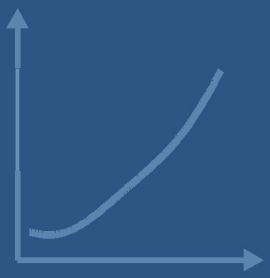
LIFT

LABORATORY FOR INNOVATION
AND THE FUTURE OF WORK



POLICY PAPER 2026-04 · OCTOBER 2026

Innovation and Catch-up in the AI Model Market

An analysis of OpenRouter data, January 2024 to August 2026

Paul Delbouve

KEY POINTS

- In AI, the best model priced at $0.10 per million tokens or less - at least fifty times cheaper than a frontier model - catches up with the latter's quality in less than a year.
- This catch-up could discourage rising investment, estimated at $715 billion in 2026 for the main AI developers, by reducing the rents generated by a technological lead.
- On OpenRouter, frontier labs processed just 28% of tokens in August 2026, down from three-quarters in spring 2025, but still earned 71% of revenue thanks to higher prices.
- Relative to its competitors, a provider releasing a frontier model sees its imputed revenue roughly double in the following months, while users of a model facing competition mainly switch to other high-end models.
- This equilibrium relies on exceptional growth in consumption and remains fragile. Frontier labs' revenue share fell from 82% to 71% between mid-July and mid-August 2026.

Cheaper competitors now match the quality of the best artificial intelligence models in less than a year. This speed could discourage labs from investing to push the technological frontier forward, since their advantage erodes quickly. Using data from the OpenRouter platform from January 2024 to August 2026, this note measures the returns to this lead and how quickly they are lost. Frontier labs processed just 28% of tokens[1] in August 2026, down from three-quarters in spring 2025, but still earned 71% of revenue. Relative to its competitors, a provider releasing a new leading-edge model sees its revenue roughly double in the following months, while users of a system facing competition mainly turn to other high-end offerings. The incentive to innovate is therefore holding up for now, but relies on very rapid demand growth.

Pushing the frontier is becoming increasingly expensive[2] (Cottier et al., 2024). Training and deploying the most capable systems requires growing amounts of computing power, specialised labour, data and energy. The main AI developers[3] are expected to spend $715 billion on data centres, chips and model development in 2026, up from $400 billion in 2025 (Sajadieh et al., 2026). These outlays precede revenue, and commercial success remains uncertain. A firm will only invest in an innovation if it can generate a sufficient return. A technological advance must therefore provide a temporary advantage, during which the firm can charge higher prices, attract new users or gain market share (Aghion et al., 2014). These rents are part of what encourages firms to undertake costly and risky research (Aghion et al., 2015). If competitors can immediately copy an advance, the advantage disappears and expected returns no longer justify the sums invested (Bloom et al., 2013), slowing breakthrough innovation.

In artificial intelligence, competitors catch up particularly quickly with the most advanced systems, or “frontier models”. Providers such as DeepSeek, Qwen and Z.ai offer similar performance at much lower prices within months of these models’ release. Usage prices, expressed per million tokens, generally exceed $5 for frontier models. Figure 1 compares two quality trajectories: that of the best model available, regardless of price, and that of the best model priced at $0.10 or less, at least fifty times cheaper. The latter catches up with the former in less than a year. In August 2025, for example, the best model in this price range matched the level reached by the frontier in December 2024. At a given quality level, prices thus fall by more than 95% in less than a year.

