AI API Pricing

Victor Dibia
Independent Researcher
USA
victor.dibia@gmail.com

Interactive version, with live figures: https://victordibia.com/explainers/ai-api-pricing/ (revision 2, last updated 2026-11-14).

1 The cheapest model at GPT-4's level got 300x cheaper. The most capable model got 3.8x cheaper.

Understanding how AI model pricing has evolved over time can be instructive. I first explored this in November 2023. OpenAI had just announced new pricing for GPT-3.5 Turbo: $1 per million input tokens. For context, this was a 20x decrease[7] compared to Davinci ($20) in March of that year.

If we revisit the data today (see chart below based on 202 models from 13 providers), the good news is that the price decrease trend still holds: GPT-4's capability cost $30.00 per million input tokens in March 2023 and $0.10 today, 300x less. But there is some clear separation.

There are now **two categories** of models - small but capable models, and large state of the art models. At the bottom are the mini and flash families. Of the 13 on sale with a score that was measured, 11 beat GPT-4, the best model of 2023, on the capability measures used here, at $0.10 to $1.00 per million input tokens. At the top are much bigger models trained on much more compute. The most capable one on sale costs $8.00 today, and averaged over each year, from GPT-4's launch in March 2023, it cost $26.18 in 2023, $14.14 in 2024, $7.02 in 2025 and $5.25 so far in 2026.

One interesting question we can answer from all of this is around if **"AI is getting cheaper"?**. The data suggests this is true but with some nuance. The small model regimes have gotten better and cheaper 300x and while state of the art or larger models have also gotten cheaper but by a much smaller factor 3.8x.**

The chart below draws two lines through the prices. The lower one follows **the cheapest model on sale at GPT-4's level or better**, where GPT-4's level is its score on the Artificial Analysis Intelligence Index[2], 6.7 against 62.3 for the most capable model today. The upper one follows **the most capable model on sale**.

Budget for the most capable model at today's price, and re-test cheaper models every few months.

- In November 2025 the most capable model on sale was GPT-5.1 at $1.25. Today it is Claude Fable 5.2 at $8.00, and a model that scores at least as high as GPT-5.1 costs $0.10 (GPT-6 Luna).
- The price of the most capable model on sale has changed 20 times since March 2023, each time because another model took the top score, and 12 of those changes were rises.
- If your workload mixes easy and hard requests, send each request to the cheapest model that can do it. Microsoft Foundry's model router and Amazon Bedrock's intelligent prompt routing are managed versions of this[10, 15].
- This changes when a provider launches its next top model at a lower price. Claude Opus 4.5 launched at $5.00 in November 2025, against $15.00 for the Opus before it, and Claude Fable 5.2 launched at $8.00 in November 2026, against $10.00 for the Fable before it.

Limits: The figures in this section are prices per million tokens. One study measures the cost of running a whole benchmark, and finds that a fixed score got 5x to 10x cheaper per year while running the best model got 3x to 18x more expensive per year[9].

2 Models within one index point of each other can differ 33x in price

If we take every model on sale today with a measured score on the index, 32 pairs of them score within one index point of each other. In the median pair the more expensive model costs 2.3x the cheaper one. In the widest pair it costs 33x, with Claude Opus 4.8 at $5.00 and GLM-5.3-Flash at $0.15.

Within five index points of the most capable model, Claude Fable 5.2 at $8.00, the cheapest model is Gemini 4 Carbon at $2.00, 4x less.

Start with the cheapest model that meets your capability bar, and move up only if it fails on your tasks.

- The best open-weight model on sale, Kimi K3.1 Pro, scores 50.7 against 62.3 for the most capable model, and costs $2.00.
- This changes if you need low latency, a specific tool-use format, data residency or an enterprise contract.

Limits: Open-weight models do not share one licence. DeepSeek ships under MIT, while the largest recent releases from Qwen and MiniMax carry their own licences with conditions on commercial use[5, 6, 11]. A downloadable model also has to be served by someone, and hosts price it differently. Llama 3.3 70B costs $1.04 per million tokens on Together, and the cheapest of the providers behind OpenRouter lists it at $0.10 in and $0.32 out[1, 14].

