LIFT

LABORATORY OF INNOVATION AND FUTURE TRANSFORMATIONS

POLICY PAPER 2026-03-EN · OCTOBER 2026

Rising AI spending and the rise of open-weight models

What OpenRouter data say about volumes, prices and the shift toward open models

Antonin Bergeaud
Paul Delbouve
Etienne Grass

KEY POINTS

— Firms' AI spending is rising even as the unit cost of a token falls sharply. Most of the increase comes from volumes, not prices.

— On OpenRouter, the share of open-weight models rises from 40% to 73% of tokens between January and August 2026, and from 11% to 34% of spending.

— Users associated with France devote 31.0% of their spending to open models, compared with 19.3% worldwide.

— Calls to open models are four times longer than in January, but their average cost is multiplied by only 3.8: the cost advantage persists.

— Reallocation toward open models accounts for 43% of the simulated decline in the cost of a constant French basket, and opens up room for deployment rather than automatic savings.

US firms surveyed in March 2026 by the Federal Reserve Bank of Atlanta expect to spend 2,068 dollars per employee on artificial intelligence this year, compared with 1,358 dollars in 2025, an increase of more than 50%. Scaling up is thus beginning to reveal a bill that early experiments had partly overlooked. Several large firms have begun to cap spending, restrict certain uses or steer their employees toward less expensive models (Financial Times). The development of agents performing repeated operations adds to this pressure by multiplying calls and tokens consumed, leading these firms to set up monitoring and control systems inspired by those used for the cloud (Wall Street Journal). Some firms that had opened access widely are now beginning to introduce usage limits and to ask which uses really justify their cost.

Should this be seen as the end of the "golden age" of AI, as one hears more and more often? Not necessarily. The explosion in spending comes first from the volumes consumed. The cost of reaching a given level of performance has, for its part, fallen sharply: the Stanford AI Index Report, published in 2026, estimates that it was divided by more than 280 between November 2022 and October 2024 for performance comparable to that of GPT-3.5. But firms obviously no longer stop at this level of performance and consume much more. They deploy more applications, use agents and send increasingly long contexts. This phenomenon recalls the Jevons paradox, or rebound effect, which is common when a fall in the cost of using a technology stimulates its consumption.

As these uses develop, firms are also learning to organize them better. The amount of the bill then depends largely on how they use the models, in terms of the type of information they send them, the operations they have them repeat and the model chosen for each task. A firm can therefore reduce its costs by changing the architecture of its uses, for example by reserving the most expensive models for the tasks that really require their capabilities. Controlling spending thus involves organizational and technological choices, beyond the mere negotiation of prices.

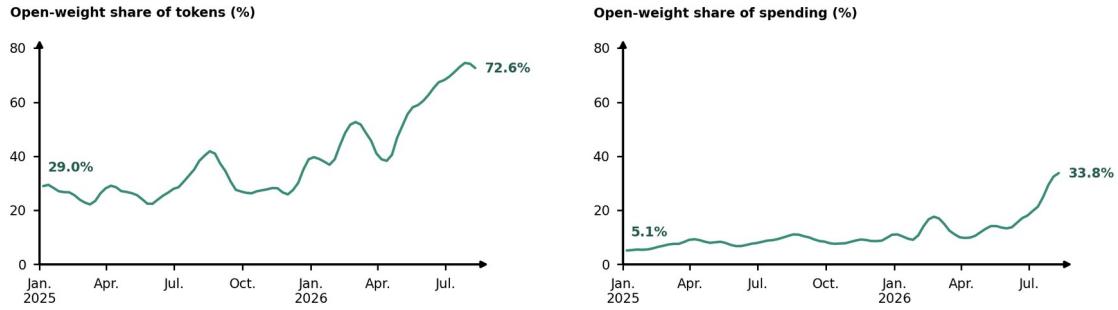
Since the beginning of 2026, observed uses suggest a change in the way models are chosen: users adapt them more closely to each use in order to reduce its cost, with no apparent loss of quality. To study this, we use data from OpenRouter, a platform widely used by developers and firms, which provides access to several hundred models through a single interface and makes it easy to switch from one to another. Its users are therefore particularly well placed to compare offerings and adapt their choices quickly. They are not representative of all AI users, but they may be ahead on practices likely to spread more widely later. We are thus interested in how their trade-offs evolve rather than in the levels observed or the market shares of the various models.

These data show a clear shift. Between early January 2026 and the week of August 10, the share of open models rises from about 40% to 73% of tokens consumed. Their share of spending also increases, from 11% to 34%. Open models therefore occupy a growing place in budgets, while accounting for a much larger proportion of the volumes processed. This gap sheds light on their economic appeal for intensive users, without directly measuring the savings achieved in performing the same task.

