STRATEGIC INTELLIGENCE DIVISION

A third off the price, 13% off the bill

Anthropic cut Sonnet from $3 and $15 per million tokens to $2 and $10, and its own tokenizer produces approximately 30% more tokens for the same text, so the bill falls about 13%. Sonnet 4.5 retires on 30 November 2026 on 61 days' notice, GPT-4o on 14 January 2027 on 92 days against a six month minimum carrying a stated exception, and both vendors withdrew custom temperature inside 48 days.

Annika Sorensen, Strategic Intelligence Division · 3 November 2026 · 11 minute read

In brief

- Anthropic's Sonnet rate fell a third between Claude Sonnet 4.5 and Claude Sonnet 5.5, from $3 and $15 per million tokens to $2 and $10, and its own tokenizer turns that into about 13% for the same text.
- The saving is smaller than the rate card says by a factor of 2.5, because the vendor states its tokenizer produces approximately 30% more tokens for the same text.
- claude-sonnet-4-5-20250929 retires on 30 November 2026 on 61 days' notice, and the five retirements before it ran between 60 and 62 days.
- OpenAI gives the four departing GPT-4 models 92 days and names one replacement priced 96% lower on input, so one migration can move the line by 13% or by 95% depending on whether the workload survives the move.
- Seventeen live Claude models carry one fixed retirement date between them, the rest carry a floor or an announcement rather than a date, and no vendor publishes a notice period as a contract term.

> *A new model can reject a deprecated parameter, and a new API version can remove it for every model.*

Anthropic, model deprecations page, read 3 November 2026

Two dates now sit inside the FY27 plan, and both were set by the supplier. claude-sonnet-4-5-20250929 is retired on 30 November 2026 on the 61 days of notice Anthropic gave on 30 September. GPT-4o, GPT-4o mini, GPT-4.1 and GPT-4.1 mini leave the OpenAI API on 14 January 2027 on 92 days of notice. Behind those two, Anthropic lists seventeen live Claude models and attaches a fixed retirement date to exactly one of them; the other sixteen carry a floor, written as not sooner than, apart from claude-mythos-preview, whose date is to be announced, and a floor is not a date a forecast can be built on.

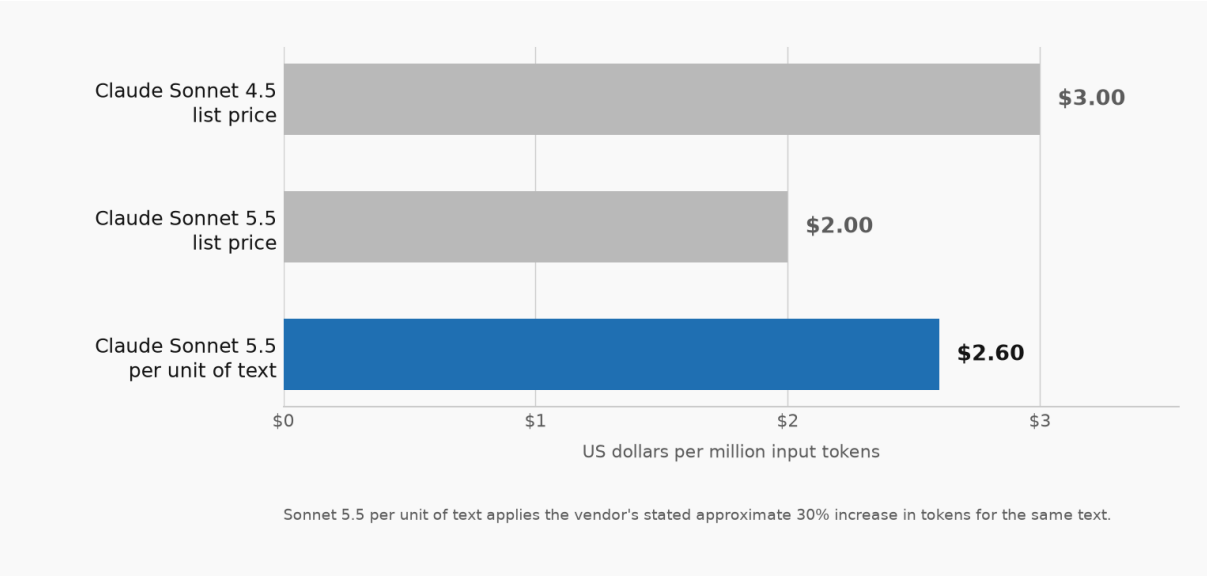
The CIO owns the migration. The CFO owns the number, and the number moved twice: the Sonnet rate fell a third on the published page, and what arrives on the invoice is about 13%. The General Counsel owns a clause that mostly does not exist, because no vendor publishes a notice period as a contract term. The head of model risk owns the part nobody has budgeted: an eval suite calibrated at temperature zero against a pinned snapshot now measures a model that will not accept the setting.

