Unproven pricing power by leading AI providers

OUTPUT TOKENS

Flagship model vs. cheapest available tier, by vendor — $ per million output tokens (log scale)

> *"Pricing power at the frontier — not usage growth — is becoming the real scoreboard for whether AI is creating value or just consuming capital."* — **Luke Lango, Innovation Investor**

List API prices, $ per million output tokens (not quality-adjusted, not enterprise-negotiated rates). "Flagship" = each vendor's top reasoning/general model at that date; "cheapest tier" = lowest-cost mini/flash/haiku-class model available across the three vendors at that date. Anthropic's and Google's small-model prices have risen at points as capability increased — a reminder that quality-adjusted price, not sticker price, is the precise lens. Sources: OpenAI, Anthropic and Google pricing pages; TokenMix, Finout, CloudZero pricing trackers (Aug 2026).

Unproven pricing power by leading AI providers INPUT TOKENS

Flagship model vs. cheapest available tier, by vendor — $ per million **input** tokens (log scale)

> *"Pricing power at the frontier — not usage growth — is becoming the real scoreboard for whether AI is creating value or just consuming capital."* — **Luke Lango, Innovation Investor**

List API prices, $ per million **input** tokens — the companion view to the output-token chart (not quality-adjusted, not enterprise-negotiated rates). "Flagship" = each vendor's top reasoning/general model at that date; "cheapest tier" = lowest-cost mini/flash/haiku-class model available across the three vendors at that date. Input-token pricing compressed further at the floor (near $0.10/M) than output pricing did, while frontier input prices fell too — Anthropic cut its own Opus line 3x between Opus 4 and Opus 4.8. Sources: OpenAI, Anthropic and Google pricing pages; TokenMix, Finout, CloudZero pricing trackers (Aug 2026).

Input vs. output: which chart is your bill?

The two token-pricing charts track different things — and most buyers are drifting toward the one where frontier pricing power has held up hardest.

Your prompt
Instructions, documents, chat history, retrieved context

INPUT TOKENS
priced 1× · buyer controls volume

Model
Processes input in parallel; generates output one token at a time

OUTPUT TOKENS
priced 3-6× · model decides volume

Answer
Visible reply + hidden "reasoning" tokens you're billed for

Input-token chart

- **What it prices:** everything sent to the model
- **Who controls the volume:** the buyer — context length is a design choice
- **Why it's cheaper:** a single parallel pass over the prompt
- **Where it compressed most:** commodity floor near $0.10/M

Your bill tracks this chart if: you run document Q&A, RAG search, "chat with your data," or anything with long context and short answers.

Output-token chart

- **What it prices:** everything generated back, including reasoning
- **Who controls the volume:** the model — a cap helps, but reasoning models self-regulate less
- **Why it's pricier:** sequential, one token feeds the next
- **Where the premium held:** frontier tier never priced below $25/M

Your bill tracks this chart if: you run coding agents, long-form drafting, or "reasoning" / extended-thinking models.

The CFO-relevant twist: adoption is moving toward agents and reasoning models — which shift spend from input to output tokens. That's the tier where frontier pricing power has proven most durable, not least. Usage growth is concentrating exactly where compression has been weakest.

Companion slide to the input-token and output-token pricing-power charts. "Reasoning tokens" are generated but often hidden from the user; OpenAI, Anthropic and Google all bill them as standard output tokens.