



POLICY PAPER 2026-04 · OCTOBRE 2026

Innovation et rattrapage sur le marché des modèles d'IA

Une analyse des données d'OpenRouter, de janvier 2024 à août 2026

Paul Delbouve

EN BREF

- Dans l'IA, le meilleur modèle vendu 0,1 dollar par million de tokens ou moins, soit au moins cinquante fois moins cher qu'un modèle de frontière, rattrape la qualité de ce dernier en moins d'un an.
- Ce rattrapage pourrait décourager des investissements croissants, de l'ordre de 715 milliards de dollars en 2026 pour les principaux développeurs d'IA, en réduisant la rente que procure une avance technologique.
- Sur OpenRouter, les laboratoires de frontière ne traitent plus que 28 % des tokens en août 2026, contre les trois quarts au printemps 2025, mais perçoivent encore 71 % des revenus grâce à des tarifs plus élevés.
- Par rapport à ses concurrents, un fournisseur qui sort un modèle de frontière voit son chiffre d'affaires imputé environ doubler dans les mois qui suivent, et les utilisateurs d'un modèle concurrencé se reportent surtout vers d'autres modèles haut de gamme.
- Cet équilibre repose sur une croissance exceptionnelle de la consommation et reste fragile. La part des laboratoires de frontière dans les revenus est passée de 82 % à 71 % entre la mi-juillet et la mi-août 2026.

Des concurrents moins chers rattrapent désormais en moins d'un an la qualité des meilleurs modèles d'intelligence artificielle. Cette rapidité pourrait décourager les laboratoires qui investissent pour repousser la frontière technologique, puisque leur avantage s'érode vite. Cette note mesure, à partir des données de la plateforme OpenRouter de janvier 2024 à août 2026, ce que rapporte cette avance et à quel rythme elle se perd. Les laboratoires de frontière n'y traitent plus que 28 % des tokens¹ en août 2026, contre les trois quarts au printemps 2025, mais ils perçoivent encore 71 % des revenus. Par rapport à ses concurrents, le fournisseur qui sort un nouveau modèle de pointe voit son chiffre d'affaires environ doubler dans les mois qui suivent, et les utilisateurs d'un système concurrencé se tournent surtout vers d'autres offres haut de gamme. L'incitation à innover tient donc pour l'instant, mais elle repose sur une demande en très forte croissance.

Repousser la frontière coûte de plus en plus cher² (Cottier et al., 2024). L'entraînement et le déploiement des systèmes les plus performants nécessitent des quantités croissantes de puissance de calcul, de main-d'œuvre spécialisée, de données et d'énergie. Les principaux développeurs d'IA³ devraient ainsi consacrer 715 milliards de dollars aux centres de données, aux puces électroniques et à la conception de modèles en 2026, contre 400 milliards en 2025 (Sajadieh et al., 2026). Ces dépenses précèdent les recettes, et leur succès commercial reste incertain. Or une entreprise n'investit dans une innovation que si celle-ci peut lui procurer un rendement suffisant. Une avancée technologique doit donc lui donner un avantage temporaire, pendant lequel elle peut pratiquer des prix plus élevés, attirer de nouveaux utilisateurs ou gagner des parts de marché (Aghion et al., 2014). Cette rente est en partie ce qui incite les entreprises à s'engager dans des recherches coûteuses et risquées (Aghion et al., 2015). Si des concurrents peuvent copier immédiatement une avancée, l'avantage disparaît et le rendement espéré ne justifie plus les sommes investies (Bloom et al., 2013), ce qui ralentit l'innovation de rupture.

Dans l'intelligence artificielle, les concurrents rattrapent particulièrement vite les systèmes les plus avancés (ou « modèles de frontière »). Des fournisseurs comme DeepSeek, Qwen ou Z.ai offrent, quelques mois après leur sortie, des performances proches à des tarifs nettement inférieurs. Le prix d'usage, exprimé par million de tokens, dépasse généralement 5 dollars pour les modèles de frontière. La figure 1 compare deux trajectoires de qualité, celle du meilleur modèle disponible, quel que soit son prix, et celle du meilleur modèle vendu 0,1 dollar ou moins, soit au moins cinquante fois moins cher. La seconde rejoint la première en moins d'un an. En août 2025, par exemple, l'offre la plus performante de cette gamme de prix égale le niveau atteint par la frontière en décembre 2024. À qualité donnée, le prix baisse ainsi de plus de 95 % en moins d'un an.

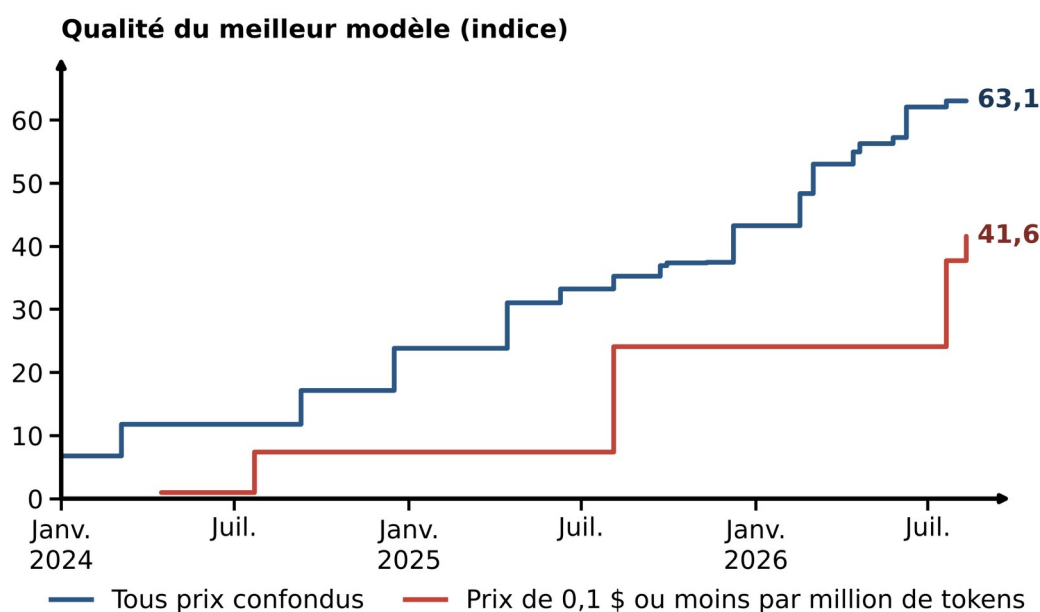
Plusieurs caractéristiques propres à l'IA peuvent expliquer cette rapidité. Les avancées circulent par les articles de recherche, les rapports techniques et les logiciels open source, si bien que les développeurs s'appuient sur des bibliothèques communes, des modèles publics et un corpus de connaissances partagé, sans refaire chaque étape de la découverte initiale (Azoulay et al., 2024). La distillation permet en outre à des modèles plus petits d'apprendre à partir des réponses ou des raisonnements de systèmes plus performants (Hinton et al., 2015 ; Mukherjee et al., 2023). Le rôle respectif de ces facteurs reste débattu, mais une fois qu'une voie technologique a été validée, s'y engager coûte nettement moins que de l'ouvrir (Guo et al., 2025). Ce rattrapage, bénéfique pour les utilisateurs qui accèdent à l'IA à moindre coût, aurait donc pu ralentir l'innovation. C'est le contraire qui se produit, dans un secteur où les investissements explosent.

¹ Un token est un fragment de texte traité par un modèle d'intelligence artificielle. On mesure l'utilisation des modèles par le nombre de tokens reçus et générés, ou consommés.

² Des chercheurs d'Epoch AI, spécialisés dans l'évaluation des coûts de l'intelligence artificielle, estiment que les coûts d'amortissement du matériel et les dépenses énergétiques liés à l'entraînement des modèles les plus avancés ont été multipliés par environ 2,4 chaque année entre 2016 et 2024.

³ Alphabet, Amazon, Meta, Microsoft et Oracle.

Figure 1. Qualité du meilleur modèle selon le prix



Notes : Qualité mesurée par l'indice de performance d'Artificial Analysis. La ligne bleue suit le meilleur modèle disponible, tous prix confondus, dont le prix dépasse généralement 5 dollars par million de tokens. La ligne rouge suit le meilleur modèle dont le prix ne dépasse pas 0,1 dollar par million de tokens. Lecture : en août 2025, ce dernier atteint une qualité de 24,1, proche de celle du meilleur modèle disponible en décembre 2024. Janvier 2024 à août 2026. Source : Artificial Analysis, OpenRouter, calculs LIFT.

Pour comprendre ce paradoxe, nous utilisons les données hebdomadaires d'OpenRouter, une plateforme qui donne accès à des centaines de modèles d'IA par une interface unique. Le Policy Paper 2026-03 du LIFT les a déjà exploitées pour montrer que la hausse des factures d'IA vient des volumes consommés et que les modèles à poids ouverts y gagnent du terrain (Bergeaud, Delbouve et Grass, 2026). Elles suivent plus de 1 000 modèles de 154 fournisseurs entre janvier 2024 et août 2026, avec leur consommation de tokens, leurs prix affichés et l'entreprise qui les produit⁴. Nous y ajoutons les évaluations indépendantes de performance publiées par Artificial Analysis. Ce suivi rapproché permet d'observer à la fois la sortie des modèles qui repoussent la frontière et l'arrivée ultérieure de concurrents moins chers aux capacités comparables, puis de mesurer l'effet de ces deux événements sur l'usage et le chiffre d'affaires des fournisseurs concernés. Ce chiffre d'affaires est imputé, c'est-à-dire calculé en valorisant les tokens consommés au prix affiché de chaque modèle. OpenRouter n'étant qu'un canal de distribution parmi d'autres, l'analyse mesure la réallocation de la demande et la pression concurrentielle sur cette plateforme, sans évaluer les revenus du secteur dans son ensemble.

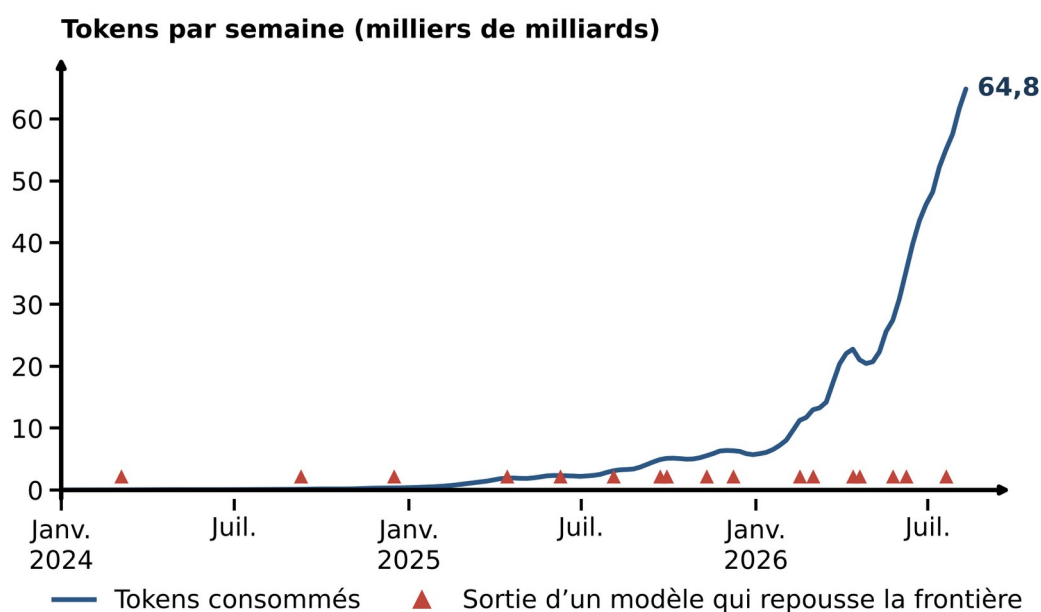
Croissance et structure du marché

La consommation hebdomadaire de tokens sur OpenRouter est passée d'environ 11 milliards en janvier 2024 à près de 65 000 milliards en août 2026. La figure 2 retrace ces volumes semaine après semaine et signale par un triangle chaque sortie d'un modèle qui dépasse le meilleur score de qualité atteint jusque-là. La progression suit une trajectoire proche d'une exponentielle, que d'autres sources relèvent aussi pour l'ensemble du marché⁵. Elle accompagne l'avancée continue de la frontière, ce qui est cohérent avec l'idée que des modèles plus capables ouvrent de nouveaux usages et élargissent l'éventail des tâches pour lesquelles l'IA devient rentable.

⁴ Les données OpenRouter portent sur 1 125 000 milliards de tokens, 7,3 millions d'utilisateurs et 116 milliards de requêtes, de janvier 2024 à août 2026.

⁵ Voir par exemple le site [Tokens per day](#), qui agrège différentes sources pour estimer l'évolution du marché total des modèles de langage.

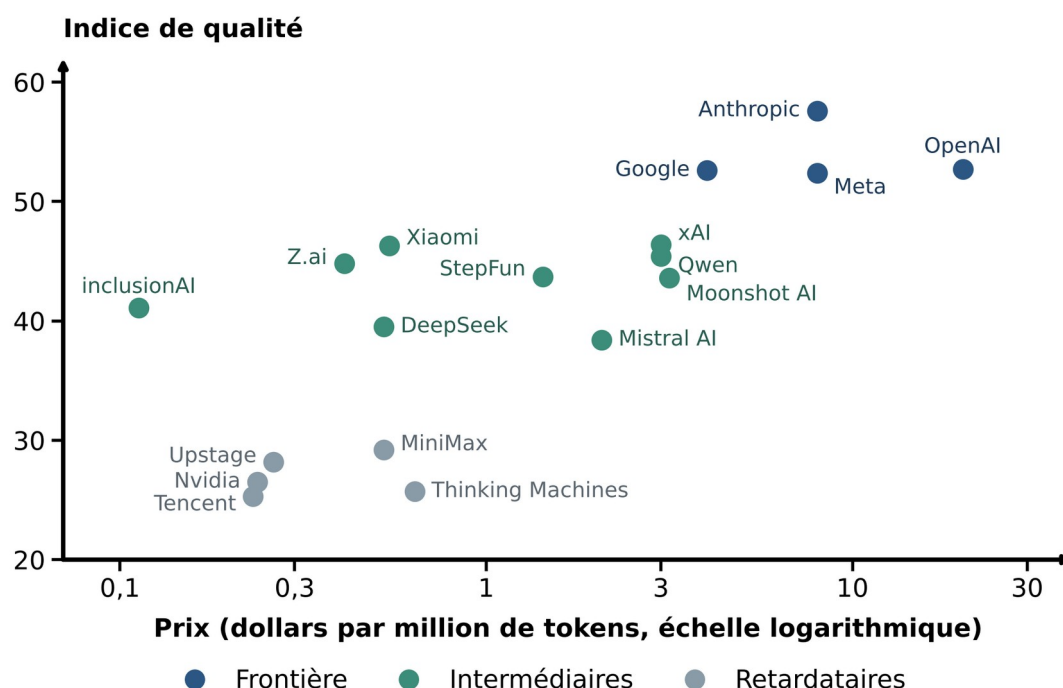
Figure 2. Consommation hebdomadaire de tokens sur OpenRouter



Notes : Tokens consommés chaque semaine sur OpenRouter, en moyenne sur la semaine courante et les trois précédentes, de janvier 2024 à la semaine du 10 août 2026. Les triangles marquent les semaines où sort un modèle qui repousse la frontière technologique, au sens de la figure 1. Source : OpenRouter, Artificial Analysis, calculs LIFT.

Placés selon leur prix et leur qualité, les meilleurs modèles des dix-huit principaux fournisseurs font apparaître trois groupes (figure 3). Un petit nombre d'entreprises (OpenAI, Anthropic, Google et Meta) lance régulièrement de nouveaux modèles de frontière et pratique des tarifs élevés. Les fournisseurs intermédiaires, dont DeepSeek ou Moonshot AI, approchent les performances récemment atteintes à des prix bien plus bas. Les retardataires, enfin, restent plus éloignés de la frontière.

Figure 3. Prix et qualité du meilleur modèle de chaque fournisseur



Notes : Chaque point représente le meilleur modèle de l'un des dix-huit principaux fournisseurs présents sur OpenRouter, à la fin de la période étudiée. La qualité est mesurée par l'indice de performance d'Artificial Analysis, le prix en dollars par million de tokens. Source : OpenRouter, Artificial Analysis, calculs LIFT.

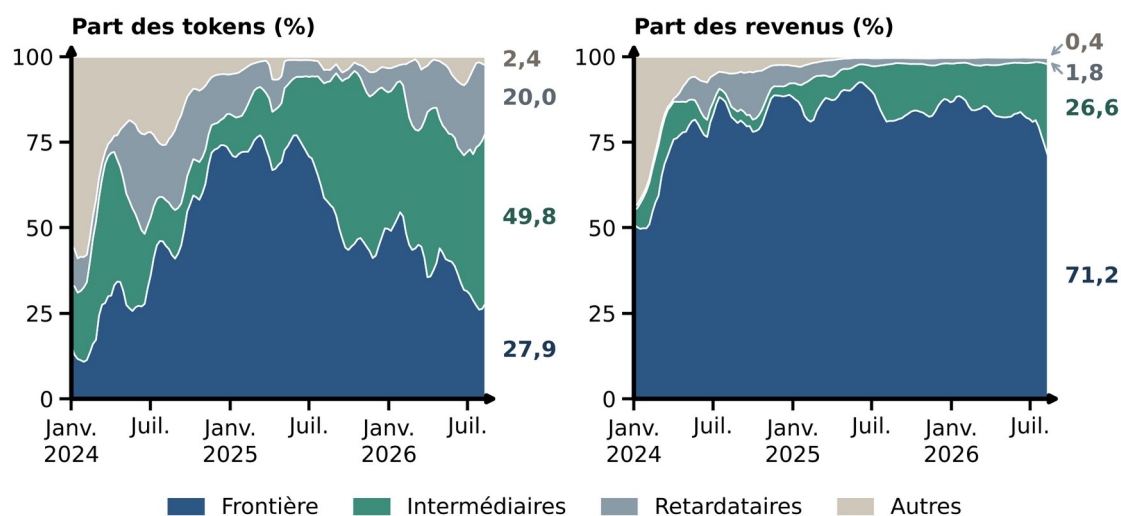
Réallocation des volumes et des revenus

Les volumes se sont nettement déplacés vers les fournisseurs moins chers, comme le montre la répartition hebdomadaire des tokens entre les trois groupes (figure 4, à gauche). Les laboratoires de frontière traitaient les trois quarts des tokens au printemps 2025 et n'en traitent plus que 28 % en août 2026, la moitié des volumes revenant désormais aux fournisseurs intermédiaires. Plusieurs d'entre eux, comme DeepSeek, Qwen ou Z.ai, publient des modèles à poids ouverts, dont le Policy Paper 2026-03 documentait la montée.

Les revenus se sont beaucoup moins déplacés (figure 4, à droite). Les laboratoires de pointe pratiquent des prix nettement plus élevés et en ont perçu plus de 80 % de janvier 2025 à juin 2026, malgré l'érosion de leur part des volumes. Un fournisseur de frontière peut ainsi perdre des parts de marché en tokens tout en voyant ses recettes croître rapidement. Sa part des revenus recule cependant de 82 % à 71 % entre la mi-juillet et la mi-août 2026, au profit des fournisseurs intermédiaires.

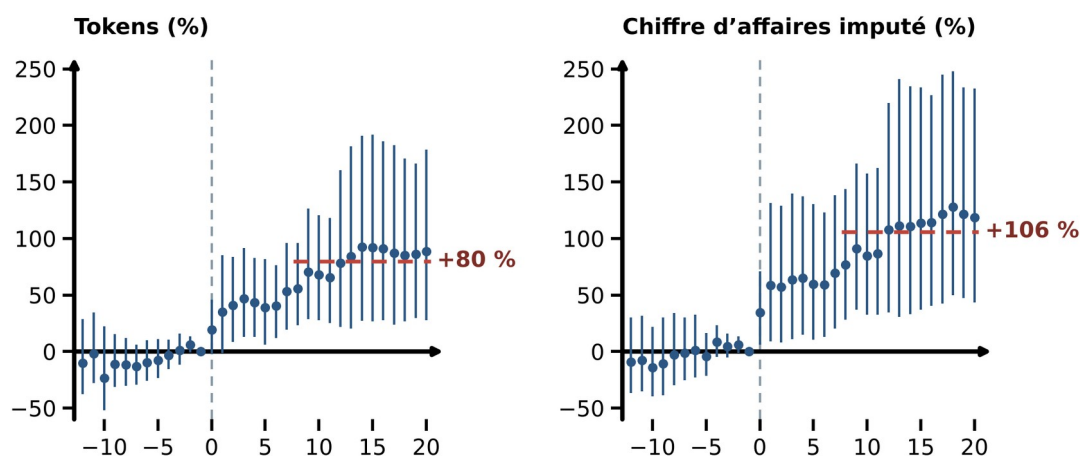
Pour mesurer ces deux mouvements, nous suivons l'activité des fournisseurs ou des modèles touchés par un événement et la comparons, avant et après, à celle d'un groupe épargné (méthode dite d'étude d'événement). Les figures 5 et 6 donnent cet écart pour chaque semaine, par rapport à celui observé juste avant l'événement. En l'absence de différence significative au préalable, les deux groupes peuvent être tenus pour comparables. Lorsqu'un fournisseur publie un modèle qui repousse la frontière, sa consommation de tokens dépasse d'environ 80 % celle des fournisseurs qui n'en publient pas, et son chiffre d'affaires imputé devient environ deux fois plus élevé que le leur, entre la huitième et la vingtième semaine après la sortie (figure 5). Lorsqu'une alternative moins chère et de qualité comparable arrive sur un segment, l'usage des modèles concurrencés recule progressivement par rapport à celui des modèles de qualité supérieure, qui échappent à cette pression, d'environ 30 % entre la huitième et la seizième semaine (figure 6). Les offres de rattrapage détournent ainsi une partie de la demande et des recettes des produits établis, un effet de captation de rente (« business stealing »).

Figure 4. Parts des trois groupes de fournisseurs dans les tokens et les revenus



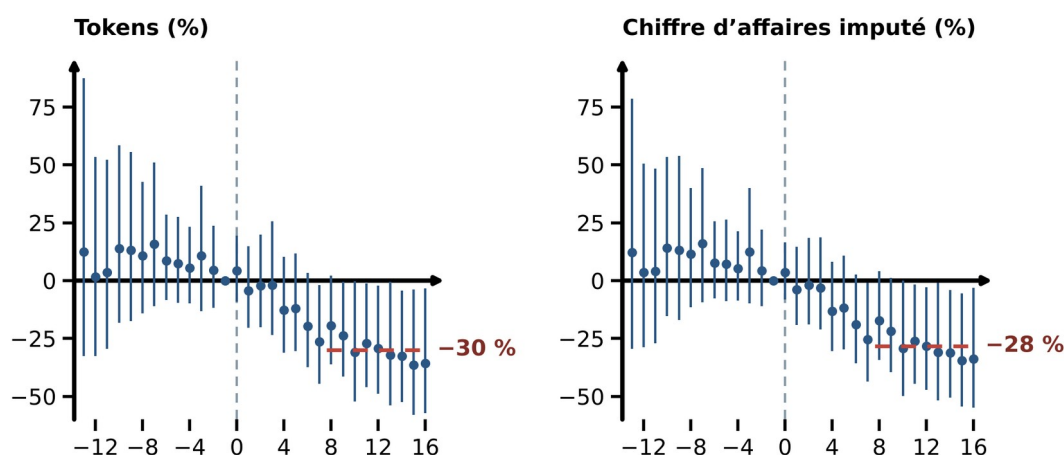
Notes : Part de chaque groupe de fournisseurs dans les tokens consommés (gauche) et dans les revenus (droite) sur OpenRouter, en %. Les groupes sont ceux de la figure 3, « Autres » regroupant les fournisseurs non classés. Revenus imputés, valorisant les tokens au prix affiché de chaque modèle. Moyenne sur la semaine courante et les trois précédentes, de janvier 2024 à la semaine du 10 août 2026. Source : OpenRouter, Artificial Analysis, calculs LIFT.

Figure 5. Effet de la sortie d'un modèle de frontière sur son fournisseur



Notes : Écart estimé, en %, entre l'activité d'un fournisseur qui publie un modèle repoussant la frontière technologique et celle des fournisseurs qui n'en publient pas, en tokens (gauche) et en chiffre d'affaires imputé (droite), selon le nombre de semaines écoulées depuis la sortie (semaine 0, ligne verticale). La semaine qui précède la sortie sert de référence. Les traits verticaux donnent les intervalles de confiance à 95 %. La ligne pointillée rouge donne l'écart moyen de la huitième à la vingtième semaine. Les coefficients, estimés sur le logarithme de l'activité, sont convertis en pourcentage. Source : OpenRouter, Artificial Analysis, calculs LIFT.

Figure 6. Effet de l'arrivée d'un modèle moins cher sur les modèles concurrencés

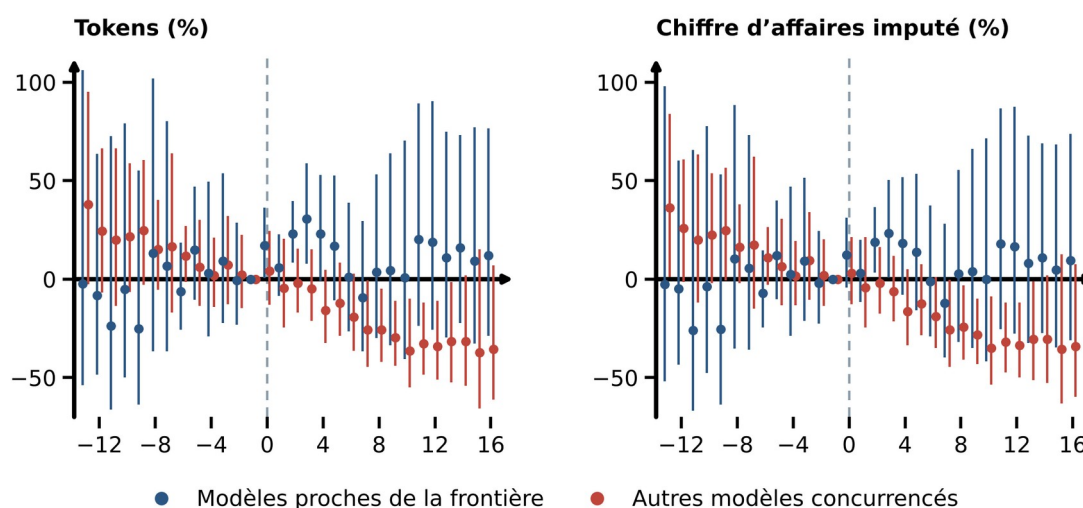


Notes : Écart estimé, en %, entre l'activité des modèles de qualité similaire à un nouveau modèle moins cher et celle des modèles de qualité supérieure, en tokens (gauche) et en chiffre d'affaires imputé (droite), selon le nombre de semaines écoulées depuis l'arrivée de ce concurrent (semaine 0, ligne verticale). La ligne pointillée rouge donne l'écart moyen de la huitième à la seizième semaine. Mêmes conventions que pour la figure 5. Source : OpenRouter, Artificial Analysis, calculs LIFT.

La demande pour les modèles de frontière

L'ampleur de cette perte dépend de la position du modèle concurrencé par rapport à la frontière. La figure 7 reprend l'estimation précédente en séparant les modèles proches de la frontière, dont la qualité atteint au moins 90 % de celle du meilleur modèle disponible avant l'arrivée du concurrent, des autres, les deux groupes étant comparés au même ensemble de référence. Pour les plus éloignés, l'apparition d'une alternative moins chère s'accompagne d'une baisse d'environ un tiers des tokens et du chiffre d'affaires imputé. Les plus proches ne reculent pas, leurs écarts estimés étant légèrement positifs dans les premières semaines puis non significatifs. Ils semblent donc mieux résister, sans être pour autant à l'abri de la concurrence.

Figure 7. Effet de l'arrivée d'un modèle moins cher, selon la distance à la frontière

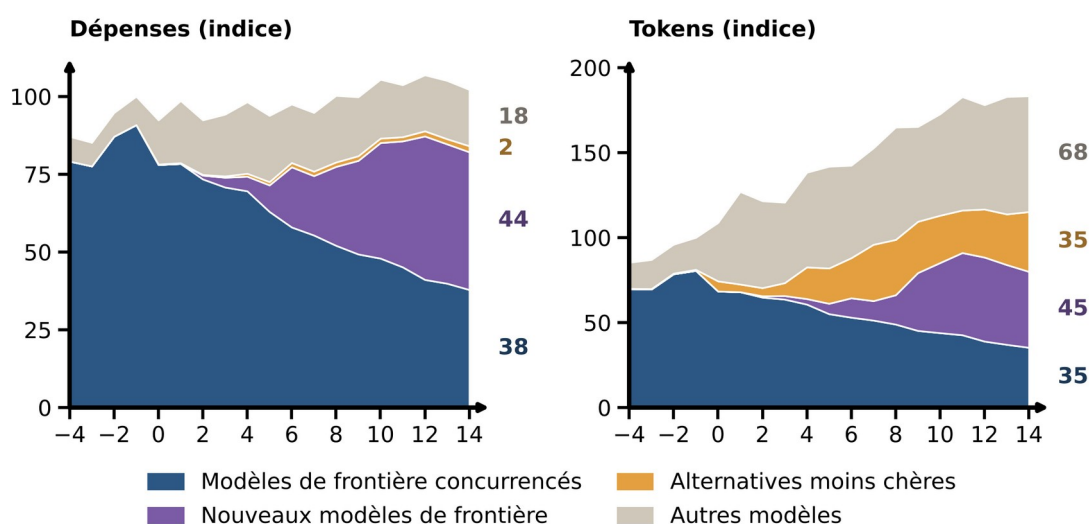


Notes : Écart estimé, en %, entre l'activité des modèles concurrencés par une alternative moins coûteuse de qualité comparable et celle d'un même ensemble de modèles non exposés, en tokens (gauche) et en chiffre d'affaires imputé (droite), selon le nombre de semaines écoulées depuis l'arrivée du concurrent (semaine 0, ligne verticale). Les modèles proches de la frontière sont ceux dont le score de performance atteint au moins 90 % du meilleur score disponible avant l'arrivée du concurrent. Mêmes conventions que pour la figure 5. Les modèles éloignés de la frontière perdaient déjà du terrain avant l'arrivée du concurrent, si bien qu'une partie de leur recul prolonge une tendance antérieure. Source : OpenRouter, Artificial Analysis, calculs LIFT.

Afin d'expliquer cette différence, nous suivons un groupe fixe d'utilisateurs qui recouraient aux modèles de frontière avant l'arrivée d'une option moins coûteuse. Leurs dépenses et leurs tokens sont répartis entre les systèmes qu'ils utilisaient, les nouveaux modèles de pointe, les alternatives moins chères et le reste de l'offre. La figure 8 exprime ces montants en indice, 100 correspondant à leur total de la semaine qui précède l'arrivée du concurrent, ce qui permet de voir si le recul d'un modèle traduit un abandon du haut de gamme ou un déplacement vers d'autres produits très performants.

Les dépenses consacrées aux systèmes initialement utilisés baissent, mais se reportent vers les nouveaux modèles de frontière (figure 8). Quatorze semaines après l'arrivée du concurrent, les alternatives moins chères traitent l'équivalent de 35 % des tokens consommés au départ, mais reçoivent moins de 2 % des dépenses initiales. Le montant consacré à l'ensemble de la frontière reste ainsi proche de son niveau antérieur. Le renouvellement de l'offre utilisée ne réduit donc pas nécessairement la demande pour les capacités les plus avancées.

Figure 8. Dépenses et tokens des utilisateurs de modèles de frontière après l'arrivée d'un concurrent moins cher



Notes : Groupe fixe d'utilisateurs des modèles de frontière, suivi avant et après l'arrivée d'un concurrent moins coûteux (semaine 0). Dépenses (gauche) et tokens (droite) de ce groupe, répartis entre les modèles de frontière initialement utilisés, et donc concurrencés, les nouveaux modèles de frontière, les alternatives moins chères et les autres modèles. Indice 100 pour le total de la semaine qui précède l'arrivée du concurrent. Source : OpenRouter, Artificial Analysis, calculs LIFT.

Ces résultats sont compatibles avec l'existence d'usages pour lesquels les gains de performance restent déterminants, même lorsque des alternatives moins chères existent. Les données ne permettent pas de distinguer ce qui relève des exigences techniques, de la fiabilité ou des habitudes, mais suggèrent que les fournisseurs de pointe conservent une clientèle prête à payer pour les meilleures capacités disponibles. Cette demande soutient les revenus de l'innovation de frontière, à condition que les nouveaux modèles restent assez performants pour la retenir.

Un équilibre fragile

L'ensemble dessine un processus de destruction créatrice sur un marché en pleine expansion. Les suiveurs à bas coût rognent les rentes des pionniers en leur prenant des clients, mais cette érosion a jusqu'ici été compensée par la croissance de la consommation, dont les recettes restent concentrées sur le haut de gamme. Lorsqu'un modèle de frontière est concurrencé, ses utilisateurs se reportent en grande partie vers d'autres offres de pointe, si bien que les laboratoires qui les produisent ont continué de croître en termes absolus. Cette configuration est cohérente avec l'effet d'« échappement à la concurrence » décrit par la théorie de l'innovation pas à pas, selon lequel la menace d'imitation pousse un leader talonné

par ses concurrents à innover pour reprendre de l'avance (Aghion et al., 2001 ; Aghion et al., 2005). Elle ne tient toutefois que si le marché croît plus vite que ne diminue la part des laboratoires de pointe.

La pérennité de cet équilibre dépend d'abord de l'adoption de l'IA. Si la croissance de la demande ralentit, les pertes de parts de marché se traduiront directement par une baisse des revenus et des rentes d'innovation. Elle dépend ensuite de la capacité des laboratoires de pointe à améliorer régulièrement leurs systèmes. Si les progrès fondés sur les architectures actuelles de grands modèles de langage (LLM) deviennent plus difficiles, leur écart de qualité avec des concurrents moins chers pourrait ne plus suffire à justifier leurs tarifs.

Pour l'instant, l'innovation de pointe résiste donc au rattrapage rapide propre à l'IA. