Open-weight models in the AI market: token share, price and the OpenRouter–Hugging Face switch

23 November 2026 · 6 min read · blogs.iictp.org/open-weight-models-token-share-pricing-openrouter-hugging-face-explained

Short answer: on 16 November 2026 Hugging Face said it is moving Inference Providers to OpenRouter's router, and cited OpenRouter data showing open-weight models rose from 26.8% of tokens served in October 2025 to 71.7% at the end of September 2026. The same post says open-weight models cost about four times less on average, while the best open-weight model trails the best proprietary one on the hardest benchmark. The figures are OpenRouter's, reported by Hugging Face.

TL;DR The short version

- OpenRouter data relayed by Hugging Face shows open-weight models' share of tokens rising from 26.8% in October 2025 to 71.7% at the end of September 2026.
- Open-weight models cost about four times less on average, while the best still trails the best proprietary model on the hardest benchmark.
- Token share is not spend or value, and the source profits from open-weight adoption, so test on your own evaluation set.

Read online, listen to it, or change the text size: https://blogs.iictp.org/open-weight-models-token-share-pricing-openrouter-hugging-face-explained

What was announced?

Hugging Face says every major open-weight release is published on the Hub and OpenRouter serves those models, routing each request across 80+ providers by live price, speed and reliability. Language-model requests through Inference Providers now go through OpenRouter, which retries on another provider if the first fails. Your code stays the same.

Market data in the post

- Open-weight share of tokens on OpenRouter: 26.8% in October 2025, 71.7% at the end of September 2026, after peaking at 77.5% in July.
- Weekly open-weight volume grew more than 60-fold over the period, against almost tenfold for closed models.
- Average price per million tokens: $0.28 input and $1.62 output for open-weight, against $1.27 and $6.36 for closed models.
- The most-used model that week was an open-weight one, DeepSeek V4.1 Flash 1017, at 59.2 trillion tokens.

The quality gap the post admits

On the Artificial Analysis Intelligence Index, the best open-weight model (Kimi K3.1 Pro) scores 46 against 65 for Claude Fable 5.2, at about a seventh of the cost per task ($3.68 against $26.54), per the post. The authors say that for many workloads the trade-off is worth it.

How the routing is said to work

OpenRouter’s rebuilt routing scores providers by a mix of price, speed and reliability instead of list price alone. Hugging Face reports OpenRouter measured 46% fewer capacity errors, 16.9% better time to first token and 16.3% better throughput.

Limitations and open questions

- These are OpenRouter’s figures, relayed by a party that profits from open-weight adoption. Token share on one router is not market share across all AI usage.
- Token counts do not equal spend or value; cheap high-volume uses can dominate.
- We did not read the OpenRouter report itself or independent data.

How to use this, step by step

1. Estimate your own token volume and price against both categories.
2. Test an open-weight model on your own evaluation set, not a leaderboard.
3. Plan for provider fallback and data-handling terms.
4. Review the contract terms of each provider that serves the model.

What it means for AI budgets and architects

Model choice is now a routing and cost decision. See build vs buy in the AI era and local LLM vs cloud AI.

Frequently asked questions

- **What changed for Hugging Face users?** Inference Providers requests for language models go through OpenRouter’s router.
- **How much cheaper are open-weight models?** About four times, on average, per the post.
- **Are open-weight models as good?** The post says the best ones still trail on the hardest benchmarks.

Related reading

- Build vs buy in the AI era
- Local LLM vs cloud AI
- FinOps for engineers
- AI Practitioner

Sources and currency

Sources were read on 23 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Hugging Face. Hugging Face and OpenRouter join forces to make open-weight models easy to run. 16 November 2026. huggingface.co/blog/openrouter. Cites OpenRouter, October 2026 data.

Published by IICTP, an independent technical credential body, on 23 November 2026. The current version of this post is at https://blogs.iictp.org/open-weight-models-token-share-pricing-openrouter-hugging-face-explained. More guides: blogs.iictp.org.