CoreWeave

What inference pricing model is right for you?

Pay-per-token pricing makes consuming inference and comparing AI cloud providers easy—but it's not always the best way to think about your inference budget.

The token pricing illusion

Not all tokens are created equal. When evaluating the cost of inference, here are the questions you need to start with.

Did the token arrive in time?

p95 or p99 time-to-first-token and inter-token latency SLOs can't be ignored

Did it land in the right context?

Outside the proper session, agent loop, or batch window, it's not a useful token

How many times did I pay for it?

Not duplicated by a client retry, fallback model call, or provider-side restart

The metric that matters

PERFORMANCE-ADJUSTED COST PER USEFUL TOKEN = **TOTAL INFRASTRUCTURE COST OVER PERIOD** / **TOKENS SERVED WITHIN SLO DURING PERIOD**

Of course, for many use cases, pay-per-token is still the right call. Let's look at some sample workloads and see where token pricing still makes sense, and where teams would get better results from GPU-billed inference.

The right pricing for the right workload

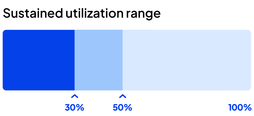
PATTERN A
Exploration and iteration

Best fit: Token pricing

You're paying for fast experimentation. Utilization would be too low for a dedicated setup.

Shape: Highly variable traffic with long idle stretches, evolving prompts, frequent model swaps. Sub-production volume, rarely above 10–20% sustained utilization.

PATTERN B
Variable production at moderate scale

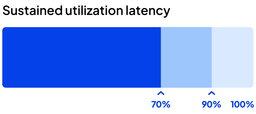


Best fit:
Depends on utilization → Run the diagnostic

If utilization is low, token pricing still holds up. If utilization is high enough that autoscaling lag starts generating SLO misses, performance-adjusted cost on dedicated capacity is more cost-effective.

Shape: Production traffic that swings hard, 3–5x peak-to-trough, with volume rising and SLOs tightening month over month. Sustained utilization in the 30–50% range and climbing.

PATTERN C
Steady-state production

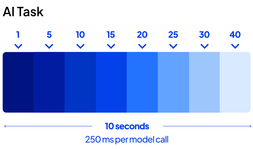


Best Fit: GPU-Billed

Your workload matches the cost model. You pay for GPU-seconds you actually use. At this utilization rate, dedicated capacity is often more economical on a performance-adjusted basis.

Shape: Predictable throughput that fills capacity and keeps it busy, RAG pipelines, classifiers, high-volume assistants, holding 70–90% sustained utilization under tight latency SLOs.

PATTERN D
Agentic at scale



Best Fit: Depends on volume → Move from token-pricing to GPU-pricing when ready

Serial calls compound latency. High-volume production agents benefit from dedicated capacity; exploratory agents still belong on token pricing. A mature agentic platform is almost always both, deliberately.

Shape: 10–50 model calls per task, which can drive far more tokens per session than a single-call estimate suggests. Serial latency dependencies: at 250ms per call, 40 serial calls add up to 10 seconds.

CoreWeave

The sticker price per million tokens isn't a one-to-one indicator of what production AI actually costs.

The performance-adjusted cost per useful token diagnostic helps close the gap. Read "The Token Pricing Illusion" to learn how you can find the right inference economics for your workload.

Read the blog