Three limitations should nevertheless be kept in mind. First, OpenRouter represents only a limited part of the market: a large share of uses goes directly through the developers' APIs or through other intermediaries. Second, the platform is particularly used to build workflows, which probably overrepresents agentic uses relative to average use. Finally, developers now seek to improve their position on it, notably through targeted promotions. These features call for caution in interpreting the levels observed and strengthen the case for reasoning mainly about their changes.[1]

[1] OpenRouter's data nevertheless have the great advantage of being accessible to researchers. We thus obtained from their teams detailed information on 1,125 trillion tokens, 7.3 million users and 116 billion requests covering the period from

Figure 1. Share of open models between January 2025 and August 2026



Notes: Share of open models in total tokens (left) and in users' spending (right), in %. Arithmetic mean of weekly shares over the current week and the three previous ones. Weeks from January 6, 2025 to August 10, 2026. Source: OpenRouter

In France, slightly greater use of open models

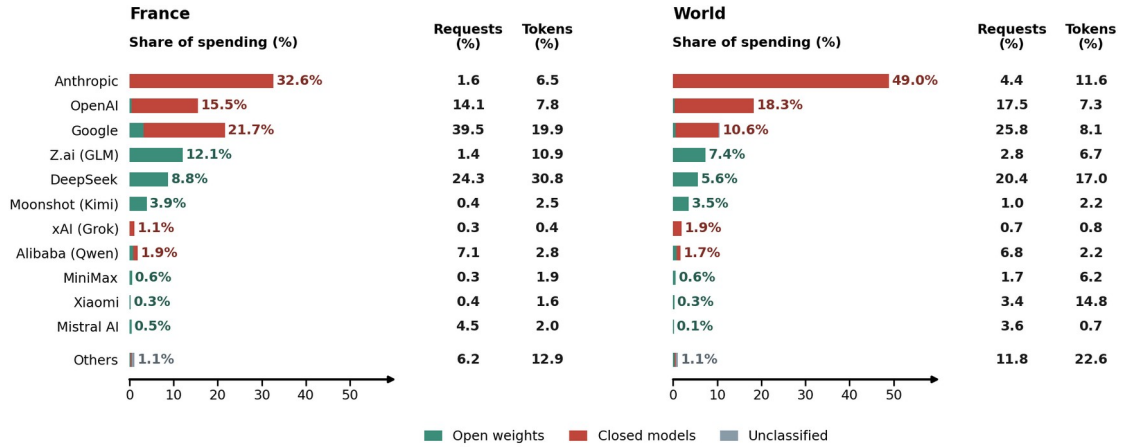
Users associated with France devote a larger share of their OpenRouter bill to open models[2]. In July 2026, these account for 31.0% of recorded French spending, compared with 19.3% worldwide. The picture by provider shows which models lie behind this difference.

Anthropic remains the leading recipient of spending in both scopes, but its share is 32.6% in France, compared with 49.0% worldwide. Google holds a much larger place in French spending, with 21.7%, compared with 10.6% worldwide. OpenAI completes the top three. Together, these three providers account for 69.8% of French spending and 77.9% of global spending observed on the platform.

January 2025 to August 2026. To our knowledge, there is no public and official database offering information at this level of detail that would be representative. Alternatives exist, notably (1) those of Artificial Analysis, which have the advantage of being built on a value analysis index, but are more aggregated. These data demonstrate the economic benefit of relying more on open models. They also show that prices have fallen since the release of the Kimi K3 model; (2) those of Ramp, drawn from its clients, which correspond to actual corporate spending data but whose representativeness is also difficult to establish; and finally (3) those of other routers such as Vercel or Requesty, which a priori have the same limitations as those of OpenRouter.

[2] An AI model is defined in part by its weights, that is, the billions of parameters learned during training that determine how it responds. In this text, we call "open" the models whose weights are accessible and can be downloaded. A firm can then run them on its own infrastructure or with a cloud provider, and generally has more latitude to adapt them to its needs. Conversely, closed models are mainly accessible as a service, notably through an API, without access to their weights. This label does not mean that everything is open. Training data, the code used to train the model or certain technical details may remain private, and licenses may impose restrictions. The more precise term is therefore "open-weight model".

Figure 2. Distribution of spending, requests and tokens by provider, France and world



Notes: July 2026. Each measure is expressed relative to the total of the scope concerned, all providers combined. The world includes France. The eleven providers shown combine the top ten by spending in each scope, in a common order based on global spending; all others are included in “Others”. Colors distinguish the status of the models, not that of the provider. Recorded input and output tokens, cache and reasoning already included. Source: OpenRouter

The larger place of open models in French spending relies in particular on Z.ai, the developer of the GLM models, and on DeepSeek. Together, these two providers account for 20.9% of spending in France, compared with 13.0% worldwide. DeepSeek holds a particularly important place in volume, with 30.8% of tokens consumed in France, compared with 17.0% worldwide. Mistral AI is also more present in French spending, but its weight remains limited: 0.49%, compared with 0.14% worldwide.

The ranking nevertheless changes depending on the measure used. In France, Google receives the largest number of requests, DeepSeek processes the most tokens and Anthropic concentrates the most spending. Worldwide, Google and DeepSeek also come first for requests and tokens. The number of calls sent to a provider therefore gives a very partial picture of its weight in the bill. This weight also depends on the pricing of the models and on the intensity of the uses entrusted to them.

