Open Source AI: Myth vs Reality, 96% of Revenue Is Closed, and MIT Weights Are Not Open Source

The word “open” in AI now describes a download button, not a licence, and definitely not a legal status. Two things published on 15 September 2026 make that gap expensive rather than academic.

Mozilla’s State of Open Source AI v1.1, published 15 September 2026, reports that 96% of 2025 model-layer revenue went to closed providers and roughly 4% to open-weight providers, based on Linux Foundation analysis of OpenRouter traffic for May to September 2025. OSI’s Open Source AI Definition 1.0 says downloadable weights alone do not qualify as open source, and the EU AI Act’s Article 53(2) exemption is narrower than most teams assume.

Two numbers from the same research week do not sit comfortably together. Chinese-dominated open-weight models handled the majority of developer token traffic on

OpenRouter in August 2026, according to Tech Times' write-up of the Mozilla report. Closed providers took 96% of the money. Those are different measurement windows, which matters, but the direction is not ambiguous: open weights won volume and lost revenue.

Getting this wrong is not a reputational problem. It is a shipped product carrying a licence you cannot comply with, or a documentation duty you assumed an exemption covered, or a self-hosting bill that lands 17 times higher than the spreadsheet said. All of those come from the same root: treating "open" as one property when it is three separate ones (licence, artefacts, regulatory status).

The 96% revenue split is a pricing story, not a quality verdict

The Mozilla figure rests on a roughly 20% open, 80% closed usage split in the May to September 2025 window. That gap between usage share and revenue share is the thing worth saying plainly, because the 96% gets quoted as a judgment on model quality. It is a judgment on who gets paid.

The Mozilla figure covers May to September 2025. The token-traffic majority is August 2026. Anyone comparing them as a single trend is putting a year-old revenue snapshot next to a current usage snapshot. My read: the gap has probably narrowed since, because open-weight capability moved fast in that year, but I have no 2026 revenue split to prove it and I am not going to pretend otherwise.

What the revenue number does tell you is where the commercial gravity sits. If your procurement logic is "the market has chosen closed, so we should too", remember that the market chose closed for billing reasons, and your billing situation is not the market's.

The four claims that keep costing teams money

THE CLAIM	THE REALITY
If the weights are on Hugging Face under MIT, the model is open source.	OSI's Open Source AI Definition 1.0 requires weights, complete training and inference code, and enough data information to rebuild a substantially equivalent model. The Register's 15 September column lays out why "open weights" and "open source" get used interchangeably and should not be. MIT weights give you permissive reuse of an artefact. They do not give you reproducibility.
THE CLAIM	THE REALITY
Open-weight releases are overwhelmingly Apache 2.0 or MIT, so licence review is a formality.	A Digital Applied audit of 30 prominent 2026 open-weight models found only 17 of 30 (57%) under unmodified Apache-2.0 or MIT-style terms. Kimi K2.5 is listed as "modified-mit" and classified restricted. For US releases above 20B parameters, 41% carry custom terms and 30% declare no licence at all.
THE CLAIM	THE REALITY
Releasing under a free and open-source licence exempts us from the EU AI Act's GPAI obligations.	Article 53(2) removes only the technical documentation and downstream-provider information duties, per the Commission guidance summary of 4 September 2026. The copyright-compliance policy and the training-data summary remain. And the exemption does not apply at all to GPAI models classified as presenting systemic risk, whatever the licence says.

THE CLAIM	THE REALITY
Self-hosting an open-weight model is obviously cheaper than paying per token.	Ornn Data's 11 September 2026 study puts rented-hardware self-hosting at $0.12 to $0.35 per million output tokens at full utilisation, and points out that self-hosting has no per-token list price at all. Vercel's example with the same Mixtral 8x7B deployment: $15.25 per million output tokens at 1 request per second versus $0.87 at saturation. Utilisation alone moves the number 17.5x.

The licence numbers deserve a second look, because the geography is the surprise. Across 178 Chinese open-weight releases above 20B parameters, 59% are Apache 2.0 and 22% MIT, per the Hugging Face 2026 Open Model Report summary. The comparable US figure is 29% Apache or MIT. If your legal team's mental model is "Chinese models are the risky licence territory", the data points the other way on licence terms specifically. Export controls, data residency and procurement policy are separate questions and I am not addressing them here.

MIT weights clear the shipping bar and fail the open-source bar

For shipping, usually yes. For claiming open source, no. Those are different bars, and conflating them is where teams get into trouble.

MIT on weights buys you permissive redistribution and modification of the artefact you downloaded. DeepSeek-V4.1-Flash, which shipped 11 September 2026 under MIT for both repo and weights, reporting 90.6 on Terminal-Bench 2.1, 74.2 on DeepSWE v1.1 and 54.8 on AutomationBench at maximum reasoning effort, is a straightforward commercial artefact under those terms. I have written before about what the MIT licence buys you and what it does not in a different context, and the same boundary applies: permissive use of what you were given, no claim on what you were not given.

