STRATEGIC INTELLIGENCE DIVISION

The price list runs on Beijing time

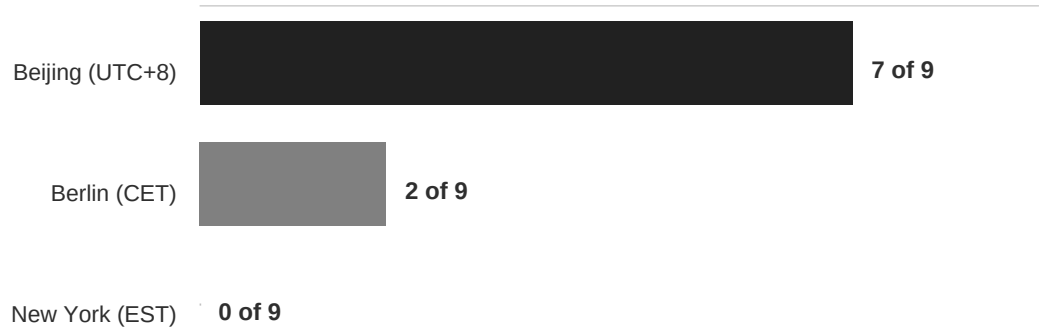
DeepSeek has priced by its own working day since 16 August 2026: peak is 09:00 to 12:00 and 14:00 to 18:00 in Beijing, and off-peak is exactly half. A nine-hour business day is seven peak hours in Beijing, two in Berlin and none in New York. Pro cached input is 12.14 times the flat rate it replaced at peak; output is 4.55 times.

Annika Sorensen, Strategic Intelligence Division · 4 November 2026 · 8 minute read

In brief

- The rate follows the caller's longitude: a nine-hour business day is 7 hours inside DeepSeek's peak in Beijing, 2 in Berlin and 0 in New York, and off-peak is exactly half on all six price lines.
- The line a replayed long prompt is billed on rose hardest, so V4 Pro cached input is 12.14 times its flat rate at peak while output is 4.55 times, and one blended multiple understates the cache line by about 2.67 times.
- The cheapest escape is a European open-weight model that undercuts DeepSeek by 47% at its peak on input and output, is about 6% dearer off-peak and 59% dearer on cached input, so it hedges the clock without winning on price.
- The licence that makes self-hosting legal carries one condition, triggered by operating a Model-as-a-Service offering, so the self-host decision is a legal question before it is an engineering one.

Hours of a nine-hour working day that fall inside DeepSeek's peak window



Hours in peak, of nine

Mosaic Singularity analysis, November 2026

Source: DeepSeek Models & Pricing, footnote 2, read 4 November 2026: peak hours are 01:00 - 04:00 and 06:00 - 10:00 UTC, Monday through Friday. Converted here to a 09:00 to 18:00 business day in each city, counting each hour that begins inside a peak window. Offsets: UTC+8 for Beijing; CET from 25 October 2026; EST from 1 November 2026. Mosaic Singularity analysis, November 2026.

A million input tokens of DeepSeek V4 Pro cost $1.32 at 10:00 in Beijing and $0.66 at 10:00 in New York, and a million output tokens cost $3.96 and $1.98. Same model, same day, same published list. The rate depends on where the caller is standing, because the peak window is the supplier's own working day and the buyer's clock is the one that moves.

The clock is the supplier's, and it is set to its own working day

DeepSeek's API changelog, in the entry dated 13 August 2026, announced the split and gave the reason: “To allocate resources more reasonably, we will adopt peak/off-peak pricing, with off-peak prices set at half of the peak-hour prices, encouraging users to schedule their tasks based on actual usage.” The new prices took effect at 16:00 UTC on 16 August 2026. The peak window runs 01:00 to 04:00 and 06:00 to 10:00 UTC on weekdays, which is 09:00 to 12:00 and 14:00 to 18:00 in Beijing, with the lunch hour at half price between them.

