A Price-Change Study of LLM API Pricing: Proprietary and Open-Source Models in PricePerToken, 2025-2026

IFX Research
Research note based on the PricePerToken public site and API
Trade the cost of intelligence.

Data retrieved: 2026-07-07 22:20 UTC

Abstract

Prepared by IFX Research, this paper studies LLM API price changes using only PricePerToken data: the current pricing catalog and the provider pricing history endpoint used by the public pricing-history page. The dataset contains 111,326 daily observations, 591 provider-model histories, and 2,108 detected metric-level price changes from 2025-07-28 through 2026-07-07. The main result is plain: pricing did not move as one market-wide downward curve. Open-source and open-weight families changed more often and with wider dispersion. Proprietary families changed less often, but single changes in expensive output prices can matter more for production budgets.

1 Introduction

API price is now a model property in its own right. It sits next to context length, latency, quality, and tool support when teams choose a model for agents, coding, search, or long-context analysis. A current price table answers one question: what does this model cost today? A price history answers a different question: how stable is that number when the budget is planned for the next month or quarter?

This paper asks three questions. First, how often do API prices change in the PricePerToken history data? Second, do proprietary and open-source models show different pricing behavior? Third, what can be seen in the most visible model families in the dataset: OpenAI GPT/o, Anthropic Claude, Google Gemini and Gemma, DeepSeek, Meta Llama, Qwen, Mistral, xAI Grok, Moonshot Kimi, Cohere Command, MiniMax, and Z-AI GLM?

The word “popular” is used in a narrow way. PricePerToken does not publish a market-share ranking in the endpoints used here, so the paper treats popular families as the large, visible model families represented in the PricePerToken catalog and history data. The paper does not add outside traffic, revenue, or usage estimates.

2 Data and Method

The source is limited to PricePerToken. The current catalog was loaded from the PricePerToken pricing endpoint [2]. The history was loaded from the provider pricing history endpoint [3], using the `provider` query parameter for every provider slug present in the catalog. The public pricing-history page [1] was used to identify the endpoint and the data fields used by the chart.

The history endpoint returns daily rows, not an event log. I reconstructed events by sorting each `provider + model` series by date and comparing adjacent rows across four fields: input price, output price, cache read price, and cache write price. A changed field becomes one metric-level event. If input and output change on the same day for the same model, that is counted as two events because the buyer sees two separate unit prices. Raw history prices are USD per token; all displayed prices use USD per 1M tokens.

The proprietary/open-source label comes from the `is_open` field in the current PricePerToken catalog. If a historical model cannot be matched to the current catalog, or if `is_open` is empty, it is labeled Unknown. Unknown models remain in coverage and event totals, but they are excluded from the binary proprietary versus open-source claims.

Table 1: Data coverage

Measure	Value
Catalog endpoint rows	574
Catalog providers	68
History endpoint rows	111,326
History provider slugs	58
Provider-model history groups	591
Detected metric-level changes	2,108
History date range	2025-07-28 to 2026-07-07

3 Proprietary Versus Open-Source Models

The binary comparison covers 358 historical provider-model pairs with a PricePerToken catalog match. The unmatched group is large enough to keep visible: it has 233 provider-model histories. Treating those rows as open or closed by hand would make the result look cleaner, but less faithful to the API.

Table 2: Price-change events by model class

Class	Models	Events	Increases	Decreases	Intro/removal	Median abs change
Proprietary	161	256	66	129	61	35.0%
Open-source	197	1,635	724	734	177	25.0%
Unknown	233	217	77	110	30	25.0%

Open-source models account for far more metric-level events than proprietary models: 1,635 versus 256. Among numeric changes, decreases make up 50.3% of open-source events and 66.2% of proprietary events. This should not be read as a simple claim that open-source models always get cheaper. The open-source group has many more variants and providers, so it produces more price observations and more revisions.