This intensity varies widely, even among open models. In France, a request sent to Z.ai’s models contains on average about 52,400 tokens, compared with 8,600 for DeepSeek, adding inputs and outputs. A limited number of calls can thus represent a very large volume of content processed and high spending, a concentration also found in usage, where a few users often account for the majority of token consumption within a firm. Likewise, there is a very wide disparity in price ranges among providers of open models. DeepSeek thus positions itself on very low-priced models, on the order of $0.5/MTokens (dollars per million tokens) on average, unlike Moonshot.ai, Z.ai or Qwen, which offer varied ranges of models whose prices can approach $1.5/MTokens and go up to $6/MTokens for the best-performing models.

Quality determines which substitutions are economically possible

The rising share of open models in tokens might suggest that firms have given up moving up in value in order to contain their spending. But they can now also obtain from open models capabilities that previously justified the use of closed models. Figure 3 illustrates this possibility of substitution. In July 2026, GLM 5.2 achieves a performance close to that of Claude Opus 4.7 on the benchmark published by Artificial Analysis, with 52.6 and 55 points respectively. Yet its observed spending per million tokens is 0.40 dollar, compared with 1.74 dollars for Claude Opus 4.7. These gaps, however, concern the uses actually observed, whose content may differ across models.

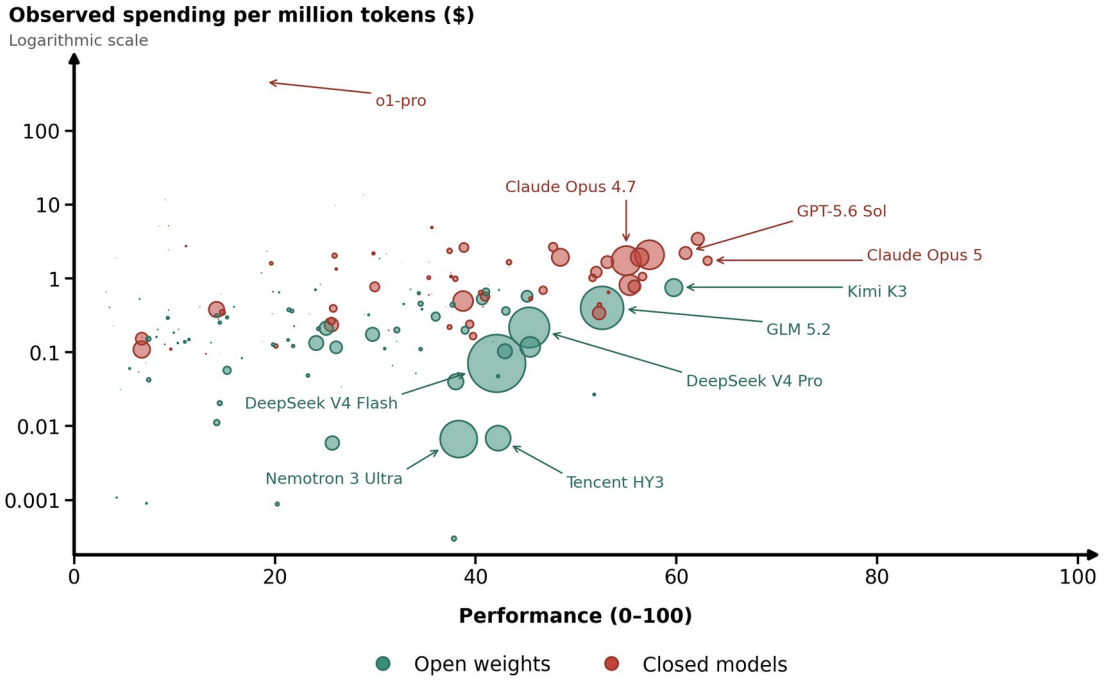
As Figure 4 shows, this decline in unit spending goes hand in hand with a shift toward better-rated models. For the 20–30 quality class, spending per million tokens falls from 1.46 dollars in March–

April 2025 to 0.28 dollar in July–August 2026. At the same time, the 40–50 class becomes the most used among those shown, while requests sent to models rated 50–60 increase rapidly. At the end of the period, the 20–30, 30–40 and 40–50 classes even show very similar average spending per token. Access to higher capabilities therefore does not systematically come with a higher unit cost.

For a firm, the benefit of switching models depends on what the most expensive model still adds to its concrete tasks. A performance lead can justify its price if it avoids errors or reduces costly human verification. When several models deliver a sufficiently reliable result, this lead loses part of its economic value. An open model can then replace a closed model without lowering requirements. But the cost of switching models, and in particular of abandoning the reflex of using the one offering the highest quality scores, can be high. As seen in Figure 2, even on a platform that makes switching models very easy, the best-rated offerings concentrate a large part of spending. This persistence does not mean that they are used for all operations: models rated at least 60 account for less than 1% of calls, but more than 15% of the bill. Outside a unified platform, additional integration constraints may slow reallocations further, although the data used here do not allow their importance to be measured.