> *“Off-peak rates are half of the peak rates. Peak hours are 01:00 - 04:00 and 06:00 - 10:00 UTC, Monday through Friday, excluding Chinese public holidays. All other hours are off-peak, including weekends and Chinese public holidays in full.”*
>
> DeepSeek Models & Pricing, footnote 2, read 4 November 2026

DeepSeek's price list has been narrowed twice since the split took effect. The Models & Pricing page of 18 August 2026 lists the two windows and stops there, so a Saturday morning was peak. The page of 10 September 2026 carries Monday through Friday. The page of 4 November 2026 adds the exclusion of Chinese public holidays, in full. Each revision carved out time that is not worked in the supplier's home market, the weekend first and then the holidays, and neither changed the hours of a Beijing business day. That is the calendar the price now runs on.

Cached input, the line long prompts are billed on, rose 12.14 times

The page carries six price lines, three for each model. Cached input is the cheapest per token, and it is the line a workload that replays the same long prompt is billed on. On V4 Pro at peak it went from $0.003625 to $0.044 a million tokens, 12.14 times. Output on the same model went from $0.87 to $3.96, 4.55 times. A cost model that applied one blended multiple across input and output would understate the cached-input line by about 2.67 times, and a pipeline that replays a long context sends most of its input tokens down that line.

One million tokens on the flat list, and on the two columns that replaced it

Line, per 1M tokens	Flat list to 12 August 2026	Off-peak, 4 November 2026	Peak, 4 November 2026	Multiple, off-peak	Multiple, peak
V4 Pro, cached input	$0.003625	$0.022	$0.044	6.07x	12.14x
V4 Pro, cache-miss input	$0.435	$0.66	$1.32	1.52x	3.03x
V4 Pro, output	$0.87	$1.98	$3.96	2.28x	4.55x
V4 Flash, cached input	$0.0028	$0.003	$0.006	1.07x	2.14x
V4 Flash, cache-miss input	$0.14	$0.15	$0.30	1.07x	2.14x
V4 Flash, output	$0.28	$0.60	$1.20	2.14x	4.29x

Source: DeepSeek Models & Pricing: Internet Archive captures of 2 July 2026 and 12 August 2026 for the flat list, 18 August 2026 for the first two-column table, and the page as read on 4 November 2026. Multiples are Mosaic Singularity analysis, November 2026, computed against the 12 August 2026 flat rate. The V4 Pro rates have not changed since 18 August 2026. The V4 Flash rates were reduced on 10 September 2026 from the $1.32 peak output first published; the Flash figures here are the current ones.

Flash and Pro have not moved together. The first peak column, published on 18 August 2026, put V4 Flash cached input at $0.014 and output at $1.32. The page of 4 November 2026 puts them at $0.006 and $1.20, because Flash was cut on 10 September 2026 and Pro was not. A cost model built from the August column is therefore still exact on Pro, the line that moved most, and 133% too high on Flash cached input.

The exposure is a function of longitude, not of architecture. One global assistant serving Beijing, Berlin and New York is billed at the top of the list for the Asian business day, at half for the European afternoon and at half through the American day. A support desk in Singapore and an agent loop that runs while a user waits cannot be moved out of that window; a nightly evaluation run can.

For an operator that resells the capability at a fixed price a task, the clock becomes a margin line rather than a cost line. The rate it pays moves by two between the morning and the night while the price it charges does not. That is why this lands in the finance forecast and not only in the platform team's configuration.

The escape is 47% cheaper at the supplier's peak and dearer off it

Mistral's model page for mistral-large-4 records a 6 October 2026 public preview, 52 billion active and 1.05 trillion total parameters, and its release post prices it at $0.70 input, $0.07 cached input and $2.10 output a million tokens, under a licence that lets a buyer run the weights itself. Against DeepSeek V4 Pro at peak those two rates are each 47% lower. Against V4 Pro off-peak they are about 6% higher on both, and on cached input Mistral is dearer on both columns, $0.07 against $0.044 at peak and against $0.022 off it. The European list is a hedge against the clock, not a cheaper list.