3 A model with a lower input price can cost more per request

You pay per request, and a request costs input tokens × input price + output tokens × output price. Across the models on sale, the output price runs from 1x to 8.3x the input price, with a median of 5x. Grok 4.7 and GPT-5.6 Terra both charge $2.00 per million input tokens, and a request with 8,000 input tokens and 4,000 output tokens costs 1.6x as much on GPT-5.6 Terra.

A reasoning model writes out its working before it answers, and that working is billed as output. OpenAI's guide says that reasoning tokens "still occupy space in the model's context window and are billed as output tokens"[13]. Across the 24 models whose token use Artificial Analysis charts, reasoning is a median 67% of output

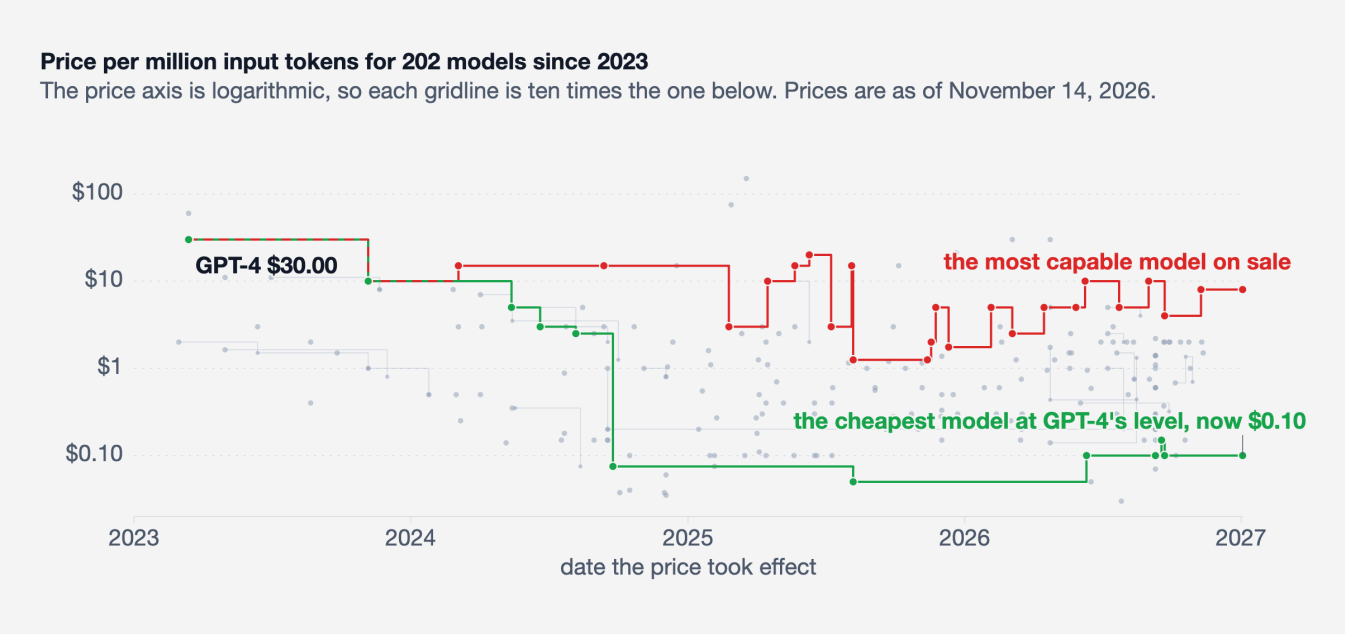

Figure 1: Each dot is one published price, and a short step joins a model's old price to its new one where it was repriced. The two lines are one line until March 2024, when Claude 3 Opus took the top score at $15.00, and they end 80x apart. The upper line dips wherever a cheaper model held the top score, and 19 of the 24 models that have held it have estimated scores. Shown: compare providers: Whole market; class: All; price of: Input tokens.

tokens, from 45% to 78%, and those models write a median of 62,200 output tokens per task on the index[2].