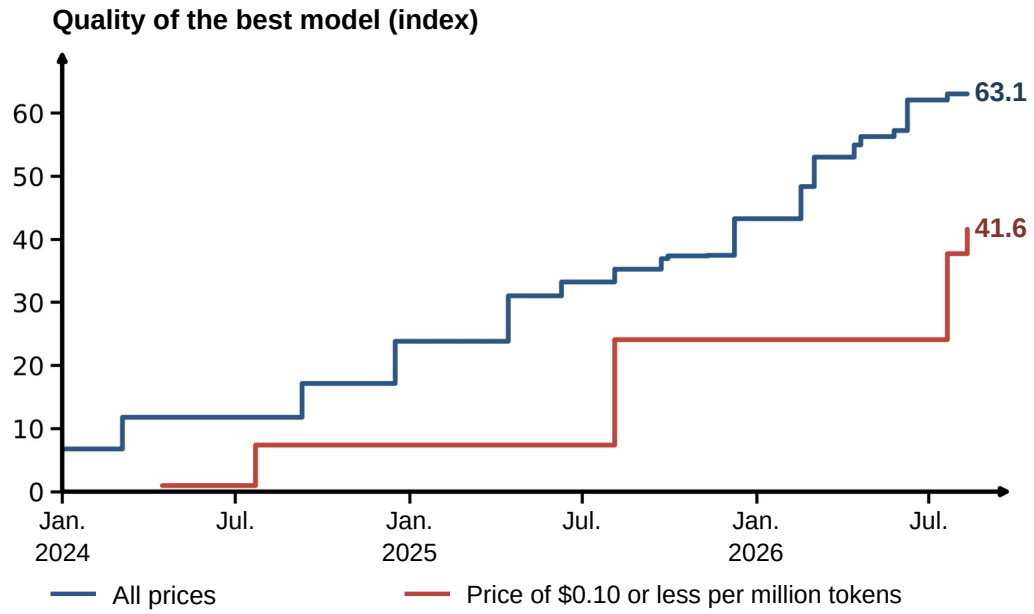
Several features specific to AI may explain this speed. Advances spread through research papers, technical reports and open-source software, allowing developers to draw on common libraries, public models and a shared body of knowledge without repeating every stage of the initial discovery (Azoulay et al., 2024). Distillation also enables smaller models to learn from the answers or reasoning of more capable systems (Hinton et al., 2015; Mukherjee et al., 2023). The relative roles of these factors remain debated, but once a technological path has been validated, following it is considerably cheaper than opening it up (Guo et al., 2025). This catch-up benefits users by lowering the cost of access to AI, but could have slowed innovation. The opposite is happening, in a sector where investment is soaring.

[1] A token is a fragment of text processed by an artificial intelligence model. Model usage is measured by the number of tokens received and generated, or consumed.

[2] Researchers at Epoch AI, who specialise in estimating AI costs, find that hardware depreciation and energy costs associated with training the most advanced models increased by a factor of around 2.4 each year between 2016 and 2024.

[3] Alphabet, Amazon, Meta, Microsoft and Oracle.

Figure 1. **Quality of the best model by price**



Notes: Quality is measured using the Artificial Analysis performance index. The blue line tracks the best available model, regardless of price; its price generally exceeds $5 per million tokens. The red line tracks the best model priced at no more than $0.10 per million tokens. For example, in August 2025, the latter reaches a quality score of 24.1, close to that of the best model available in December 2024. January 2024 to August 2026. Source: Artificial Analysis, OpenRouter, LIFT calculations.

To understand this paradox, we use weekly data from OpenRouter, a platform providing access to hundreds of AI models through a single interface. LIFT Policy Paper 2026-03 already used these data to show that rising AI bills are driven by consumption volumes and that open-weight models are gaining ground (Bergeaud, Delbouve and Grass, 2026). The data track more than 1,000 models from 154 providers between January 2024 and August 2026, including token consumption, listed prices and the company producing each model[4]. We add independent performance assessments published by Artificial Analysis. This detailed tracking allows us to observe both releases of frontier-pushing models and the subsequent arrival of cheaper competitors with comparable capabilities, and then to measure the effects of these two events on the usage and revenue of the providers concerned. Revenue is imputed: it is calculated by valuing consumed tokens at each model's listed price. Since OpenRouter is only one distribution channel among others, the analysis measures demand reallocation and competitive pressure on this platform, without assessing revenue for the sector as a whole.