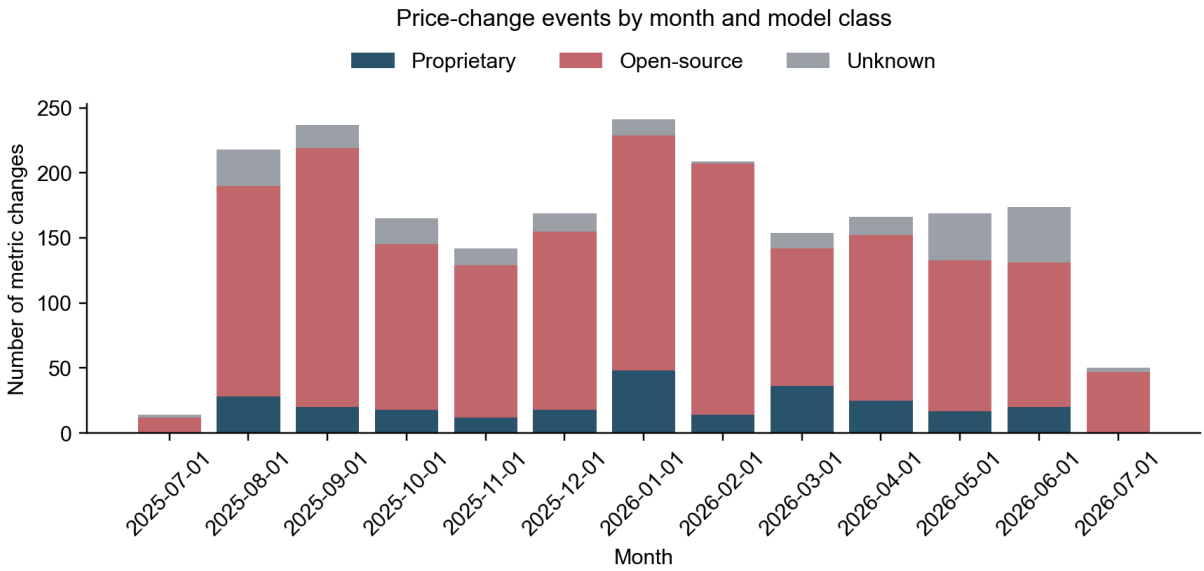

Figure 1: Monthly metric-level price changes by model class. The bars show when PricePerToken observed a new value in the daily history series.

The first-to-last view is calmer than the event count. Many model-metric pairs end at the same unit price at which they started. The median input net change is +0.0% for proprietary models and +0.0% for open-source models. The median output net change is +0.0% and +0.0%. The mean is less useful here because small absolute changes in cheap models can be huge in percentage terms.

Table 3: First-to-last net change by model-metric pair

Class	Metric	Pairs	Cheaper	Flat	Costlier	Median
Proprietary	Input	160	13.1%	81.2%	5.6%	+0.0%
Proprietary	Output	160	14.4%	80.6%	5.0%	+0.0%
Open-source	Input	196	24.0%	43.4%	32.7%	+0.0%
Open-source	Output	196	19.4%	45.9%	34.7%	+0.0%

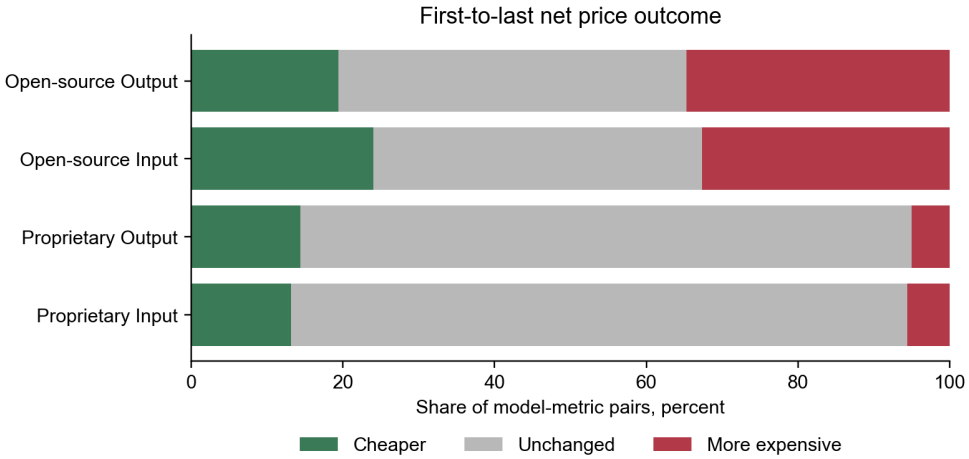

Figure 2: First-to-last direction of price movement. Flat means the last observed price equals the first observed price.