Past 2,700 output tokens the cheapest of the four models in the figure is Grok 4.7, not Gemini 4 Hydrogen. Gemini 4 Hydrogen charges $1.50 per million input tokens and Grok 4.7 charges $2.00.

Compare models by cost per request, using token counts you have measured.

- Run your own workload through each candidate model, log the billed input and output tokens, then multiply by each model's prices.
- A reasoning effort setting, which tells a model how much to reason, changes how many output tokens a request uses, so it changes the bill[13].
- This changes if your inputs are long and your outputs short. Retrieval and classification sit on the flat part of every curve, where the input price decides the cost.

4 Prompt caching cuts the cost of a 20-turn agent run by 66%

An agent is a program that calls a model in a loop, and it resends its whole context on every turn: the system prompt, the tool schemas and every turn before it. Input tokens therefore grow with the square of the number of turns while output grows in a straight line. In the run below, which is calculated from the list prices of Claude Sonnet 5.5, input is 78% of the bill by turn 20.

With prompt caching, the provider keeps the start of the prompt that it has already processed (the prefix) and charges a fraction of the input price to read it again, 5% of it on Claude Sonnet 5.5[3].

By turn 20 the cached run costs $0.22 against $0.64 without caching, a saving of 66%.

Keep the prompt prefix stable so it caches, and budget at the cached rate.

- Put everything static first (system prompt, tool schemas, examples), and never interpolate a timestamp, a request id or a shuffled tool order into it. One changed byte near the front invalidates everything after it[4].
- An estimate made from the list input price puts a 60-turn run at 5.5x its cached cost.
- This changes if your loop is slow or bursty. Cache entries expire (five minutes by default on Claude, with a one-hour option that costs more to write), so a run that idles pays full price again[4].

Limits: The run assumes that every turn reads from the cache and that the context only grows. An agent that prunes or summarises its context changes the prefix, so its later turns read less from the cache. Writing to the cache is priced differently by each provider: Anthropic charges 1.25x the input price for a five-minute entry, OpenAI charges 1.25x on GPT-5.6 and later and nothing on earlier models, and Google bills for the time that a cache you create explicitly is stored[3, 8, 12].

Every price has a date and a link to its source in the full table.

Figure 2: Each point is a model on sale, placed by its index score and its input price. A filled point has a measured score and a ring has an estimated one. Green points are open-weight models, whose weights can be downloaded, and blue points are models sold only through an API. Shown: as of: Nov 2026.

