



AI · MARKETS

The cheapest model that passes

One week of frontier launches showed the top models converging on price. What's scarce now is access, distribution and your own tests.

BY MAGAZINER · OCTOBER 2026

IN SHORT

In late September, Google, OpenAI, Anthropic and xAI all shipped or priced frontier models, and several landed at nearly the same price. The most capable models were gated. The biggest consumer spike came from distribution, not a benchmark. The practical question has shifted from “which model is smartest?” to “which is the cheapest one that passes my tests?”

ON 1 OCTOBER, ACCORDING to third-party reporting, calling Google’s new Gemini 4 Argon on the public API returned “not found”. Google had announced it the day before, with a price list, but it was going first to vetted cyber defenders. That same week, four labs shipped or priced frontier models. Read the launch posts one after another and a pattern shows up. The prices bunch together. The most capable versions are locked away. And the biggest surge of users went to the company with the most distribution, not the best benchmark.

My argument is that the frontier has stopped being the scarce thing. When several labs sell similar capability at similar prices, the advantage moves to three places: who gets access to the gated models, who owns the screens people already use, and who has honest tests of their own work.

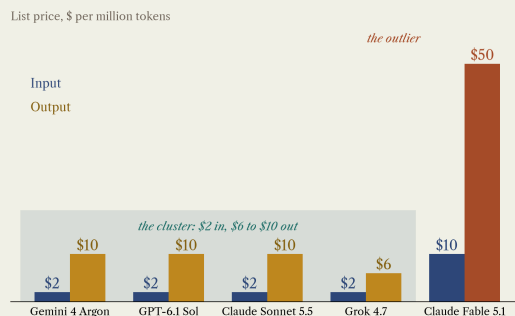


1

THE PRICES BUNCHED UP

Look at list prices per million tokens. Google’s Gemini 4 Argon launched at an introductory \$2 in and \$10 out, rising later to \$4 and \$20. OpenAI’s GPT-6.1 Sol is \$2 and \$10. Anthropic’s Claude Sonnet 5.5 is \$2 and \$10, and Anthropic says it uses fewer tokens per task. xAI’s Grok 4.7 is \$2 and \$6. Anthropic’s top model, Claude Fable 5.1, is the outlier at \$10 and \$50.

Four of these five frontier models list at about the same price



Source: company price lists, late September 2026. Argon at introductory price (later \$4 and \$20).

Each of these companies claims strong results on coding and professional work. Some claims are striking: Anthropic says Sonnet 5.5 jumped from 10.3 percent to 70.6 percent on Terminal-Bench 4.0. But they're mostly self-reported, and they're all aimed at the same shelf.

When the price is the same, the deciding question becomes whether a model clears your own test. That is a different buying habit from "get the smartest one."



2

THE BEST MODELS WERE GATED

The most capable models weren't on the public menu. Google shipped Argon first to trusted cyber defenders and said it was in the US government's voluntary pre-release process. Paid API customers come next, with no public date. Anthropic's Mythos, the same weights as Fable with looser safeguards for cyber and life sciences, is also limited to vetted programmes.

The reason is that cyber capability at this level cuts both ways. The consequence is that access has become a product feature. A roadmap that assumes a gated model next quarter needs a plan for getting access, or a design that works with the public ones.

Access has become a product feature.



3

DISTRIBUTION MADE THE SPIKE

The clearest surge of the week wasn't a model. It was Meta's Muse assistant. It launched in the US and Canada on 8 September and was number one on the US App Store by the 18th. Sensor Tower projected it past five million downloads in about 22 days, faster than ChatGPT, Grok or Claude reached that mark on Sensor Tower's numbers.

But Sensor Tower also says Muse took up to half of Meta's daily house-ad impressions for two weeks. Downloads jumped 73 percent in a day as Meta ramped those ads. Estimates differ between tracking firms, and chart position is not retention.

That's the point. Meta can put an agent in front of its users at a scale no app store ranking can match. "Rented reach, owned relationships" makes the same argument at length: reach rented from a platform can be turned up or down overnight.



4

THE PLATFORMS ARE SELLING WORKPLACES

OpenAI's DevDay showed what a lab looks like when it stops selling only answers. Always-on agents with their own cloud computers. Background coding. A shared workspace for people and agents. Plumbing so agents can live in Slack and Teams. Meta's Muse and Anthropic's Claude Code are versions of the same bet.

The fight is over where the agent lives, what it can touch and who approves irreversible actions. That is the scaffolding coding already had (see "Coding came with a judge"), now being built for everyone.



5

WHAT THIS MEANS FOR BUILDERS

Most teams should buy the model and build the harness: the tools, tests, memory, approval rules and human review. Put model choice behind an interface so you can switch. Keep a set of tests built from your real tasks, and change models when the numbers say so. Watch cache pricing as closely as list price, because agents reuse a lot of context.

I don't know which lab will lead next quarter. That's rather the point. If the frontier keeps converging, the durable advantages are the ones the labs can't ship for you: your own tests, your own data, and a clear rule for what an agent may do. ■

Sources: Google DeepMind's Gemini 4 Argon announcement (30 September 2026); OpenAI DevDay (29 September); Anthropic's Claude Fable 5.1 and Sonnet 5.5 releases; xAI's Grok 4.7 release; Sensor Tower and TechCrunch on Meta's Muse; Artificial Analysis. Benchmark claims are the companies' unless noted.

Illustrations and charts were made for this article. Set in Libre Caslon and Josefin Sans.