4 Input, Output, and Cache Pricing

The data show why a single “price per token” number is no longer enough. Input and output still carry most events, but cache read and cache write prices also change. A model with stable headline input and output prices can still become cheaper or more expensive for workloads that reuse long prompts.

Table 4: Metric-level event counts by price field

Metric	Events	Inc.	Dec.	Introduced	Removed
Input	809	378	431	0	0
Output	743	377	366	0	0
Cache read	432	109	158	114	51
Cache write	124	3	18	69	34

Cache pricing is a separate economic layer. Retrieval-heavy and agentic workloads can pay a different effective price from short chat workloads even when they use the same model. PricePerToken exposes that split directly: cache read and cache write prices move independently from input and output.

5 Popular Model Families

The most active popular families by event count are Qwen (549), DeepSeek (263), Z-AI GLM (222), Moonshot Kimi (161), Meta Llama (120). Across all families in the corpus, the single most active family is Qwen (549 events). The busiest month is 2026-01 (241 events).

Table 5: Popular families: activity and median first-to-last net change

Family	Models	Events	Inc.	Dec.	Median input net	Median output net
Qwen	54	549	219	254	-30.4%	-20.0%
DeepSeek	18	263	110	127	+0.0%	+0.0%
Z-AI GLM	12	222	77	127	-7.1%	+0.0%
Moonshot Kimi	9	161	70	87	+0.0%	+0.0%
Meta Llama	23	120	77	26	+100.0%	+42.9%
Mistral	26	105	56	21	+0.0%	+0.0%
Google Gemma	8	97	43	45	+33.3%	+104.2%
MiniMax	6	91	31	54	+0.0%	+0.0%
OpenAI GPT/o	7	85	24	54	+0.0%	+0.0%
Google Gemini	17	41	4	17	+0.0%	+0.0%
Anthropic Claude	18	26	3	0	+0.0%	+0.0%
xAI Grok	3	7	2	4	+0.0%	+0.0%
Cohere Command	1	6	4	2	+0.0%	+0.0%

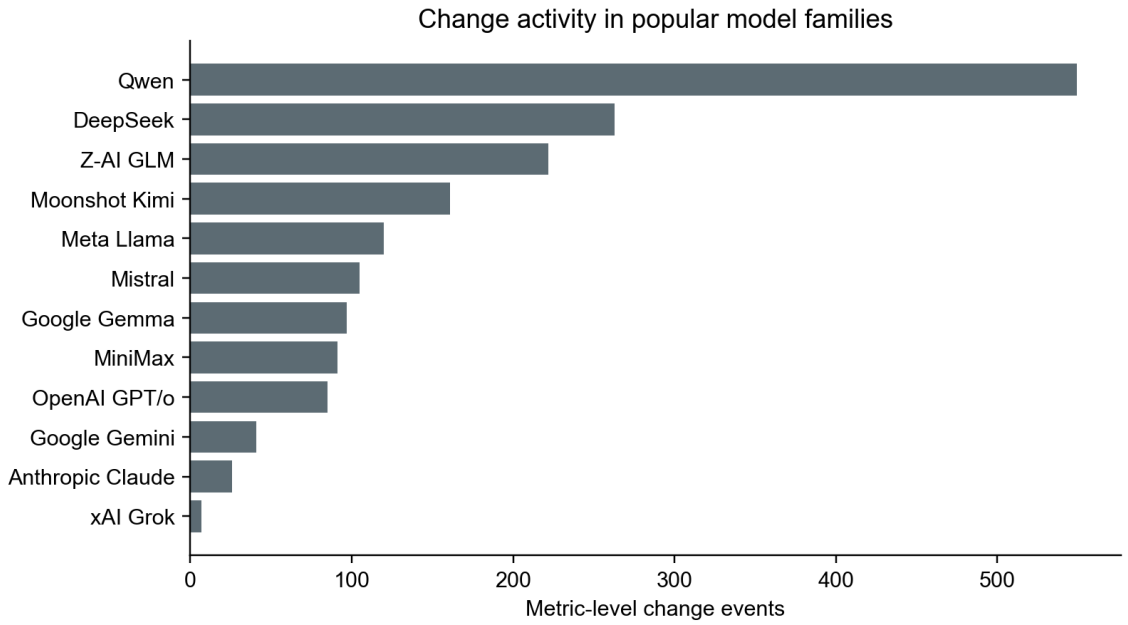

Figure 3: Price-change activity in popular model families. One event is one changed metric, so a same-day input and output change counts twice.

