JULY 6, 2026

The AI price war, explained: what falling model prices mean for buyers

Model prices are collapsing because the labs are undercutting each other on purpose. Mostly good for buyers — if you avoid the trap underneath it.

By **Dave Taylor**

The AI price war means the models you run your business on are getting cheaper fast, and that shift is mostly working in your favour. Inference prices have fallen a median of roughly 50x per year — as fast as 200x in a single year for some models — per Epoch AI's tracking of LLM inference price trends. The labs are cutting prices on purpose to win your workload off each other. For a buyer, that means the same automation costs less every quarter you keep running it. The trap is assuming a cheaper model per token is a cheaper model per finished job. Sometimes it is. Often it isn't, and the difference is where the real money hides.

What's actually driving AI prices down?

Two forces are pushing model prices down at once, and only one of them is the obvious one. The obvious force is engineering — better chips, better serving, smaller models that match last year's big ones. That's the 50x-a-year median Epoch AI tracks, and it falls rapidly but unequally: the fastest models dropped roughly 900x in a year, the slowest around 9x. Prices don't fall on a smooth curve you can plan against. They fall in lurches, model by model, whenever a lab ships something more efficient.

The second force is competition, and it's the one that changes how you should buy. The labs aren't only passing on cheaper compute — they're deliberately undercutting each other to capture the workloads that will run for years. A model that costs less to run is worth more as a land-grab than as a margin, so the price you see isn't just what inference costs the lab. It's partly a bid for your business. That has a practical consequence: prices can move faster than the underlying economics alone would explain, and they can move because a competitor blinked rather than because anything technical changed. A cost curve you could forecast; a bidding war you can't. That is why the fall is so uneven — one lab drops a price to answer another, the whole market lurches, and the timing is set by strategy rather than silicon. You can't plan a budget around a number that changes when a rival ships.

The cheap model is the strategy now

The clearest signal that this is a price war, not just a cost curve, is that the labs are pricing their *workhorse* models to win, not their flagships. Anthropic launched Claude Sonnet 5 at a $2/$10 introductory rate per million tokens, input and output, running through 31 August 2026 — cheaper than its own Opus 4.8 at $5/$25 — and it explicitly positioned Sonnet 5 as the model meant to run your agents at scale, not the top-of-range showpiece. The standard price after the intro is $3/$15. Read that pricing as a message: the money is in the model you'll run a million times, so that's the one they'll fight over.

You can see the same shape on the other side of the market. OpenAI runs a ladder from GPT-5.5 at $5/$30 down to GPT-5.4 nano at $0.20/$1.25, per OpenAI's published pricing — a spread of more than twenty-five times between the top and bottom rungs for the same family. GPT-5.6 is in limited preview and not yet a released price point, so the live ladder is what matters. The ladder itself is the strategy: a capable-enough model on every rung, so no workload has a reason to leave for a competitor. A task that only needs the nano tier never reaches for a rival's cheap model, because there's one in-house at a fifth of the price.

When two labs both build their whole line-up around the cheap, high-volume tier, that isn't a coincidence of cost. It's a fight over who runs your agents for years — so a buyer who sees it stops shopping on flagship prices and starts shopping on the tier they'll actually run.

What the buyer gains — and the trap underneath it

The gain is real and it compounds. An automation you built last year on last year's model can be re-pointed at this year's cheaper one and do the same work for a fraction of the token cost. The savings aren't a one-off discount you negotiate once — they arrive on their own, every time the price war produces a new workhorse tier, for as long as you keep the automation running. If you built it to swap models cleanly, you bank the drop. If you hard-wired one model in, you leave that money on the table.

The trap is that a cheaper model is only cheaper if quality holds. Price per token is not price per finished result. A model that costs half as much per token but retries twice, reasons for longer, or needs a human to correct its output can cost you *more* per completed job than the pricier one it replaced. You pay in tokens you didn't expect and in staff time cleaning up. We covered the mechanics of that in why your AI token costs keep rising even as prices fall: the unit price drops while the token count per task climbs, and the bill goes up anyway. A "cheap" model that quietly does this is pseudo-cheap. It looks like a saving on the pricing page and lands as a loss on the invoice.

Don't build your workflow around one price

The mistake that costs the most is treating today's price as a fixed assumption you can design around. It isn't fixed. It's an input that moves — usually down, sometimes in a lurch, occasionally when a model you depend on gets deprecated and the migration is on you. Epoch AI's own read on the trend is that prices fall "rapidly but unequally," which is a polite way of saying you cannot predict which model will be cheapest six months out. If your workflow assumes one specific model at one specific price, every one of those moves is a rebuild instead of a swap.

The buyer-side fix is to build the model as a swappable part, not a foundation. That means your prompt, your evaluation set, and your integration don't care which model answers — you can point them at a new one, run your test set, and see whether quality holds before you cut over. We wrote the decision framework for this in how to choose an AI model for a business task: pick on cost per accepted result, not per token, and keep the choice reversible. Reversibility is the whole point. A model you can swap in an afternoon lets you treat every price cut as free money; one welded into your logic turns the same cut into a project you have to justify, staff, and schedule. When the next drop lands — and at 50x a

year, one always does — a firm built this way captures it before lunch. A firm that hard-wired one model re-tests for a fortnight, decides the disruption isn't worth it, and never saves at all.

What the price war means for an SMB

For a smaller firm, the price war is where AI automation finally makes plain financial sense — but only if you build to catch the falling prices instead of locking yourself to the current one. This is the frame we build to at CxD. We build AI automations for SMBs, and we treat model price as an input that moves, not a fixed cost baked into the design. The same automation — the same prompt, the same tests, the same wiring into your systems — can be re-pointed at a cheaper model as prices fall, and for a smaller firm running a workflow thousands of times a month, that re-point is where the savings actually land.

That matters more for an SMB than for a large enterprise, because an SMB doesn't have a platform team to re-engineer a workflow every time the market shifts. A big company can absorb the rebuild cost; a ten-person firm can't, so in practice it just doesn't rebuild. If the automation is built to swap models cleanly, a solo founder or a lean ops team captures a 50x-per-year price trend without touching the logic — the market does the cost-cutting for them. If it isn't, each price cut is a project nobody has time for, so the firm quietly keeps paying last year's rate for this year's work, month after month, and never notices the drift because the bill looks the same as it always did. The difference between those two outcomes isn't the model you pick today. It's whether the automation was designed to let you change your mind cheaply — which is the part the price war rewards, and the part most build-it-once automations skip.

Where this leaves you

The AI price war is good news for buyers, with one condition attached. Prices are falling fast and on purpose, the labs are fighting hardest over the everyday workhorse models rather than the flagships, and that fight puts money back in your pocket every quarter — but only if two things are true. Your automation has to be built to swap models cleanly, and you have to judge a model on cost per finished result, not cost per token, so a pseudo-cheap model doesn't quietly eat the saving. Get both right and the market works for you: the same automation gets cheaper on its own, indefinitely, without a rebuild. Get either wrong and you either keep paying last year's rate or you chase a cheaper price straight into a bigger bill. The single thing to take away is that the model is the moving part, and the money is in building so it can move.

If you're running AI automations — or planning one — and want them built to catch falling prices instead of paying this year's rate forever, book a 30-minute working session and we'll look at where your workflow is exposed to a single model's price.