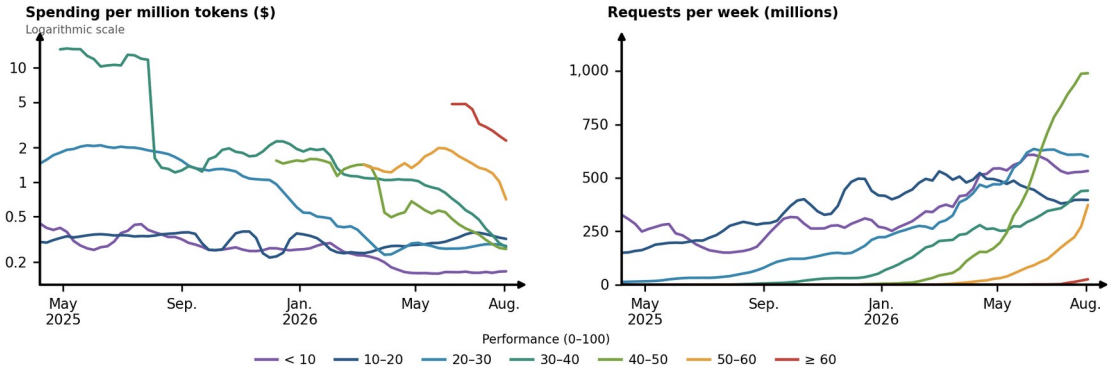
To assess the economic cost of AI, however, one must look beyond the price of tokens and reason in terms of the total cost of a usable result, including in particular rework and human review. Studies that value a complete workflow rather than a token show that, for realistic volumes, the cost of errors and of checking them can exceed the token bill by two to three orders of magnitude when a model lacks reliability. Reliability on the tasks actually performed by the organization then becomes a more relevant selection criterion than the score obtained on a benchmark alone. Savings can therefore come from a finer allocation of operations across models, but also from eliminating unnecessary calls, reusing results already obtained or reducing human review. Cost control thus depends on the organization of the entire processing chain as much as on the model used for each call.

Figure 3. Performance and usage cost of models in July 2026



Notes: The area of each bubble is proportional to the tokens consumed. Green: open weights; red: closed models. Spending per million tokens is the ratio of recorded spending to input and output tokens, cache included, shown on a logarithmic scale. The 191 rated and classified models shown cover 94.6% of tokens in the scope considered. Source: OpenRouter and Artificial Analysis performance scores.

Figure 4. **Usage cost and diffusion by performance class**



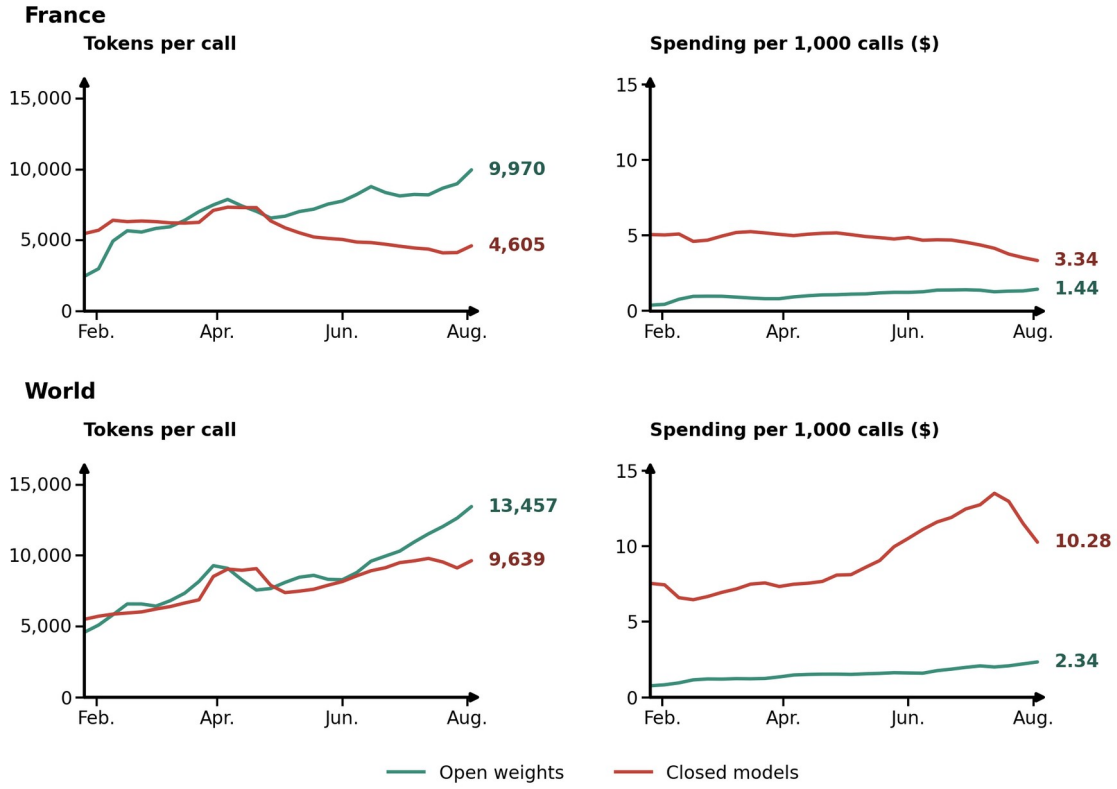
Notes: Open and closed models combined. Four-week smoothing: ratio of the sums of spending and tokens on the left, average of weekly requests on the right. Data from March 17, 2025 to August 9, 2026. Costs on a logarithmic scale, excluding activity identified as billed directly to providers; requests for all calls. Models without a score excluded, coverage presented separately. The composition of models and calls may vary within classes. Source: OpenRouter and Artificial Analysis performance scores.