The rate card fell a third and the bill falls 13%

Anthropic's price table lists Claude Sonnet 4.5 at $3 per million input tokens and $15 per million output tokens. Claude Sonnet 5.5 is listed at $2 and $10. That is 33.3% off both sides of the meter. The tokenizer takes part of it back. The release note that launched Claude Sonnet 5 on 30 June 2026 says the model uses a new tokenizer that produces approximately 30% more tokens for the same text, and adds that the exact increase depends on the content and workload shape. A corpus that billed one million tokens on 4.5 bills 1.3 million on 5.5. At $2 a million that is $2.60, against $3.00 before. Output runs the same way: $13.00 against $15.00. The bill falls 13.3%, roughly two fifths of the third the rate card shows, and a vendor

handed a 30% token increase would have had to cut the per-token price by 23.1% to leave the customer where it started.

Input cost for the same text, Claude Sonnet 4.5 against Claude Sonnet 5.5



Source: Anthropic pricing page, read 3 November 2026; Anthropic Claude Platform release notes, 30 June 2026; Mosaic Singularity analysis, November 2026. The third bar applies the vendor's stated approximate 30% increase in tokens for the same text to the $2.00 list price. Output tokens behave identically at $15.00, $10.00 and $13.00.

Two qualifications belong in the same breath. The 30% is the vendor's approximation and not a measurement of any one workload, and a corpus of structured records will tokenize differently from a corpus of prose. And the rate card did not have to fall at all: the same pricing page records that Claude Sonnet 5 launched on 30 June 2026 at $2 and $10 as introductory pricing, that the figure became the standard price on 10 August 2026, and that the previously scheduled increase to $3 and $15 on 1 September 2026 will not occur.

The other side of the same reprice runs 95% the other way

OpenAI's deprecation table names one recommended replacement, gpt-6-luna, for all four GPT-4 models leaving on 14 January 2027. The pricing page lists gpt-6-luna at $0.10 input and $0.50 output per million tokens, against GPT-4o at $2.50 and $10.00 and GPT-4.1 at $2.00 and $8.00. That is 96% off input and 95% off output, conditional on the workload surviving a move to a different model. Source: OpenAI deprecations page and pricing page, read 3 November 2026; Mosaic Singularity analysis, November 2026.

Both vendors took the temperature dial inside 48 days

OpenAI's changelog entry of 3 September 2026 for GPT-6 Astra states that the model does not support custom temperature or top_p values or log probabilities. Anthropic's API version 2026-10-21, released on 21 October 2026, does the same at the API level: the parameter table on its deprecations page records temperature, top_p and top_k as removed in that version, where they return a 400 error. Forty eight days apart, both vendors withdrew the setting an enterprise uses to make a model's output repeatable. It is a 400 and not a warning, so a pipeline built at temperature zero stops rather than drifts.

The pin does not save it, and this is the part that will be discovered in production. On Claude 4.7 and later models, and on Claude Mythos Preview, the same page says earlier API versions return the error when the parameters are set to a non-default value. Pinning the API version therefore preserves the parameter on

older models and does nothing on the newer ones. The model decides, not the version header, and the model is the vendor's.

