise engineering budgets. The reason is not the model itself. It is the way the model is hosted, scheduled, and executed on your hardware.

Most teams provision for peak, over-size GPUs, and leave capacity idle. Workloads that could run on smaller hardware are pinned to the largest instances. Workloads that could share a node are isolated. The result is a bill that is two to ten times higher than the actual compute required.

The 50-90% opportunity

Rival measures the real cost of running an open-weight model on your existing cloud and hardware, then optimizes execution without changing the model, the API, or the application. Across production deployments, the measured reduction in compute spend consistently falls between 50% and 90%.

Where the savings come from

- Right-sizing: matching each workload to the cheapest hardware that can meet its latency and throughput target.

- Batching and scheduling: packing requests so GPUs are active, not waiting.
- Quantization and speculated decoding: preserving output quality while using fewer FLOPs per token.
- Dynamic scaling: removing nodes during low-traffic periods instead of renting them around the clock.

Case study: DeepSeek at scale

A growth-stage technology company was self-hosting DeepSeek on a standard cloud GPU fleet. Monthly spend had reached \$35,000 and was climbing. They did not want to change models, retrain, or introduce a new vendor lock-in.

After applying Rival's execution layer on top of their existing infrastructure, measured monthly spend dropped to \$5,000. The same throughput, the same model weights, the same cloud account. The only difference was how efficiently the hardware was being used.

\$35,000 / month → \$5,000 / month · 86% reduction · zero model changes

What Rival does differently

Traditional AI infrastructure treats compute as a provisioned resource. Rival treats it as a measured runtime. That means every deployment is continuously benchmarked against cost, latency, and quality, and workloads are routed to the optimal execution path.

- No migration required: keep your existing cloud accounts, models, and endpoints.
- No model changes: run the same open-weight weights you already validated.
- Verified savings: we measure before and after on your own baseline.
- No lock-in: if we cannot beat your current cost, there is nothing to buy.

How to get a number for your stack

The fastest way to validate the opportunity is a 20-minute cost review. Bring one workload and your current monthly spend. Rival will benchmark the workload on your

existing hardware, model the optimized execution path, and return a measured before-and-after number. Most teams have a verified answer within two weeks.

Next step

If your team is spending more than \$10,000 per month on self-hosted AI compute, the savings are almost certainly material. Download this brief, share it with your platform or infrastructure lead, and book a cost review to get a number tailored to your environment.

Visit rival.io/graphx-campaign to book a 20-minute cost review.