Longer calls, whose cost advantage persists

Calls to open models have become much longer, while remaining less expensive than calls to closed models. In France, over the last four weeks of our sample, which ends on August 9, 2026, a call to an open model contains about 10,000 tokens, compared with 4,600 for a closed model. Yet spending per 1,000 calls amounts to 1.44 dollars for open models, compared with 3.34 dollars for closed models, that is, 57% less. Worldwide, the gap in spending per call reaches 77%. These comparisons concern the consumption observed, without assuming that the models perform the same tasks.

This relative advantage nevertheless coexists with a rise over time in the cost of open calls. In France, their average length is multiplied by four between the first weeks of 2026 and the end of the period. Over the same time, spending per 1,000 calls rises from 0.38 to 1.44 dollars, a 3.8-fold increase. The slight decline in the average cost per token therefore does not offset the lengthening of calls caused by the massive use of agentic AI from 2026 onward. Open models remain less expensive per call than closed models, but each open call now costs more than at the beginning of the year.

Figure 5. **The cost advantage of open models persists despite longer calls**



Notes: Ratios of sums over four complete weeks, from January 5–February 1 to July 13–August 9, 2026. Calls identified as billed directly to providers are excluded from requests, tokens and spending; calls with zero spending are kept. Input and output tokens, cache and reasoning already included. Variable populations and models. Source: OpenRouter.

The French progress of open models thus concerns mainly the intensity of calls. Across all recorded flows, their share of requests rises from about 48% to 54%, while their share of tokens rises from 34% to 71%. Part of this gap also comes from the shortening of closed calls. The sharp rise in open volumes therefore corresponds to a much more limited shift in the number of calls.

Our data allow only a partial decomposition of these tokens, but they give an indication of the origin of this lengthening. Inputs reported as “cached” account, in accounting terms, for 79% of the increase in tokens per open call in France and 92% worldwide. Reasoning tokens contribute less than 1% of this increase in both scopes. Most of the increase therefore comes from contexts that are longer or reused from one call to the next, much less from an increase in the tokens devoted to reasoning.

In practice, an application can send the same file at several stages of a processing chain. A first request may ask for it to be analyzed, a second for a piece of information to be extracted from it and a third for the result to be checked. The content of the file is then counted in the tokens at each call, even when it is reused from a cache. The total volume can thus increase much faster than the amount of new information actually transmitted. This increase is also passed on less than proportionally to the bill, because inputs served from the cache are generally billed at a fraction of the standard input price, often around one tenth. This helps explain why, while the average length of open calls is multiplied by four, their average cost is multiplied by only 3.8.

This change may correspond to economically useful uses. A longer context window makes it possible, for example, to send a complete file directly, without first having to select the relevant passages. Reusing this context at several stages can also simplify the design of an application or a workflow. But some repetitions may bring no gain and, on the contrary, constitute a source of potential savings. The number of tokens alone therefore does not tell whether the firm is performing more useful work or simply mobilizing more resources to obtain a comparable result.

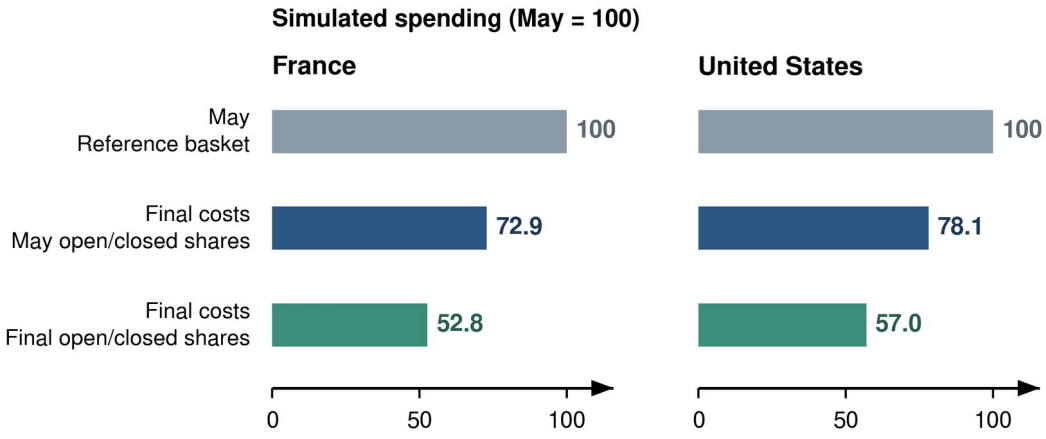
Unit savings open up room for deployment

Rising spending can mask a decline in the cost needed to finance a given volume. To assess its magnitude, we build a basket of tokens from May 2026 consumption, then keep its volume and its distribution across performance classes. We calculate what this same basket would cost at the average costs observed at the end of the period. Open and closed models are combined in this comparison.