30 November 2026

claude-sonnet-4-5-20250929 retires 61 days after the notice of 30 September 2026, and the five consecutive Anthropic notice periods ending with it ran between 60 and 62 days: 62 days from 19 December 2025, 60 from 19 February 2026, 62 from 14 April 2026, 61 from 5 June 2026 and 61 from 30 September 2026. Source: Anthropic model deprecations page, read 3 November 2026; Mosaic Singularity analysis, November 2026.

Notice runs from 60 days to 16 months, and none of it is a contract term

What each vendor has published, and what it actually gave

Vendor	Models	Announced or released	Retired	Interval	Published notice policy
Anthropic	claude-sonnet-4-5-20250929	30 September 2026	30 November 2026	61 days	At least 60 days for publicly released models, stated in the deprecations documentation rather than in a contract term
OpenAI	GPT-4o, GPT-4o mini, GPT-4.1, GPT-4.1 mini	14 October 2026	14 January 2027	92 days	At least 6 months for generally available models, unless safety or compliance concerns require a faster timeline
OpenAI	GPT-5.3-Codex, GPT-5.1, GPT-5.4-Nano	1 October 2026	1 April 2027	6 months, 182 days	At least 6 months for generally available models, at least 3 months for specialised variants
Google Cloud	gemini-2.5-pro	17 June 2025, release date	20 October 2026	16 months	At least 12 months after release, and not moved to an earlier date than what is listed

Source: Anthropic model deprecations page; OpenAI deprecations page; Google Cloud model versions page; all read 3 November 2026. Intervals are computed from the dates in each row by Mosaic Singularity analysis, November 2026.

OpenAI publishes the most specific notice policy of the three, and it is written as a policy rather than a term. The page states minimum notice of at least six months for generally available models, three months for specialised variants of them and as little as two weeks for preview models, and it qualifies the whole list twice: "Unless safety or compliance concerns require a faster timeline", and, where they do, "we will provide as much notice as reasonably possible." The batch announced on 1 October 2026 was given the six months the policy states, retiring 1 April 2027. The batch announced on 14 October 2026, four models that are neither previews nor specialised variants under the page's own definitions, was given 92 days, retiring 14 January 2027, and the entry records no reason and names no limb. That is the whole of the protection: a published minimum, an exception the vendor applies, and no statement of whether the exception is what produced 92 days. The one route past a shutdown date the page does describe is a commercial conversation, because it says only that "In some cases, developers may be able to provision dedicated capacity for continued access after a model's shutdown date." Google forecloses one direction in a single line, that retirement timelines may be extended but "won't be moved to an earlier date than what is listed", and gemini-2.5-pro ran from 17 June 2025 to 20 October 2026. Anthropic's floor is the shortest of the three and the narrowest: "at least 60 days' notice before model retirement for publicly released models", to customers with active deployments, stated in the deprecations documentation rather than in a contract term. The same page adds that partner platforms such as Amazon Bedrock and Claude on Google Cloud managed by Google set their own retirement schedules, so the same model can carry a different date by

channel.

The saving is real and the buyer cannot yet commit to it

Take the two directions together and the CFO's problem is not the price. It is that the same line item can move by 13% or by 95% in one migration, and which happens depends on whether the workload can run on the model the vendor named. A summarisation job that survives the move to gpt-6-luna costs 96% less per input token. A job that does not survive costs a re-architecture, and 92 days is the whole of the notice in which to find out which one it is.

The re-validation is where the money goes, and no vendor page helps with it. An eval suite built at temperature zero on a pinned snapshot has to be rebuilt, because the parameter is refused on the current generation and the snapshot has a retirement date. In a regulated firm that is model risk evidence rather than a test run: the validation record names a version, and the version has an expiry. The enterprise that could evidence a controlled system now has to evidence it against a model it does not control, on a schedule it did not set, and it has to do so with the sampling parameters that made the original measurement reproducible removed from the interface.

Two obligations land on 30 November 2026, two more on 14 and 31 January 2027

The retirement date is a deadline for two of the four seats, not the finish line. The CFO and the CIO owe theirs on 30 November 2026, the day claude-sonnet-4-5-20250929 retires; the General Counsel owes on 14 January 2027 and the head of model risk on 31 January 2027.