Qwen and DeepSeek produce many events because their families are broad: base, instruct, reasoning, coder, VL, and several size tiers. Mistral also spans open and closed models, which is why the analysis uses PricePerToken's `is_open` field at the model level instead of assigning a whole brand to one class. Anthropic Claude has fewer numeric changes in this window; its history is more stepwise, with price movement tied to model variants and cache terms. OpenAI has fewer events than Qwen, but the active rows are commercially important: GPT-5 variants, GPT-OSS, image models, audio models, and codex/chat variants.

Table 6: Most active models inside popular families

Family	Model	Class	Events	Inc.	Dec.
OpenAI GPT/o	gpt-oss-120b	Open-source	43	11	29
Moonshot Kimi	kimi-k2.5	Open-source	40	14	26
DeepSeek	deepseek-v3.2	Open-source	35	9	21
Moonshot Kimi	kimi-k2.6	Open-source	34	12	22
DeepSeek	deepseek-chat-v3-0324	Open-source	33	15	14
Z-AI GLM	glm-5.1	Open-source	32	11	21
Z-AI GLM	glm-4.6	Open-source	30	11	15
DeepSeek	deepseek-r1-0528	Proprietary	28	18	6
Qwen	qwen2.5-vl-72b-instruct	Open-source	28	15	10
Google Gemma	gemma-3-27b-it	Open-source	27	8	14
Z-AI GLM	glm-4.5	Open-source	27	12	13
Z-AI GLM	glm-4.7	Open-source	27	11	10

6 First and Last Observed Prices

Figure 4 plots each input or output model-metric pair by its first and last observed price. Points below the dotted line became cheaper; points above it became more expensive. The log scale is needed because the corpus mixes sub-dollar open-weight APIs with expensive frontier and reasoning models.

