

AI This Week: Sovereignty Got a Price Tag, From Apertus at CHF 0.70/M to Anthropic's $1.5T Listing

CHF 0.70 per million tokens. That is what a Swiss sovereign NLP stack on Apertus costs hosted, in a week whose other end of the scale was a $1.5 trillion vendor listing on Nasdaq.

Owning your AI stack now has a line item. Hosted Apertus runs at CHF 0.70 per million tokens, or €2.15 per GPU-hour if you operate the hardware yourself (post of November 18). On November 19, Anthropic priced its IPO at $83.00 per share, about $1.5 trillion fully diluted, according to NPR. Both numbers go in the same spreadsheet.

A published price and a public valuation can now sit side

by side

Six posts went up here between November 16 and November 21, written separately, without a shared plan. Read together they answer one question: what does sovereignty cost. For about two years that word was mostly a posture in European AI conversations. You were for it or against it, and the argument ran on jurisdiction, values and risk appetite. Now there is a published price for running a Swiss-origin model on European infrastructure, and a public market valuation for the largest independent alternative. You can defend the choice to a board in numbers.

The comparison is uncomfortable for both camps. CHF 0.70 per million tokens is not free, and €2.15 per GPU-hour is a rate rather than a total: utilisation, on-call engineering and model updates sit on top of it. And $1.5 trillion is a valuation. Nobody pays it. It does say something about the pricing power buyers will face over the next five years, and about how much capital the vendor has to defend its position.

Hosted Apertus wins below the throughput crossover

The two figures from the November 18 post are the ones to anchor on: CHF 0.70 per million tokens hosted, €2.15 per GPU-hour self-operated. The crossover point is a function of sustained throughput. Below a certain daily token volume the hosted rate wins, because you are not paying for idle silicon. Above it the hourly rate wins, because GPU time you already bought is marginal-cost-free.

My read: most teams asking the sovereignty question are nowhere near that crossover. Bursty internal workloads, a few thousand documents a day, no appetite for running an inference cluster. For them the hosted rate is the answer and self-hosting is a negotiating position, which is still worth having. The ones who genuinely clear the threshold usually know it, because their current API bill already made someone in finance ask a question.

Moving to aws-eusc is an account design problem

The November 21 piece on Moving to the AWS European Sovereign Cloud Without Breaking Your Account Structure lands on a point vendor messaging keeps losing: this is not a lift-

and-shift. Moving into aws-eusc requires reworking your account structure. Organisational units, identity, logging destinations, and the assumptions your IaC makes about where a region lives. If your migration plan is a line item called "re-point to new region", the plan is wrong.

I would call that the largest hidden cost in this week's set of numbers. Token prices are visible and comparable. Reorganising an AWS Organization that grew by accretion over six years is neither, and that work eats platform engineering time without producing a feature anyone can demo.

Speech and image procurement turns on where a model can legally run

Two of the week's posts are straight procurement work. The gpt-transcribe vs Parakeet v3 vs Voxtral vs Scribe comparison from November 20 exists because transcription now has four credible suppliers, and a CTO signing a contract needs to know which of them can run where. The November 19 post covers the open-weight image models FLUX.2 klein, MiniMax H3-Image and Qwen-Image-2.1-Turbo, assessed on what a Swiss CTO can legally ship, which is narrower and more useful than asking which one looks best in a cherry-picked gallery.

The legal filter is the part I would not skip. Open weights and permissive licensing are different things, and "we can download it" is not the same claim as "we can put its output in a customer-facing product in this jurisdiction". That distinction is where open-model adoption plans quietly fail at legal review.

Two of three runtime firewalls tie you to a hyperscaler

The November 16 post compares Bedrock Guardrails, Azure Prompt Shields and LlamaFirewall as runtime controls. The tie to sovereignty is direct: two of the three bind you to a hyperscaler control plane, one does not, and that choice outlives the model choice underneath it.

The November 17 post is the odd one out and the one I enjoyed writing most. Its argument

is that the AI freeze lawsuit fails because a court cannot operationally define "more capable". Injunctive relief needs a boundary a defendant can comply with and a judge can test. "Do not train anything more capable than what exists today" has no such boundary, because capability is measured on benchmarks chosen after the fact and gamed before it. The policy energy behind the freeze is real. The legal instrument is not available.

Brussels set a date while Nasdaq set a price

The IPO took the headlines, and the European Commission spent the same week on the less dramatic work. The Anthropic offering closed on November 20 with roughly $93.5 billion in gross proceeds including full exercise of the underwriters' option, and InformationWeek reports that the combined total with selling-stockholder shares and a concurrent private placement reached about $119.5 billion, with Microsoft, Nvidia, Google, AMD, Broadcom, Micron and TSMC in the placement. Your model vendor's balance sheet is now partly owned by your chip vendor and your cloud vendor.

In Brussels, the Commission opened its Apply AI Summit on November 16 with close to 1,000 policymakers, startups and researchers attending. It is collecting feedback until January 15, 2027 on a draft template for deployers' Article 27 fundamental-rights impact assessments, an obligation expected to bite from December 2, 2027. AI Act enforcement by the AI Office and national authorities started August 2, 2026, and the June amendments entered into force on July 27, 2026.

Separately, Microsoft announced a family of seven models at Ignite 2026 aimed at lower-cost enterprise workloads, with Microsoft-Decision-1.1 generally available per the Ignite Book of News. I have not evaluated any of them. The direction is the one everybody is moving in: smaller, cheaper, task-shaped models for the enterprise work that never needed frontier capability.

The Article 27 template matters more here than the IPO does

December 2, 2027 sounds distant. It is not, if the assessment requires you to document data flows, jurisdictions and processing locations for systems you are procuring now. A

deployment decision made in early 2027 on a stack you cannot describe in those terms becomes rework in late 2027. That, rather than ideology, is the strongest practical argument for the sovereignty numbers mattering.

My take: the week's accidental theme is that the sovereignty debate has moved to finance and legal, and the engineering question has narrowed to whether you can give them a number and a jurisdiction for each component. **CHF 0.70 per million tokens is a number.** "We use a US API with European data residency promises" is not one that survives an Article 27 template. The teams with an easy 2027 are the ones who can produce a component-by-component table this quarter, even if half the rows say they are staying where they are.

One thing I am genuinely unsure about: whether self-hosting survives contact with Microsoft's seven-model enterprise push and whatever Anthropic does with $93.5 billion of fresh capital. Cheap, small, hosted models compress the economics of running your own hardware from the other direction. The €2.15 per GPU-hour figure is defensible today. I would not assume it still reads as a bargain in eighteen months, and I would write any self-hosting business case with an exit that does not require throwing away the integration work.



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