In France, a basket worth 100 dollars in May is valued at about 53 dollars at end-of-period average costs within the same performance classes. The US result is similar, at 57 dollars. But if we keep exactly the model versions of the initial basket and the number of tokens assigned to each, the indices remain close to 100 in both countries. The decline observed with a constant performance distribution therefore does not correspond to a uniform reduction in the cost of the models already in use.

The progress of open models accounts for a substantial share of this decline. If we keep, within each performance class, the shares of tokens sent to open and closed models in May, while applying the average costs observed at the end of the period, the French basket is valued at 73 dollars. With the shares observed at the end of the period, it is worth 53 dollars. Reallocation toward open models thus accounts for 20 of the 47 points of decline, or 43%. In the United States, it accounts for 21 points out of 43, or 49%.

Figure 6. Reallocation toward open models accounts for 43% of the simulated decline in the cost of the French basket



Notes: Reference from May 4 to 31, 2026; end of period from July 13 to August 9. May token volume kept within each five-point performance class. The two simulations apply end-of-period average costs per token by openness status, with unclassified models isolated, keeping their initial shares or using their final shares within each class. Their difference measures reallocation at final costs; the contribution of unclassified models is less than 0.01 point. Source: OpenRouter

The rest of the decline comes from changes in average costs within open and closed models. These price cuts are not independent of the diffusion of open models, since the availability of open-weight substitutes, Chinese as well as American, for example with the Nemotron family, constrains the pricing of closed models. In this respect, the 43% attribution is a lower bound on the total effect of open weight.

These changes in average costs may reflect price cuts, reallocations between models in the same category or changes in the composition of calls. The increased use of open models therefore explains part of the decline in the cost of the basket, but not all of it.

A more complete decomposition is presented in Table 1, which distinguishes the models already used in May from those that were absent from the initial basket. The two exercises partly capture the same movements. The arrival of a new open model, for example, can both increase the share of open

models and shift usage toward a model absent in May. The contributions from the two decompositions should therefore not be added together.

These amounts measure an accounting possibility, not savings actually achieved by each firm. The classes group models by their performance on the benchmark, without guaranteeing that they are interchangeable for all tasks. Nor does the calculation hold constant the users or the composition of calls, in particular the share of cache and outputs. It therefore measures neither the cost of strictly identical work nor the causal effect of adopting open models.

Table 1. Decomposition of the price of AI use between May and July 2026

Basket cost decomposition	France	United States
May reference	100.00	100.00
Final costs with initial shares by status	72.91	78.07
Final costs with final shares by status	52.76	57.00
Total decline of the basket, in points	47.24	43.00
Of which reallocation toward open models, in points	20.15	21.07
Of which costs at constant shares, in points	27.09	21.93
Of which change in unclassified shares, in points	−0.0038	0.0011
Share of the decline attributed to reallocation toward open models	42.7%	49.0%

Notes: Contributions are expressed in index points, on initial spending normalized to 100. The slight discrepancies in sums in the table come from rounding.

How much additional activity can unit savings finance?

At the costs observed at the end of the period, the budget that financed the French May basket would buy 90% more tokens, compared with 75% for the US basket. The progress of open models contributes strongly to this margin. In France, applying final costs while keeping the initial shares of open and closed models would finance only 37% more tokens. Reallocation toward open models raises this capacity to 90%. The calculation keeps the distribution across performance classes constant; it does not measure an equivalent increase in the number of successful tasks.

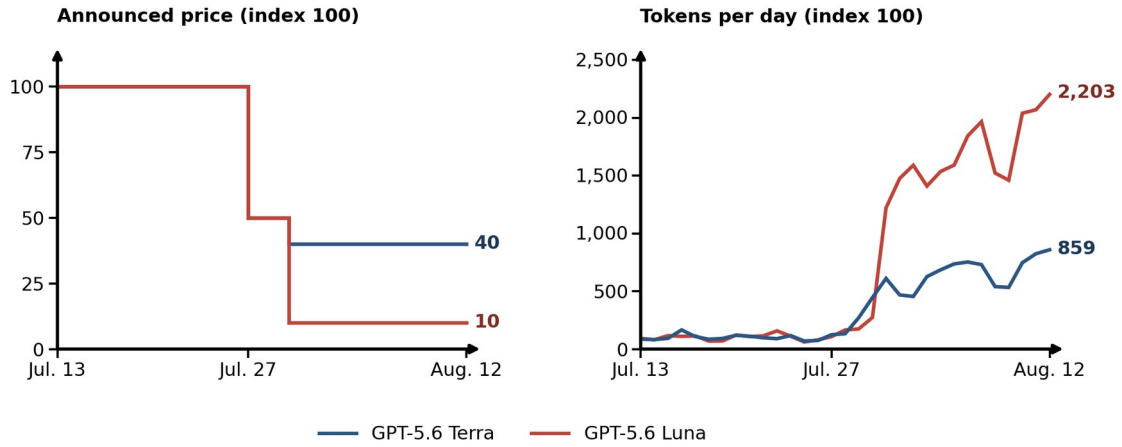
Part of this margin may be absorbed by the additional consumption it makes attractive. Applications that were previously too costly become feasible and other uses emerge. Likewise, for uses already deployed, users can send more context or multiply calls. Lower costs can therefore reduce spending, be entirely reinvested in more usage, or even lead to higher spending.