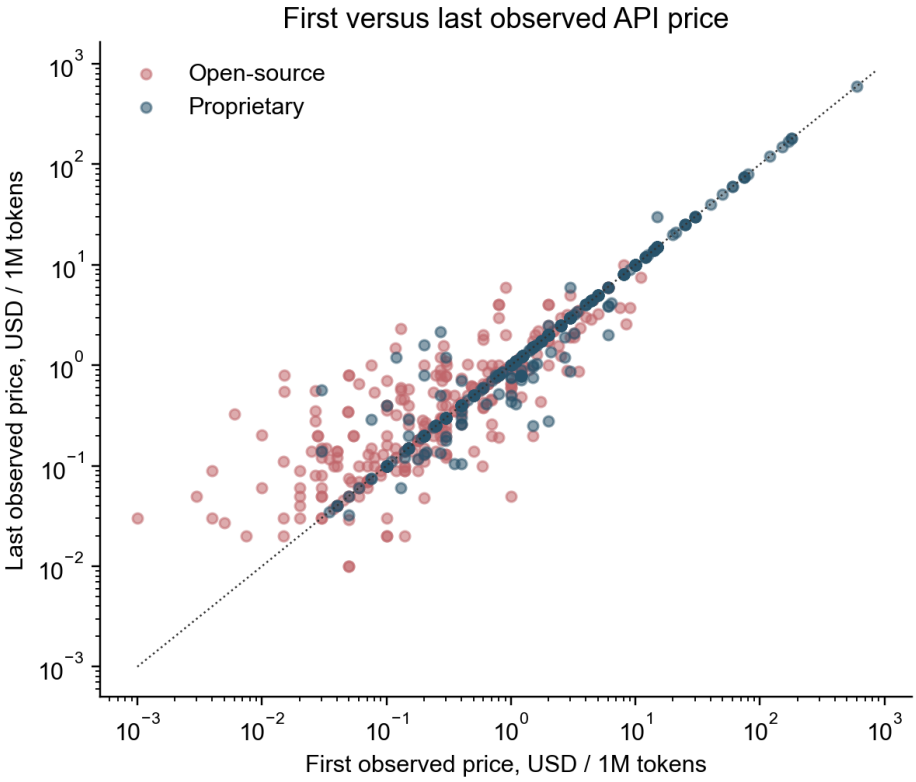

Figure 4: First and last observed API price, USD per 1M tokens.

The pattern is uneven. Cheap open-source APIs often show large percentage moves on small dollar bases. A move from 0.10 to 0.05 USD per 1M tokens is a 50% cut, but the budget effect is small unless volume is huge. A smaller percentage change in a proprietary output price can matter more for workloads that generate many tokens. Percent change and dollar exposure need to be read together.

7 Separate Comparisons

7.1 OpenAI, Anthropic Claude, and Google Gemini

The closed frontier families behave like managed product lines. Prices usually sit still within a variant and then move in steps when a new variant or cache term appears. OpenAI has the richest history in this group because the catalog contains many GPT/o variants: GPT-4.x, GPT-5.x, mini, nano, pro, image, audio, codex, and chat aliases. Claude has fewer changing rows, but cache read/write pricing matters. Gemini and Gemma should be separated: Gemini is the closed Google API line, while Gemma is treated as open-source where PricePerToken marks it as open.

7.2 Meta Llama, Qwen, DeepSeek, and Mistral

The open-source and open-weight families show more price discovery. Qwen has the highest event count among popular families in this dataset, followed by DeepSeek and Z-AI GLM. These counts reflect many model variants and pricing tiers, not one model changing every day. Meta Llama shows more increases than decreases in event count, while Qwen and DeepSeek have more decreases than increases. Mistral is mixed, so the brand name alone is a weak classifier.

7.3 xAI Grok, Moonshot Kimi, Cohere Command, MiniMax, and Z-AI GLM

This group sits between the broad open-weight families and the more stable closed frontier lines. Z-AI GLM is highly active in this window, with more decreases than increases. Moonshot Kimi has many events concentrated in a smaller set of models. MiniMax also shows repeated revisions, often across input, output, and cache fields together. Cohere Command has fewer model rows in the active set, so its table result should be read as a small-sample signal.

8 Limitations

The study is intentionally narrow. It uses PricePerToken only. It does not add external usage share, vendor announcements, private contracts, or cloud-provider discounts. The popular-family list is therefore a study frame, not a claim about total market usage.

There are also data limits. The history endpoint is daily. If a model changed price twice inside one day and ended where it started, this method would not see the round trip. The open/proprietary label is taken from the current catalog, so models that disappeared from the catalog remain Unknown rather than being classified by hand. Finally, PricePerToken reports observable public API prices; negotiated enterprise prices are outside the data.

9 Conclusion

The PricePerToken history from 2025-07-28 to 2026-07-07 does not support a one-line story that all LLM API prices simply fall over time. It shows three pricing regimes.

First, proprietary frontier families tend to have fewer price changes. When they move, the dollar effect can still be large because output prices are often higher. Second, open-source and open-weight families change more often. The higher event count comes from many variants and provider rows, and from active repricing in families such as Qwen, DeepSeek, Mistral, Llama, Kimi, MiniMax, and Z-AI GLM. Third, cache pricing is now part of model economics. Input and output prices alone miss a real part of the bill for long-context and agent workloads.

For buyers, the practical rule is simple: record price history, not only current price. Current price is enough for a one-off comparison. Price history is needed for a budget, a routing policy, or a model-selection process that has to survive more than one billing cycle.

References

[1] PricePerToken. LLM API Pricing History and Trends. https://pricepertoken.com/pricing-history.

[2] PricePerToken API. Current pricing catalog. https://api.pricepertoken.com/api/pricing/?show_all=true.

[3] PricePerToken API. Provider pricing history endpoint. https://api.pricepertoken.com/api/provider-pricing-history/?provider=openai.