> *"The model is free to use, modify, fine-tune, and commercialize, with a single condition: businesses that operate a Model-as-a-Service offering must obtain a separate commercial license from Mistral before using the model commercially."*
>
> Mistral, release post for Mistral Large 4, 6 October 2026

The exit narrowed on the other side in the same weeks. On 4 November 2026 the Hugging Face model hub carried ten models under Qwen-4, Qwen3-4B, Qwen3.5-4B and Qwen3.8-27B among them, every one a Qwen3 generation model and none of them Qwen-4 weights. Renting got dearer by the hour in the weeks that owning became possible in Europe, and the Chinese open-weight route an APAC buyer would price against it had no Qwen-4 weights on the hub that day.

Four decisions, four seats, one date

By 31 January 2027 the Chief Financial Officer re-forecasts the inference line and the CIO or Chief AI Officer re-times the estate, because a signed FY27 plan cannot absorb a rate that moves by two inside a business day. The General Counsel works to the same 31 January 2027, and that work gates the only option that removes the clock: the licence condition has to be read before capital commits to self-hosting, not after. The regional owner in Asia Pacific carries the peak column until then, on 7 of its 9 business hours. All four arrive as one approval and not four: whether to absorb the clock by re-timing the estate or remove it by funding a self-hosted path is one capital decision, and it reaches the board with the FY27 plan on 31 January 2027, on the Chief Executive's recommendation.

What each seat owes before the FY27 plan is signed

Seat	Decision to take	By when	Why this one, from this record
Chief Financial Officer	Re-forecast the FY27 inference line on the two published columns, and restate cached input at $0.044 a million at peak rather than $0.003625	31 January 2027, when the FY27 AI budget is signed	Cached input rose 12.14 times at peak and output 4.55 times, so one blended multiple understates the cache line by about 2.67 times
CIO or Chief AI Officer	Move every batch, evaluation and embedding-refresh job out of 01:00 to 04:00 and 06:00 to 10:00 UTC, and record which interactive workloads cannot leave the local business day	31 January 2027, in the same plan	New York spends 0 of 9 business hours in peak and Beijing 7 of 9, and the difference is when the work runs rather than what it runs on
General Counsel	Read the Model-as-a-Service condition in the Mistral Ecosystem Open Weight License v 26.10 before a product built on those weights is offered to a third party	Before the FY27 plan commits capital to a self-hosted path on 31 January 2027	The licence grants modification and commercialisation and attaches one condition, triggered by how the model is served rather than how it is used
Regional CFO or CIO, Asia Pacific	Price the local business day at the peak column, or move the interactive inference that must run inside it onto a self-hosted or fixed-price path	31 January 2027, in the regional AI budget line	An APAC operation is in peak for 7 of its 9 business hours, so its whole day is bought at the top of the list while a US-hours operation buys at half

Source: Mosaic Singularity analysis, November 2026, from the price columns read on 4 November 2026 and the licence read on the same date. The 31 January 2027 date is the point at which FY27 AI budgets and inference-siting decisions are set.

What to watch, and the dates that settle it

- By 31 January 2027: whether the next revision of the page still carries two columns. A return to a single rate would make the clock a promotion rather than a price.
- Before 30 April 2027: whether a second Chinese lab publishes a peak clock. Until one does, the exposure belongs to one supplier, and a buyer can answer it by not buying from that supplier.
- At every reading to 30 April 2027: whether the peak window itself moves. It has been narrowed twice since 16 August 2026, weekends first and then Chinese public holidays, and a third narrowing would take the European morning out of peak as well.
- Before 30 April 2027: whether a pro column is still priced at $1.98 and $3.96. The page of 10 September 2026 carried a note that requests to deepseek-v4-pro would be routed to V4.1 Flash and billed at the Flash price from 12:00 Beijing time on 14 September 2026; the page of 4 November 2026 carries no such note and still prices it.

The question these dates settle is not which model to buy. It is whose clock the workload runs on, and that is sited, not selected.