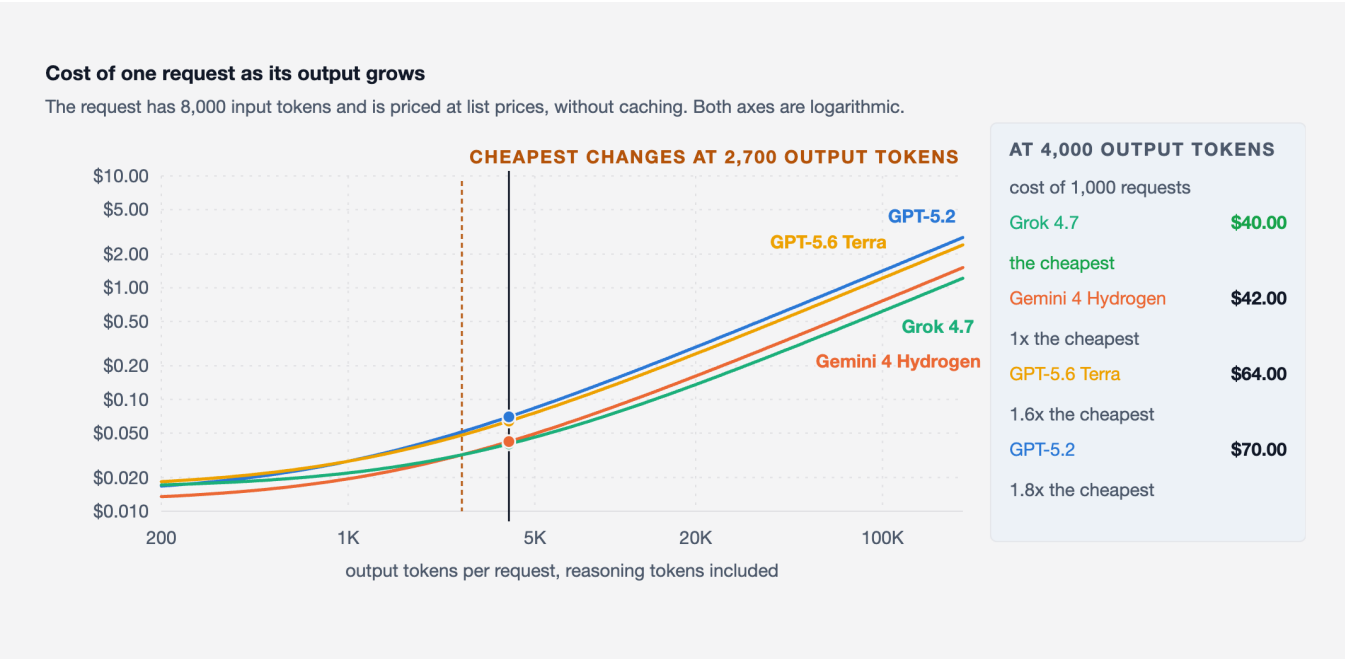

Figure 3: Each curve is one model. A curve is flat on the left, where the input cost dominates, and rises on the right, where the output cost does. Shown: output tokens per request: 4,000; input tokens per request: 8,000.

Figure 4: The solid part of each line is what the run has cost so far, and the faint part is where it is heading. Caching costs less from turn 2 on. Shown: turns completed: 20.

5 Scope and conventions

The dataset holds 208 models from 13 providers and 285 price points, as of November 14, 2026. Figure 1 draws the 202 that were released in 2023 or later and have a price above zero. Prices are list prices on the standard tier: the peak rate where a provider prices by time of day, the short-prompt rate where it prices by context length, and an introductory or promotional price while it applied (ten of the price points). Figures 1 to 3 leave out two lower rates: the batch rate, for requests that can wait, and the cached rate. The full table gives the cached rate.

The two lines of Figure 1 and the points of Figure 2 cover the models that have both a price and a score. Of the models on sale, twelve have a price and no score, from Qwen3.7-Flash at $0.03 to GPT-5.4 Pro and GPT-5.5 Pro at $30.00, and two are listed as free (GLM-4.5-Flash and GLM-4.7-Flash).

The lower line of Figure 1 reached $0.05 with GPT-5 nano. On June 11, 2026, OpenAI gave notice that it will shut down the model's only dated version on December 11, 2026. A deprecated model does not count as on sale, so the line rose on that day. Counting GPT-5 nano would make the fall 600x.

Figure 1 sorts models into three classes by name. A model is small if the provider's name for it contains the word mini, nano, lite, flash, haiku, small, micro or instant, or a size of 8B or less, or if it is OpenAI's Luna. It is large if the name contains the word pro, opus, max, large, premier, ultra, fable or mythos, or the size 405B, or if it is an o-series, GPT-4.5, Astra or Sol model. Every other model is mid.

The scores are from the Artificial Analysis Intelligence Index v4.3.2, read on one day, September 28, 2026, because the index is revised. Between readings on September 10, 2026 (v4.3) and September 28, 2026 (v4.3.2), 134 of 633 entries changed, where an entry is one model at one reasoning setting. Scores cover 157 of the 208 models, and 98 of those are estimates. Each model has one score, the highest that Artificial Analysis measured for it at any reasoning effort, or its estimate where nothing was measured. Artificial Analysis has since replaced that version with v5.0.1, which scores current models only, so GPT-4 has no score on it. The ten models released after September 28, 2026 have a score that this page estimated from their v5.0.1 score of November 14, 2026, using the 30 models measured on both versions. For a model left out of that fit, the estimate is off by about 2.4 index points. A model that has had several dated versions under one name has the score of the version served on each date.

The seven models from 2023 that have a score sit between 5.0 and 7.0 on the index, against 62.3 at the top, and all of those scores are estimates. On GPQA Diamond, 198 science questions that Artificial Analysis has run on models from 2023 and from 2026, GPT-4 scored 35%, and the 16 small models on sale that have been run on it score 36% to 95%.