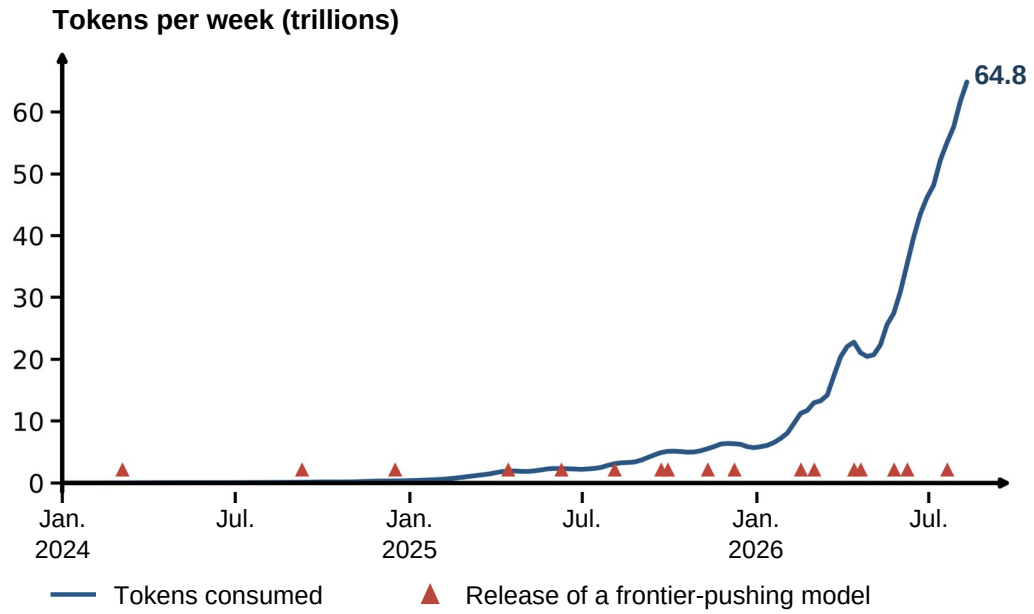
Market growth and structure

Weekly token consumption on OpenRouter rose from around 11 billion in January 2024 to nearly 65 trillion in August 2026. Figure 2 tracks these volumes week by week and marks with a triangle each release of a model that surpasses the highest quality score achieved up to that point. Growth follows a roughly exponential path, which other sources also report for the market as a whole[5]. It accompanies the continued advance of the frontier, consistent with the idea that more capable models open up new uses and broaden the range of tasks for which AI becomes cost-effective.

[4] The OpenRouter data cover 1,125 trillion tokens, 7.3 million users and 116 billion requests, from January 2024 to August 2026.

[5] See, for example, Tokens per day, a website that aggregates several sources to estimate developments in the total language-model market.

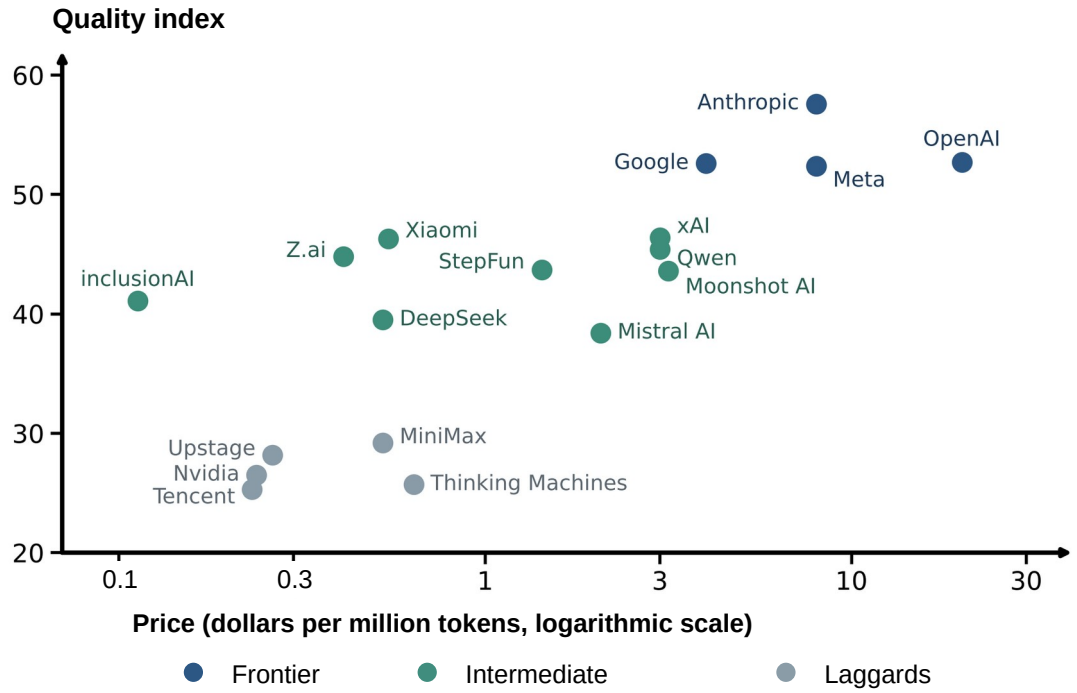
Figure 2. Weekly token consumption on OpenRouter