What MIT weights do not buy you: the ability to reproduce the model, the ability to audit

what went into it, any guarantee that the training data was lawfully sourced, or standing to call the result open source under OSAID 1.0. If your compliance story depends on knowing what the model learned from, permissive weights answer none of it.

The word “modified” in a licence field should stop a review. Modified-MIT is not MIT. The Digital Applied audit flagging Kimi K2.5 as restricted despite the MIT-adjacent label is the concrete case: the name on the tin and the terms inside diverge, and nobody catches that from a repo card at a glance.

So classify every model you depend on into three buckets before it reaches production: permissively licensed weights (Apache 2.0 or unmodified MIT), restricted or custom terms, and undeclared. **Undeclared is not permissive by default.** Read the licence file in the repo, not the label on the model card.

Apertus shows full openness, and why almost nobody ships it

Apertus is the reference point. Released 2 September 2025 by EPFL, ETH Zurich and CSCS in 8B and 70B sizes under Apache 2.0, it publishes weights, architecture, training recipes, data-reconstruction code and intermediate checkpoints, according to the ETH Zurich release. That is the shape OSAID 1.0 describes: the artefact plus the path to rebuilding something substantially equivalent.

It is also an academic and public-infrastructure release, which is the pattern. Publishing data-reconstruction code and intermediate checkpoints costs a commercial lab its training-pipeline advantage and exposes its data provenance to scrutiny. I expect the fully-open category to stay dominated by publicly funded consortia and a handful of labs treating openness as strategy rather than product. That is a prediction, not a finding.

The audit of the 60 most-downloaded Hugging Face repos on 17 September 2026 shows what the middle ground looks like in practice: 44 Apache 2.0, 6 MIT, 4 custom, 3 Meta or Google model licences, 1 OpenRAIL, 2 unstated. Two unstated among the sixty most-downloaded repositories in the ecosystem. The licence field is not a required field in practice, and download counts do not care.

> **Download availability tells you the model exists. It tells you nothing about whether you can ship it.**

Break-even swings 20x depending on which API you compare against

The break-even numbers are where the open-versus-closed debate usually collapses into a spreadsheet. Digital Applied's September 2026 cost guide puts single-H100 break-even against a cheap API at roughly 5.7 billion tokens per month at 60 to 70% sustained utilisation. Against a premium API, that drops to around 256 million tokens per month.

The spread between those two figures is more than 20x, and it comes entirely from what you are comparing against. Self-hosting to beat a premium API is a normal engineering decision at moderate scale. Self-hosting to beat a commodity API requires volume most teams do not have, plus the utilisation discipline to keep the hardware busy.

That utilisation point is the one I would press hardest. The Vercel example (same model, same hardware, 17.5x cost swing between 1 request per second and saturation) means **your self-hosting cost is mostly a function of your traffic shape**, not your model choice. Bursty internal tooling with idle nights sits near the worst case. Steady batch processing sits near the best.

MY TAKE

These myths persist because "open" is a single word doing three jobs, and a download button is the only part anyone checks. The 96% revenue split is real, and so is the roughly 20% open usage share it sits against, which means price per token is not the binding constraint for most buyers. Integration cost, procurement comfort and utilisation risk are. My expectation: revenue share moves toward open weights more slowly than usage did, and the shift shows up first in high-volume, predictable-load workloads where the break-even maths is unambiguous, not in general chat assistants.

Three questions that settle a model choice faster than a vendor call

Before you commit to a model, in this order.

What does the licence file say, not the model card? Unmodified Apache 2.0 or MIT means you proceed. Custom, modified, or absent means legal review before integration, and "absent" is the worst of the three because there is nothing to review.

Are you claiming open source publicly, or just using open weights? If the former, OSAID 1.0 sets the bar and almost nothing meets it. Saying "open weights" costs you nothing and is accurate.

Which Article 53(2) duties do you still carry? The copyright-compliance policy and the training-data summary survive the exemption. If the model is classified as presenting systemic risk, the exemption does not apply at all. The Mozilla report is a reasonable starting point for the market picture, but the regulatory text is what binds you.

One thing I am genuinely unsure about: whether the licence discipline shown in Chinese releases (59% Apache 2.0, 22% MIT across 178 models above 20B) holds as those labs commercialise. The US pattern of drifting toward custom terms at 41% suggests permissive licensing is a phase that ends when a business model arrives. It may or may not repeat.

The distinction I would ask every team to hold onto is boring and load-bearing: open weights is an artefact property, open source is a definitional claim under OSAID 1.0, and the Article 53(2) exemption is a regulatory status that depends on neither of them cleanly. Mixing those three is how a permissive-looking model ends up blocking a release. If you are working through a licence classification or a self-hosting break-even and want a second pair of eyes on the assumptions, send me the shortlist and I will tell you where I think it breaks.

WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →