



AI This Week: Open Weights Got Cheap and Ownable, From Decision-1 at \$0.042/M to Nvidia Circling Reflection AI



AI This Week: Open Weights Got Cheap and Ownable, From Decision-1 at \$0.042/M to Nvidia Circling Reflection AI

What happens to an Apache 2.0 license when the lab that published it gets bought? Wednesday I wrote about Reflection AI shipping 501B open weights. Friday, Nvidia was reported in talks to acquire the lab.

The license survives. Apache 2.0 is irrevocable for weights already published, so the Beam checkpoint you pull today stays yours regardless of who owns Reflection AI. What an acquisition can end is the roadmap: the next checkpoint, the bugfix release, the long-context variant. Between October 5 and 11, 2026, open weights got cheap enough to run on every request path (Microsoft's Decision-1 at \$0.042 per million input tokens, zero for output) and valuable enough that Nvidia, which reportedly put \$800 million into Reflection AI in a \$2 billion round in October 2025, is now reported to be discussing buying the whole thing.



Five open-weight stories between October 5 and 11

Microsoft shipped Decision-1 at \$0.042/M in, nothing out

A 9B model post-trained from Qwen3.5-9B for single-pass decision scoring, available via Microsoft Foundry. Zero-cost output is the detail that matters: the model returns a decision rather than an essay, so Microsoft can charge nothing for it. I wrote the full breakdown in [Microsoft Ships Decision-1: a 9B Model Built on Qwen3.5 That Costs \\$0.042/M In and Nothing Out](#) on October 10. Consequence: routing, moderation and triage calls that used to need a frontier model now cost close to nothing at scale.

Reflection AI shipped Beam: 501B total, 23B active, Apache 2.0

A mixture-of-experts model released October 8 under a permissive license, built at a \$25 billion valuation. Details in [Reflection AI Ships Beam: 501B Open-Weight MoE, 23B Active, Apache 2.0, Built at a \\$25B Valuation](#). Consequence: 23B active parameters means inference cost tracks a mid-size model while capacity tracks a very large one, which is the whole argument for self-hosting instead of renting.

Nvidia reportedly entered talks to buy Reflection AI

Reported October 10 to 11 by [startupfortune.com](#) and [Gate News](#). Reported structures under discussion: acqui-hire, technology licensing, a larger equity investment, or expanded chip and infrastructure support. No signed agreement, and no purchase price has been reported. The widely repeated \$25 billion is Reflection AI's valuation, not a deal price.



Bloomberg: tracked AI CSAM variants went from under 12 to over 500

Bloomberg reported that variants built on open-weight models jumped by that margin, covered in [Bloomberg: Open-Weight Models Now Drive AI CSAM Cases, Tracked Variants Jump From Under 12 to Over 500](#) on October 9. Consequence: the same irrevocability that protects your Beam checkpoint protects every fine-tune built from it, including the ones nobody wants.

December 2 is the Article 50 marking deadline

EU AI Act transparency and marking obligations for machine-readable marking of synthetic content ship on December 2, 2026. I laid out what to build in [How to Ship Article 50 Marking Before December 2, 2026: A CTO's Guide to the EU Transparency Code](#) on October 5. Consequence: if you self-host open weights, the marking obligation is yours, because there is no vendor between you and the output.

That last point is the one I would underline. The cheaper open weights get, the more of the compliance surface moves from your vendor's lawyers to your own engineers. A hosted API can ship provenance metadata as a platform feature. A checkpoint you downloaded and serve on your own GPUs cannot. Seven weeks to December 2 at time of writing.

An acquisition does not revoke weights already published

No, and it is worth being precise, because the question will come up in your next architecture review.

Apache 2.0 grants a perpetual, irrevocable copyright and patent license to anyone who received the artifact. A change of control at the publishing entity does not retroactively unwind grants already made. If you pulled Beam's weights on October 8, you hold those rights.

What an acquirer controls is everything that has not shipped yet. Future checkpoints can



be released under a different license, or not released at all. Fine-tuning recipes, evaluation harnesses and tooling that lived in the same repos can go quiet. The research team can be reassigned. Nvidia's reported options include acqui-hire and technology licensing, and in both of those the people and the IP move while public release cadence is beside the point.

My read: I would treat any open-weight model whose publishing lab is in acquisition talks as a model with a frozen upgrade path until proven otherwise. That is not a reason to avoid it. It is a reason to do the sizing work now. If Beam gets no successor, is the checkpoint you have good enough for the next stretch of your workload, and do you have the GPU budget to run it yourself? If the answer to either is no, what you are adopting is a vendor relationship with extra steps. None of the reported Nvidia structures have been confirmed by either company, so this is risk management rather than prediction.