References

[1] Together AI. 2026. Llama 3.3 70B on Together AI ($1.04 per 1M tokens, input and output) (accessed 2026-09-28). https://www.together.ai/models/llama-3-3-70b

[2] Artificial Analysis. 2026. Artificial Analysis Intelligence Index v4.3.2 (ten evaluations; a score that Artificial Analysis has estimated is labelled 'Estimate (independent evaluation forthcoming)') (read 2026-09-28). https://artificialanalysis.ai/models

[3] Anthropic. 2026. Pricing - Claude Platform Docs (cache read 0.05x the base input price on Claude Opus 5.5 and Claude Sonnet 5.5 and 0.1x on most other models, 5-minute cache write 1.25x) (accessed 2026-11-14). https://platform.claude.com/docs/en/about-claude/pricing

[4] Anthropic. 2026. Prompt caching - Claude Platform Docs ('Changes at each level invalidate that level and all subsequent levels'; 'By default, the cache has a 5-minute lifetime') (accessed 2026-09-29). https://platform.claude.com/docs/en/build-with-claude/prompt-caching

[5] Alibaba Cloud. 2026. Qwen3.8-2.4T-A95B licence file (Qwen3.8-Max License: a hosted-inference business above a revenue threshold 'shall obtain a separate license from Qwen' for commercial use) (accessed 2026-09-28). https://huggingface.co/Qwen/Qwen3.8-2.4T-A95B/blob/main/LICENSE

[6] DeepSeek. 2026. DeepSeek-V4.1-Flash model card (licence: MIT) (accessed 2026-09-28). https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash

[7] Victor Dibia. 2023. Top 5 AI announcements and implications (OpenAI DevDay, November 2023): the 'Cost Reduction' section, $0.02 to $0.001 per 1K input tokens between March and November 2023, a 20x decrease (accessed 2026-09-28). https://newsletter.victordibia.com/p/top-5-ai-announcements-and-implications

[8] Google. 2026. Context caching, Gemini API (explicit caches are billed for 'the amount of time cached tokens are stored') (accessed 2026-09-28). https://ai.google.dev/gemini-api/docs/generate-content/caching

[9] Gundlach, Lynch, Mertens, and Thompson. 2025. The Price of Progress: Price Performance and the Future of AI ('the price for a given level of benchmark performance has decreased remarkably fast, around 5x to 10x per year'; 'the price of running frontier models is rising between 3x to 18x per year') (arXiv preprint, accessed 2026-09-28). https://arxiv.org/abs/2511.23455

[10] Microsoft. 2026. Model router for Microsoft Foundry ('Smaller and cheaper models are used when they're sufficient for the task, but larger and more expensive models are available for more complex tasks') (accessed 2026-09-28). https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/model-router

[11] MiniMax. 2026. MiniMax-M3 licence file ('to deal in the Software for non-commercial purposes') (accessed 2026-09-28). https://huggingface.co/MiniMaxAI/MiniMax-M3/blob/main/LICENSE

[12] OpenAI. 2026. Prompt caching guide ('For GPT-5.6 and later, cache writes cost 1.25x the standard, uncached input-token rate') (accessed 2026-09-28). https://developers.openai.com/api/docs/guides/prompt-caching

[13] OpenAI. 2026. Reasoning models guide ('reasoning tokens … are billed as output tokens') (accessed 2026-09-28). https://developers.openai.com/api/docs/guides/reasoning

[14] OpenRouter. 2026. Meta: Llama 3.3 70B Instruct on OpenRouter ('Llama 3.3 70B Instruct costs $0.10/M input tokens and $0.32/M output tokens', the lowest of the listed providers) (accessed 2026-09-28). https://openrouter.ai/meta-llama/llama-3.3-70b-instruct

[15] Amazon Web Services. 2026. Amazon Bedrock Intelligent Prompt Routing (routes each prompt between two models of one family) (accessed 2026-09-28). https://aws.amazon.com/bedrock/intelligent-prompt-routing