- CFO: re-forecast the FY27 AI line on cost per unit of text rather than cost per token, and apply the vendor's stated token increase to every Sonnet 4.5 workload before crediting the price cut. The published third is 13% on Anthropic's own figures, and a 2.5 times error in the wrong direction is a budget that looks funded and is not. By 30 November 2026.
- CIO: inventory every call that sets temperature, top_p or top_k and every pipeline pinned to claude-sonnet-4-5-20250929, and move both before the retirement date rather than after the first 400. The parameters are already refused on Claude 4.7 and later whatever API version is sent. By 30 November 2026.
- General Counsel: put a notice term into the next model services agreement, with the notice period, the migration support and the treatment of a tokenizer change written into the schedule. OpenAI's published policy carries a safety and compliance exception and Anthropic's 60 days sits in documentation rather than in a contract, so neither page is the protection. By 14 January 2027.
- Head of model risk: restate the validation evidence so that it names the model version, the retirement date of that version and the sampling settings that were available when the test ran. An eval run that cannot be reproduced on demand is not evidence, and after 30 November 2026 the original cannot be rerun. By 31 January 2027.

What a board should put to its management

By 30 November 2026 the board resolves the first question below against an artefact, not a presentation: the register of production models carrying a fixed retirement date, the seat accountable for each migration, and the date each migration is funded in the FY27 plan.

- Which models in our production estate have a fixed retirement date, and which are running on a floor that we have mistaken for a date?
- What does the FY27 AI budget assume about the price per unit of work, and on whose tokenizer was that assumption measured?
- If every workload we run on GPT-4o has to move to the replacement OpenAI named, which of them have been re-validated, and what happens to the others on 14 January 2027?
- Which of our contracts would survive a supplier withdrawing a parameter that our controls depend on, and who has read that clause?
- On which date did we last prove that a model in production behaves as our validation record says it does?

What to watch, and when

- 30 November 2026: claude-sonnet-4-5-20250929 retires. An extension granted to any customer would be the first public evidence that these dates move for the people paying.
- 14 January 2027: the four GPT-4 models leave the OpenAI API. An extension granted before that date would show the shutdown date was not final, and the page would still not say whether the six month minimum or the safety and compliance exception produced 92 days.
- Before 31 January 2027: whether Anthropic announces the retirement of claude-haiku-4-5-20251001, whose floor of 15 October 2026 has passed, or claude-opus-4-5-20251101, whose floor is 24 November 2026. Neither has a date.
- By 30 June 2027: whether any frontier vendor publishes a notice period as a contract term with migration support behind it. If one does, the reading that control has moved to the vendor fails, and this piece should be retired with it.

Primary documents

- Anthropic, Claude Platform release notes, entries of 9 June 2026, 30 June 2026, 28 September 2026, 30 September 2026 and 21 October 2026.
- Anthropic, model deprecations, model status table, parameter deprecation table and deprecation history.
- Anthropic, pricing, model pricing table and footnotes.
- OpenAI, deprecations, notice policy, 1 October 2026 and 14 October 2026 entries.
- OpenAI, changelog, 3 September 2026 entry on GPT-6 Astra.
- OpenAI, pricing, standard tier table, and the GPT-6 Luna model page.
- Google Cloud, model versions and lifecycle, model tables and retirement dates.

Sources of the figures

- Anthropic pricing page, read 3 November 2026; Anthropic Claude Platform release notes, 30 June 2026; Mosaic Singularity analysis, November 2026. The third bar applies the vendor's stated approximate 30% increase in tokens for the same text to the $2.00 list price. Output tokens behave identically at $15.00, $10.00 and $13.00.
- OpenAI deprecations page and pricing page, read 3 November 2026; Mosaic Singularity analysis, November 2026.
- Anthropic model deprecations page, read 3 November 2026; Mosaic Singularity analysis, November 2026.

- Anthropic model deprecations page; OpenAI deprecations page; Google Cloud model versions page; all read 3 November 2026. Intervals are computed from the dates in each row by Mosaic Singularity analysis, November 2026.