The price elasticity measures this response of volumes to price. An elasticity of −0.1 means that a 1% price cut generates, all else equal, about 0.1% additional consumption. Spending then falls, but less than the price. The threshold of −1 corresponds to locally constant spending: the increase in volumes offsets the price decline. With a stronger response, for example an elasticity of −1.5, spending increases. We are then in the realm of the Jevons paradox.

This elasticity is difficult to estimate in a market where models and uses change rapidly. A decline in spending per token may result from a lower price, but also from the choice of different models. The increase in the traffic of a cheaper model may, for its part, reflect a shift from its competitors rather than an increase in total consumption. It is therefore necessary to exploit price cuts on identified versions and to follow the same accounts across all models.

We use the July 2026 promotions on GPT-5.6 Terra and Luna. The successive cuts of July 27 and 30 reduce their effective prices by 60% and 90% (Figure 7), respectively, on OpenRouter via the standard OpenAI service. Each account's exposure is calculated from its usage before the announcement. If an offering represents 40% of its initial bill and its price is halved, the cost of its initial basket theoretically falls by 20%. We then compare the change in the total consumption of more and less exposed accounts, taking into account their initial activity and the changes common to each country.

Figure 7. The July promotions are accompanied by a sharp rise in the volumes of the models concerned



Notes: Standard OpenAI service; cells in which all four BYOK fields are filled in and equal to zero. Announced price on the left, daily input and output tokens on the right. Base 100: price before July 27 and average daily tokens from July 13 to 26. Data from July 13 to August 12, 2026, without smoothing. Prices concern non-cached input and output; the quantity includes cached tokens. Closed models, single episode with concurrent service changes. Descriptive comparison, without causal identification or tracking of a constant population. Sources: OpenRouter, OpenAI and authors' calculations.

Tracking 13,052 accounts in six countries yields an estimate of about −0.36. We therefore use −0.4 as the central elasticity for the simulations. This choice remains uncertain: the episode is short and the results are sensitive to the weight of large consumers. Its transposition to the French basket is a working assumption, not a precise measure of the elasticity. We therefore accompany it with a scenario with no volume response and a strong-response scenario, at −1.5.

In Table 2, unit savings are distinguished from the consumption they can finance. With open/closed shares unchanged within each performance class, the change in average costs brings the cost of the French basket down from 100 to 72.9. Reallocation toward open models then reduces it to 52.8. It thus accounts for about 43% of the decline in unit cost. For each scenario, we calculate the response of volumes to these two cost levels in order to isolate what reallocation toward open models adds.

Table 2. Unit savings, for initial spending of 100

Contribution to spending	No response ε = 0	Central ε = −0.4	Strong response ε = −1.5
Change in average costs, at initial open/closed shares	−27.1	−27.1	−27.1
Additional savings from reallocation toward open models, at fixed volume	−20.1	−20.1	−20.1
Increase in spending due to volumes, without this reallocation	0.0	+9.8	+44.2
Additional effect of the volume response, after reallocation toward open models	0.0	+5.6	+40.7
Final spending	52.8	68.1	137.7

Notes: Basket of May 4–31, 2026, valued at the costs of July 13–August 9, with a constant distribution across performance classes. Reallocation valued at final costs, then volume response simulated at constant elasticity. Average costs combine prices and the composition of usage. The attribution depends on this order of calculation; it does not measure a causal effect of openness. The extreme scenarios are not statistical bounds. Any discrepancies are due to rounding. Source: OpenRouter, authors' calculations.

In the central scenario, spending falls by 32%, despite a 29% increase in tokens consumed. Without progress in the share of open models, it would fall by only 17%, while volumes would increase by 13%. Reallocation toward open models thus lowers the spending index from 82.7 to 68.1, that is, 14.6 fewer points of spending after taking into account the additional consumption it generates. Part of the unit savings is therefore reinvested in more usage, while the rest actually reduces spending.

The strong-response scenario illustrates another possibility. Spending then reaches 137.7, compared with 117.1 without progress in open models: unit savings generate an expansion of volumes strong enough to push spending above its initial level. The progress of open models can thus greatly increase deployment possibilities while raising spending. This result depends on the assumed response of users and does not mean that adopting open models necessarily leads to higher spending.

These simulations isolate the additional consumption generated by lower costs. Growth in usage independent of this decline may add to it. Rising spending therefore remains compatible with a sharp decline in unit cost. Its interpretation depends on what the additional volumes finance: more useful tasks, more context or simply redundant calls. To understand the change in spending, one must thus distinguish the decline in unit cost, the response of volumes and the extension of uses.

Implications for deployment choices

The costs and volumes observed on OpenRouter are not sufficient to determine the optimal strategy of each organization. They nevertheless suggest a decision framework for users of language models: measure the total cost of using AI to automate a task at the expected level of quality, avoid unnecessary calls and token volumes, then decide whether the savings and gains thus generated should reduce spending or finance new uses. Building on these results, this final section presents several principles and avenues for reflection to inform organizations' choices about deploying AI models.

Controlling spending starts with the cost of a task performed at the required level of reliability. This cost includes all the necessary attempts and the time spent on human review. Savings on calls can be canceled out by the cost of additional corrections. A more expensive model can, on the contrary,

avoid this rework. Performance scores help select the offerings to test, but only an evaluation on the organization's own tasks makes it possible to assess this trade-off.

Optimization must also address the amount of processing needed for each result. Reusing an answer or selecting the useful information can avoid redundant calls. The cache reduces the cost of some repetitions, without removing the need to examine their usefulness. Stopping rules can limit loops that no longer improve the result. These choices can lower consumption per task, even when the organization increases the number of tasks entrusted to AI. Switching models is one way to reduce the bill; the design of the application can directly reduce its volumes.

A final trade-off, which our data do not allow us to quantify, concerns the operating mode. For high-volume, predictable activity, dedicated inference capacity can be compared with pay-per-use, in particular when the model weights are available and its license allows it. The cost then shifts partly from spending per call to capacity that must be financed and operated. Available simulations by integrators give orders of magnitude. For about thirty billion tokens per month, running an open model with thirty billion parameters in-house costs, over three years, about one tenth of the cost of frontier APIs. The same exercise shows, however, that renting these same weights from a hosting provider remains cheaper than running them oneself, and that the break-even point against APIs lies at around a few hundred million tokens per month. The savings therefore come from switching to open models more than from owning the infrastructure, whose justification rests on data residency, the right to fine-tune and price stability.

The value of dedicated capacity depends on its utilization rate and on the skills needed to operate it. Irregular activity may be better suited to a pay-per-use service. The OpenRouter data do not quantify the savings from this change in architecture.

The ability to compare solutions itself has a fixed cost (whose amortization depends on usage volumes). Reusable business-specific tests can reduce the work required for each new trade-off. The portability of applications preserves the possibility of switching providers when the gain justifies the checks. These investments can be profitable for intensive uses and disproportionate for a small organization. Pooling evaluations on well-defined tasks would help small firms and public administrations compare offerings without redoing all the work.

These principles have a particular implication for public buyers. A comparison in cost per validated task would bring spending closer to the expected service, with explicit security and reliability requirements. It would leave open the possibility of choosing a more expensive closed model when its advantage is useful, or an open model when its capabilities suffice. The share of open models and the number of tokens consumed are not, in themselves, performance objectives.

The savings thus generated can be kept or devoted to new uses. Time freed up does not automatically become budget savings; it can also improve the service or make it possible to absorb more activity. The decision to extend deployment should specify which of these gains is expected and how it will be verified. Rising spending can be justified by more useful work, while falling spending can mask costs shifted onto employees. Jointly monitoring the full cost and the results obtained makes it possible to decide which uses deserve to be extended.

Authors

Antonin Bergeaud, HEC Paris and LIFT.

Paul Delbouve, Paris School of Economics, FIL and LIFT.

Etienne Grass, Capgemini.

The authors thank Justin Summerville, of OpenRouter, for access to the data. The views expressed in this paper are those of the authors alone and not of Capgemini.

About this translation

This document is a machine translation of LIFT Policy Paper 2026-03, « L'augmentation des factures d'IA et la montée de l'open weight », published in French in October 2026 and available at heclift.github.io/website. The translation was kept as close as possible to the original and reviewed by the authors. In case of discrepancy, the French version prevails.

About LIFT

The Laboratory of Innovation and Future Transformations (LIFT) is a research laboratory at HEC Paris that studies the impact of artificial intelligence on firms, employment, innovation and growth. It brings together academic researchers and practitioners across six research areas, produces original data and regularly publishes notes, studies and indicators for public and private decision-makers.

LIFT publications

Policy Papers present empirical results and their implications for public and private decision-making in a short format. Working Papers present research in progress. LIFT trackers provide regularly updated indicators on the diffusion of AI. The views expressed are those of the authors alone and do not necessarily reflect those of their employers and institutions or of the laboratory's partners.

How to cite

Bergeaud, A., P. Delbouve and E. Grass (2026), "Rising AI spending and the rise of open-weight models", LIFT Policy Paper 2026-03-EN, HEC Paris.

Contact

LIFT – HEC Paris, 1 rue de la Libération, 78351 Jouy-en-Josas
lift@hec.fr · heclift.github.io/website

Published under a Creative Commons CC BY 4.0 license. Reproduction is permitted with attribution.