Notes: Tokens consumed each week on OpenRouter, averaged over the current week and the three preceding weeks, from January 2024 to the week of 10 August 2026. Triangles mark weeks in which a frontier-pushing model is released, as defined in Figure 1. Source: OpenRouter, Artificial Analysis, LIFT calculations.

Plotting the best models from the eighteen main providers by price and quality reveals three groups (Figure 3). A small number of firms - OpenAI, Anthropic, Google and Meta - regularly launch new frontier models and charge high prices. Intermediate providers, including DeepSeek and Moonshot AI, approach recently achieved performance levels at much lower prices. Laggards, finally, remain further from the frontier.

Figure 3. **Price and quality of each provider's best model**



Notes: Each point represents the best model from one of the eighteen main providers on OpenRouter at the end of the study period. Quality is measured using the Artificial Analysis performance index; prices are in dollars per million tokens. Source: OpenRouter, Artificial Analysis, LIFT calculations.

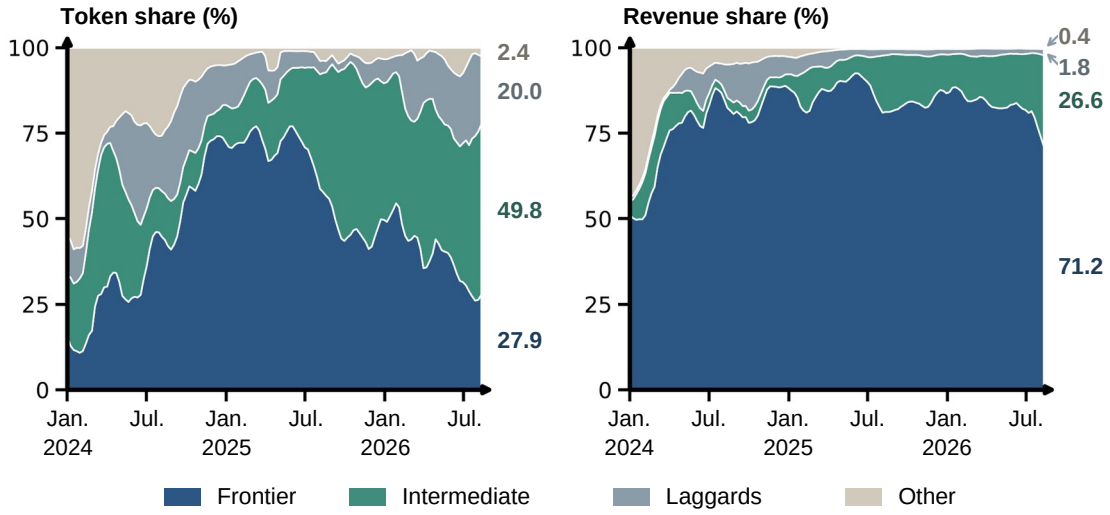
Reallocation of volumes and revenue

Volumes have shifted markedly towards cheaper providers, as shown by the weekly distribution of tokens across the three groups (Figure 4, left). Frontier labs processed three-quarters of tokens in spring 2025, but just 28% in August 2026, with intermediate providers now accounting for half of volumes. Several of these providers, including DeepSeek, Qwen and Z.ai, release open-weight models, whose rise was documented in Policy Paper 2026-03.

Revenue has shifted much less (Figure 4, right). Frontier labs charge substantially higher prices and earned more than 80% of revenue from January 2025 to June 2026, despite the erosion of their share of volumes. A frontier provider can therefore lose token market share while still seeing its revenue grow rapidly. Its revenue share nevertheless fell from 82% to 71% between mid-July and mid-August 2026, to the benefit of intermediate providers.

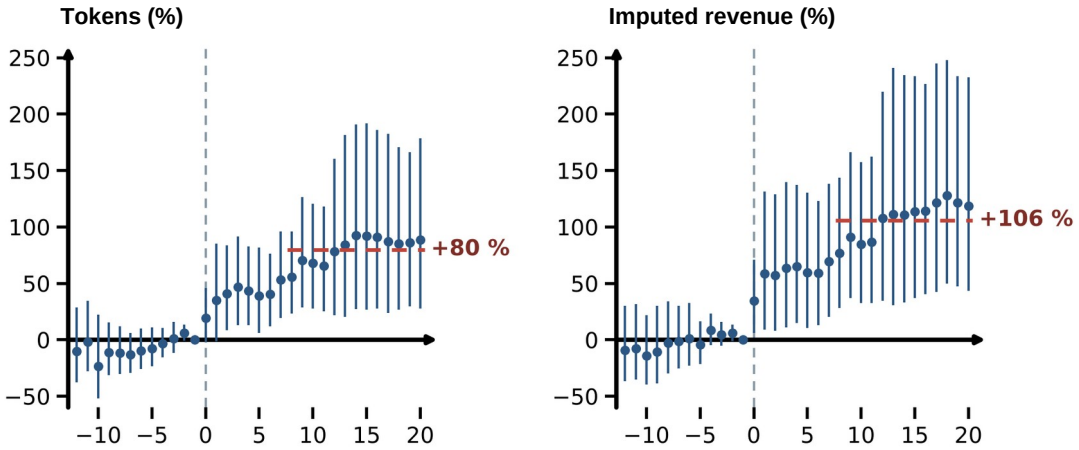
To measure these two developments, we track the activity of providers or models affected by an event and compare it, before and after, with that of an unaffected group, using an event-study method. Figures 5 and 6 report this gap for each week, relative to the gap observed just before the event. In the absence of a significant prior difference, the two groups can be regarded as comparable. When a provider releases a frontier-pushing model, its token consumption exceeds that of providers that do not by around 80%, and its imputed revenue becomes roughly twice as high, between the eighth and twentieth weeks after release (Figure 5). When a cheaper alternative of comparable quality enters a segment, usage of the exposed models gradually declines relative to that of higher-quality models, which escape this pressure, by around 30% between the eighth and sixteenth weeks (Figure 6). Catch-up offerings thus divert part of the demand and revenue from established products, a "business-stealing" effect.

Figure 4. **Token and revenue shares of the three provider groups**



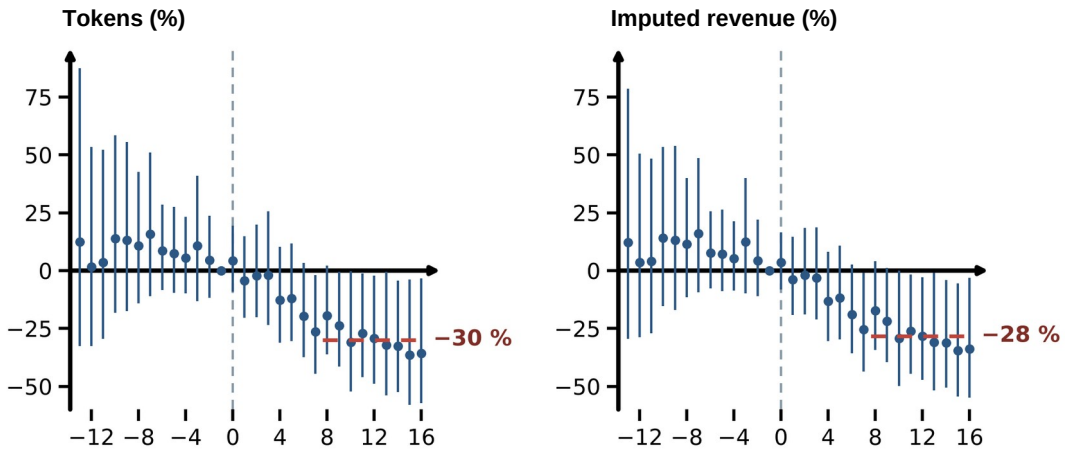
Notes: Share of each provider group in tokens consumed (left) and revenue (right) on OpenRouter, in %. The groups are those in Figure 3; “Other” includes unclassified providers. Revenue is imputed by valuing tokens at each model’s listed price. Average over the current week and the three preceding weeks, from January 2024 to the week of 10 August 2026. Source: OpenRouter, Artificial Analysis, LIFT calculations.

Figure 5. **Effect of a frontier-model release on its provider**



Notes: Estimated gap, in %, between the activity of a provider releasing a frontier-pushing model and that of providers that do not, in tokens (left) and imputed revenue (right), by weeks since release (week 0, vertical line). The week before release is the reference period. Vertical bars show 95% confidence intervals. The red dashed line shows the average gap from the eighth to the twentieth week. Coefficients, estimated on log activity, are converted into percentages. Source: OpenRouter, Artificial Analysis, LIFT calculations.

Figure 6. **Effect of a cheaper model's arrival on exposed models**

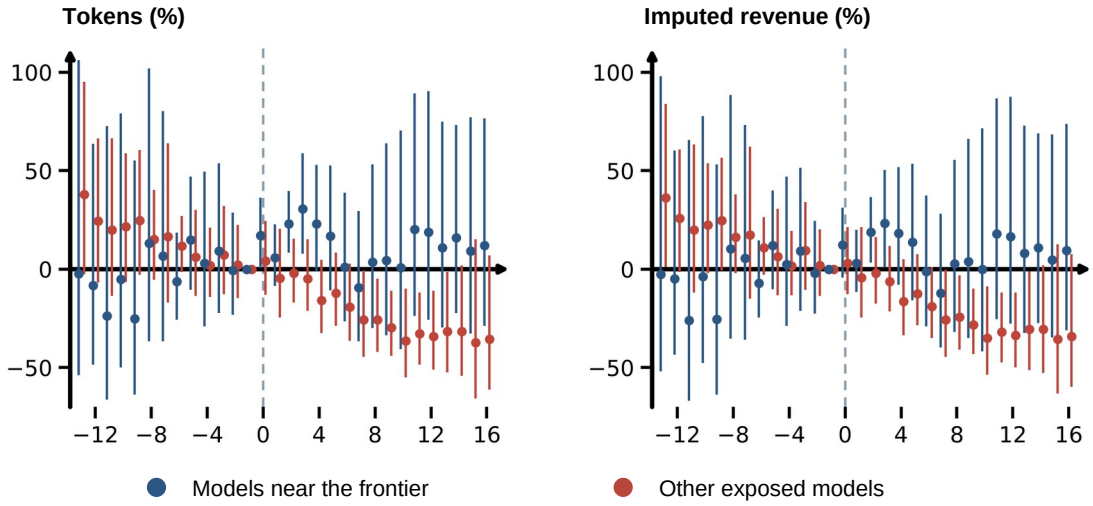


Notes: Estimated gap, in %, between the activity of models of similar quality to a new, cheaper model and that of higher-quality models, in tokens (left) and imputed revenue (right), by weeks since this competitor's arrival (week 0, vertical line). The red dashed line shows the average gap from the eighth to the sixteenth week. Other conventions are as in Figure 5. Source: OpenRouter, Artificial Analysis, LIFT calculations.

Demand for frontier models

The size of this loss depends on the exposed model's position relative to the frontier. Figure 7 repeats the previous estimation, separating models near the frontier - those whose quality reaches at least 90% of the best model available before the competitor's arrival - from the others, with both groups compared with the same reference group. For models further from the frontier, the appearance of a cheaper alternative is associated with a decline of around one-third in tokens and imputed revenue. Models closer to the frontier do not decline: their estimated gaps are slightly positive in the first few weeks and then become statistically insignificant. They therefore appear more resilient, without being immune to competition.

Figure 7. **Effect of a cheaper model's arrival, by distance to the frontier**

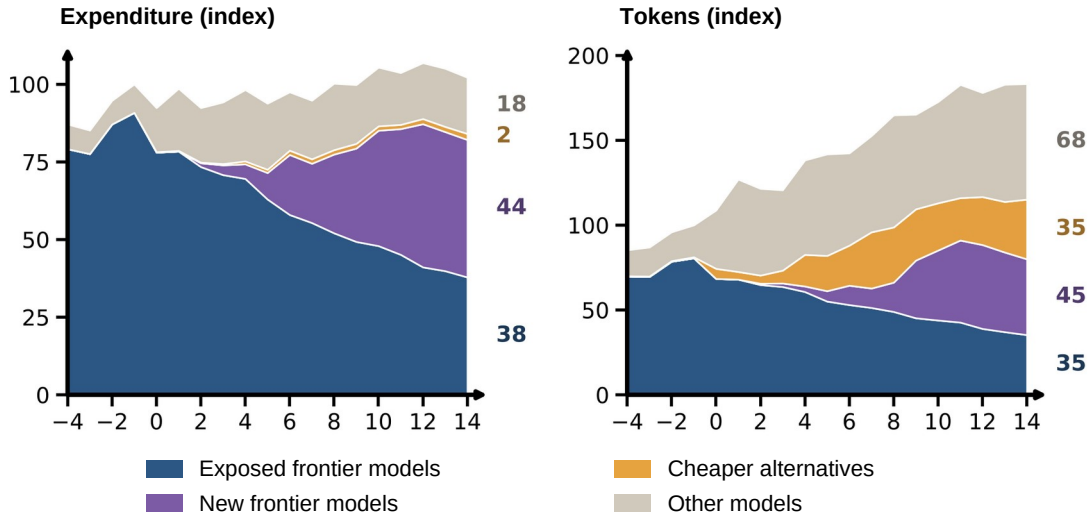


Notes: Estimated gap, in %, between the activity of models exposed to a cheaper alternative of comparable quality and that of the same set of unexposed models, in tokens (left) and imputed revenue (right), by weeks since the competitor's arrival (week 0, vertical line). Models near the frontier have a performance score of at least 90% of the best score available before the competitor's arrival. Other conventions are as in Figure 5. Models further from the frontier were already losing ground before the competitor arrived, so part of their decline extends an earlier trend. Source: OpenRouter, Artificial Analysis, LIFT calculations.

To explain this difference, we track a fixed group of users who used frontier models before a cheaper option arrived. Their expenditure and tokens are divided between the systems they were using, new frontier models, cheaper alternatives and the rest of the available models. Figure 8 expresses these amounts as indices, with 100 corresponding to the group's total in the week before the competitor's arrival. This shows whether a model's decline reflects a move away from the high end or a shift towards other highly capable products.

Expenditure on the systems initially used falls, but shifts towards new frontier models (Figure 8). Fourteen weeks after the competitor's arrival, cheaper alternatives process the equivalent of 35% of the tokens initially consumed, but receive less than 2% of initial expenditure. Spending on frontier models as a whole thus remains close to its previous level. Changes in the models used therefore do not necessarily reduce demand for the most advanced capabilities.

Figure 8. Expenditure and tokens of frontier-model users after a cheaper competitor arrives



Notes: Fixed group of frontier-model users, tracked before and after a cheaper competitor arrives (week 0). Expenditure (left) and tokens (right) for this group are divided between the frontier models initially used, which are therefore exposed to competition, new frontier models, cheaper alternatives and other models. The total in the week before the competitor's arrival is indexed to 100. Source: OpenRouter, Artificial Analysis, LIFT calculations.

These results are consistent with the existence of uses for which performance gains remain decisive, even when cheaper alternatives are available. The data cannot distinguish between technical requirements, reliability and habits, but suggest that frontier providers retain users willing to pay for the best capabilities available. This demand supports revenue from frontier innovation, provided that new models remain capable enough to retain these users.

A fragile equilibrium

Taken together, these findings describe a process of creative destruction in a rapidly expanding market. Low-cost followers erode pioneers' rents by taking customers from them, but this erosion has so far been offset by growth in consumption, with revenue remaining concentrated at the high end. When a frontier model faces competition, its users largely switch to other leading-edge offerings, allowing the labs producing them to continue growing in absolute terms. This pattern is consistent with the "escape-competition" effect described by step-by-step innovation theory, whereby the threat of imitation pushes a leader closely pursued by competitors to innovate in order to regain its lead (Aghion et al., 2001; Aghion et al., 2005). It only holds, however, if the market grows faster than frontier labs' market share declines.

The sustainability of this equilibrium depends first on AI adoption. If demand growth slows, losses in market share will translate directly into lower revenue and innovation rents. It also depends on frontier labs' ability to improve their systems regularly. If progress based on current large language model (LLM) architectures becomes more difficult, their quality advantage over cheaper competitors may no longer be sufficient to justify their prices.

For now, frontier innovation is therefore withstanding the rapid catch-up characteristic of AI. These initial trends do not establish what will happen over the longer term, as investment may yield diminishing returns and the quality gap between the best models and low-cost alternatives could narrow. Two indicators can be used to track this equilibrium: frontier labs' revenue share, which has been declining since mid-July 2026, and the time it takes the cheapest models to match the quality of the best ones, currently less than a year. If these threats materialised, frontier labs' revenue would no longer justify investment on this scale, lending weight to concerns about an AI funding "bubble".

References

Aghion, P., U. Akcigit and P. Howitt (2014). "What Do We Learn from Schumpeterian Growth Theory?" In P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*, vol. 2, 515-563. Elsevier.

Aghion, P., N. Bloom, R. Blundell, R. Griffith and P. Howitt (2005). "Competition and Innovation: An Inverted-U Relationship". *Quarterly Journal of Economics*, 120(2), 701-728.

Aghion, P., C. Harris, P. Howitt and J. Vickers (2001). "Competition, Imitation and Growth with Step-by-Step Innovation". *Review of Economic Studies*, 68(3), 467-492.

Aghion, P., P. Howitt and S. Prantl (2015). "Patent Rights, Product Market Reforms, and Innovation". *Journal of Economic Growth*, 20(3), 223-262.

Artificial Analysis (2026). Comparative assessments of AI model quality and prices. https://artificialanalysis.ai

Azoulay, P., J. L. Krieger and A. Nagaraj (2024). "Old Moats for New Models: Openness, Control, and Competition in Generative AI". NBER Working Paper, National Bureau of Economic Research.

Bergeaud, A., P. Delbouve and E. Grass (2026). "L'augmentation des factures d'IA et la montée de l'open weight" [Rising AI Bills and the Rise of Open-Weight Models]. *LIFT Policy Paper 2026-03*, HEC Paris.

Bloom, N., M. Schankerman and J. Van Reenen (2013). "Identifying Technology Spillovers and Product Market Rivalry". *Econometrica*, 81(4), 1347-1393.

Cottier, B., R. Rahman, L. Fattorini, N. Maslej, T. Besiroglu and D. Owen (2024). "The Rising Costs of Training Frontier AI Models". arXiv:2405.21015.

Guo, D., D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu et al. (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". arXiv:2501.12948.

Hinton, G., O. Vinyals and J. Dean (2015). "Distilling the Knowledge in a Neural Network". arXiv:1503.02531.

Mukherjee, S., A. Mitra, G. Jawahar, S. Agarwal, H. Palangi and A. Awadallah (2023). "Orca: Progressive Learning from Complex Explanation Traces of GPT-4". arXiv:2306.02707.

OpenRouter (2026). Platform usage data, January 2024 to August 2026. https://openrouter.ai

Sajadieh, S., L. Fattorini, R. Perrault, Y. Gil, V. Parli, L. Santarlasci, J. Pava et al. (2026). "The AI Index 2026 Annual Report". Stanford Institute for Human-Centered Artificial Intelligence.

Tokens per day (2026). Estimates of the total language-model market based on several sources. https://tokensperday.com

Author

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

The author thanks Justin Summerville of OpenRouter for access to the data, and Pierre Clarke for his review and help with the writing.

About LIFT

The Laboratoire de l'Innovation et du Futur du Travail (LIFT), or Laboratory for Innovation and the Future of Work, is an HEC Paris research lab dedicated to analysing the impact of artificial intelligence on firms, employment, innovation and growth. It brings together academic researchers and practitioners around six research areas, produces original data, and regularly publishes policy notes, studies and indicators for public- and private-sector decision-makers.

LIFT publications

Policy Papers present empirical findings and their implications for public and private decision-making in a short format. Working Papers present research in progress. LIFT trackers provide access to up-to-date indicators on the diffusion of AI. The views expressed are solely those of the authors and do not represent the positions of their employers or institutions, or those of the laboratory's partners.

Suggested citation

Delbouve, P. (2026), "Innovation and Catch-up in the AI Model Market", LIFT Policy Paper 2026-04, HEC Paris.

Contact

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

Published under the Creative Commons CC BY 4.0 licence. Reproduction is permitted with attribution to the source.

