Very cheap is basically free: pricing a coding model per step

A free model id in my config silently had no hosts, and the real price of a cheap coding model depends on which hosts you will accept. The per-step arithmetic, and why I want a price ceiling instead of a free switch.



Published
1 November 2026

Reading time
3 minutes

About
Mullet

Tags
AI

More on: Mullet

Mullet's ladder starts with free models. On 1 November I found that my free rung had nothing behind it, and Mullet had given no warning. While I was checking that, I worked out what the cheap rung costs per step. The answer is small enough that I question where I draw the line between free and paid.

A free model with no hosts

My config named a free model: `nvidia/nemotron-3-super-120b-a12b:free`. OpenRouter's endpoint API still answered HTTP 200 for that id, but it listed 0 endpoints. The live model list had 14 free-suffixed ids, and this was not one of them. The paid id of the same model had 3 hosts.

So the id resolved and there was nothing behind it: the free rung had no host to send to, and Mullet gave no warning. Check model ids against the live list, not against a config that has not complained.

A caution on free endpoints: they may log prompts. Mullet uses them only on public repositories.

What a cheap host costs

GLM-5.3-Flash on OpenRouter has 40 hosts. The cheapest is $0.0285 per million tokens in and $0.0923 out. The next lists $0.03 and $0.16. Both of those are 4-bit (fp4) hosts, which Mullet leaves out by default because 4-bit copies code noticeably worse (my judgement).

The cheapest 8-bit hosts, per million tokens in and out:

- $0.042 and $0.30
- $0.075 and $0.25
- $0.08 and $0.50
- $0.084 and $0.28
- $0.0885 and $0.295

The vendor's own host (Z.AI) is $0.15 and $0.50, and so are several other 8-bit hosts.

On its paid OpenRouter rungs Mullet also asks only for hosts that do not keep or train on prompts. The price list does not show that policy, so a host that looks cheap may not qualify. The free rung is the exception, which is why it only runs on public repositories.

A step, priced

My Strata run used 1.14 million tokens in and 30 thousand out. Input is about 38 times output (arithmetic), so the input price matters most.

At the cheapest 8-bit host a step that size comes to about $0.057 (1.14 times 0.042, plus 0.03 times 0.30). At the vendor's own prices it is about $0.19 (1.14 times 0.15, plus 0.03 times 0.50). So $0.06 to $0.19 across the 8-bit hosts listed. At the cheapest 4-bit host it would be about $0.035.

All of that is my arithmetic from listed prices on 1 November. It is not billed. I made no paid call, and prices change. It also prices every input token at the full rate, though the ledger counts 1,038,284 of those 1,141,486 as cache reads, and the listed hosts show a cache-read price that is lower at some of them, so a real bill could be lower.

A small bug

The same Strata run showed up in a Mullet report as "$2.52 estimated" for a model that costs me nothing per token. Mullet did not recognise the provider as local, so it priced the tokens at a fallback of $2 in and $8 out per million. The sum matches: 1.14 times 2 plus 0.03 times 8 is 2.52 (arithmetic).

Mullet 2.5.0 adds a setting, `ladder.local_providers`. Listing `strata` there makes a model from that provider count as the local rung, and local attempts are reported as costing nothing. The ledger rows for that earlier run still say $2.52.

What I want next

Mullet's ladder treats free and paid as two kinds of rung. At these prices the gap between them is about six to nineteen cents a step (my arithmetic above), while a free endpoint can cost something else: it may log prompts.

What I want is for the user to set a price ceiling, per million tokens, and for "free" to mean at or below it. Zero would mean free only. Budgets differ, so the ceiling should be a setting, not a decision made for everyone. It is not built yet.

The code is at https://github.com/tomgould/mullet-public (https://github.com/tomgould/mullet-public).