

The token is dying. The platform *isn't.*

Will frontier-models-as-a-service fail? Both sides of that fight are right — about different layers.

Four forces are eating the *raw-token* business.

Open weights. Private LLMs. Private-cloud inference.
Desk-side hardware.

Each one is real. None is unbounded. Here's the evidence
—

13 weeks

is all it takes for open models to catch any given closed release. They trail the frontier by ~4 months overall — at **\$0.23 vs \$1.86** per million tokens.

Chinese open models went from under 2% to a *majority* of OpenRouter token volume in 18 months.

SOURCES: MIT SLOAN • EPOCH AI • OPENROUTER

65%

of national governments will hold sovereign-AI policies by 2028 (Gartner). On-prem vLLM, sovereign clouds, and air-gapped clusters moved open models into **regulated procurement**.

In finance, healthcare, and government, compliance overrides the TCO spreadsheet.

100–500M

tokens per month is the self-hosting break-even against frontier APIs. Above it, every locally routed token is one the API never bills again.

Below it, the “oversight tax” — idle GPUs, inference engineering, ops — eats the savings alive.

\$4,699

buys a desk-side box that runs 100–200B-parameter models. Dell’s GB300 desktops reach a **trillion** parameters. MoE made frontier-class quality fit in local memory.

The trajectory: local *good-enough* — one to two generations behind, forever cheaper.

Industrial costs. Software *multiples*.

~30%

Gross margin on frontier inference —
vs 75–85% for software (Epoch AI)

\$20.9B

One lab's 2025 operating loss on
\$13.07B booked revenue (FT-verified)

\$725B

2026 hyperscaler capex — the cross-
subsidy keeping the race funded

Inference is profitable. Training the next model consumes all of it.

*If an open model matched yours tomorrow,
would your customers leave?*

RAW TOKEN API

Yes.

A feature with a
13-week half-life.

EMBEDDED PLATFORM

No.

*The harness
is the moat.*

And yet — the buyers keep choosing *closed*.

~87%

of enterprise workloads still run on **closed models** — and open-source share **fell** from 19% to 11% last year (Menlo Ventures).

\$2.5B

run-rate for one coding agent, nine months after launch. The labs are converting tokens into **platforms** before the window closes.

Even open-weight champions blinked: Alibaba and Meta closed their frontier flagships.

STRATIFIED HYBRID

Open wins the *volume*.
Frontier keeps the *revenue*.

The thesis is right on tokenomics, wrong on mortality. Raw tokens commoditize — the labs transform into platform companies, or die trying.

If you own an AI budget —

- 01** **Stratify.** Route ~80% of traffic to open or mid-tier models. Save 30–50%.

- 02** **Respect the break-even.** No self-hosting below ~100–500M tokens/mo.

- 03** **Standardize the seams.** The vendor's moat is a base URL. Keep it swappable.

- 04** **Default regulated work to sovereign.** A capability lag is a compliance cost.

- 05** **Underwrite the platform transition, not the token margins.**

Not “*will the labs fail?*”

But: “*will their platform transition outrun their product’s commoditization?*”

That race is still open. Full 6-page brief in the comments.