Decision-1 and Haiku 5.5 are priced for different jobs

They landed three days apart and they are not competing for the same work, which is the interesting part.

Model	Shipped	Input / M	Output / M	Positioning
Microsoft Decision-1 (9B, from Qwen3.5-9B)	Oct 10, 2026	\$0.042	\$0	Single-pass decision scoring via Microsoft Foundry
Anthropic Claude Haiku 5.5	Oct 7, 2026	\$0.10	\$0.50	Fastest and cheapest Anthropic small model, prompts up to 100,000 tokens

Input is roughly 2.4x cheaper on Decision-1, and output is the difference between a bill and no bill at all, because Decision-1 emits a decision rather than prose.

So the architecture question stops being "which small model do I standardise on" and becomes "how many classes of call does my system actually make". Scoring, routing and yes-or-no gating are one class, and they are now effectively free per call. Generation, summarisation and anything a human reads is another class with a real output bill. **Teams running everything through one general-purpose model are paying generation prices for classification work.**



The honest caveat: I have no benchmark data comparing Decision-1 and Haiku 5.5 on the same task. What I have covers pricing, parameter counts and positioning. Price gaps this wide usually survive contact with benchmarks, but “usually” is doing a lot of work in that sentence, and you should run your own eval before rewiring a routing layer on the strength of a price list.

Azure’s 18-region gateway failure and a prerelease flag that lied

Two other items from the week belong in the same argument, because both are about what happens when the thing you depend on stops behaving as documented.

Azure’s Sweden Central AI outage was followed roughly 24 hours later by a gateway failure across 18 regions. That is wider than a regional incident’s usual blast radius. It is the kind of correlated failure that makes the single-provider inference story harder to defend, and it is also, bluntly, an argument for keeping a self-hosted open-weight model warm for the calls you cannot drop.

Microsoft Agent Framework’s `~prerelease` flag exposed a gap between documented behaviour and shipped behaviour. Small thing, same category. The documentation is a claim about the artifact; the artifact is the artifact.

And in CI/CD, I wrote on October 10 about wiring Strix-style autonomous remediation into pipelines, where the whole design question is the gate that stops the agent from pushing its own patch to production. Cheap inference makes autonomous remediation economically obvious. It does not make it safe. Those are separate decisions, and the cost curve will keep pushing teams to collapse them into one.

Two things I did not cover this week: Anthropic launched a Cyber Mission and Critical Infrastructure Defense Program on October 8 with 11 founding partners including Accenture, CrowdStrike, Dragos, Palo Alto Networks and Rockwell Automation. Separately, Arena was reported to have raised a \$200M Series B at a \$3.1B valuation, a figure that remains provisional pending company or top-tier financial confirmation.



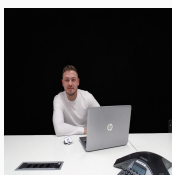
Capability and valuation moved, governance did not

Three things moved in the same direction between October 5 and 11, and the fourth did not move at all.

Open-weight capability got cheap: Decision-1 at \$0.042/M input, Beam at 23B active parameters out of 501B total. Open-weight labs got valuable: Nvidia reportedly in talks over a company it already funded with \$800 million. Open-weight abuse got measurable and ugly: under 12 to over 500 tracked CSAM variants, per Bloomberg. ****The governance layer is the thing that did not move.**** Article 50's December 2 deadline covers marking of synthetic content, a labelling obligation on output, and it says nothing about who may redistribute a fine-tune of a permissively licensed checkpoint.

My take: I do not think the answer is restricting open weights, and I would not pretend the CSAM number is an argument for that, because the models already shipped and Apache 2.0 does not come back. The answer is that organisations serving open-weight models are now doing a job that used to belong to a platform: provenance marking, abuse detection, output policy. That job has headcount attached to it, and almost nobody has budgeted for it, because the pitch for open weights was that it saves money.

My honest uncertainty is timing. I cannot tell you whether the Nvidia deal closes, in what form, or whether Reflection AI keeps publishing weights afterwards. What I can say is that a chip company owning a leading open-weight lab creates an incentive to keep weights open (weights sell GPUs) that sits awkwardly next to the incentive to differentiate (closed weights sell platforms). I lean toward the first incentive winning in the short term and the second winning as soon as inference margins matter more than chip volume. If you are sizing a self-hosting decision this quarter, [write to me and we will look at your actual workload](#) rather than the press releases.





AI This Week: Open Weights Got Cheap and Ownable, From Decision-1 at \$0.042/M to Nvidia Circling Reflection AI

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 →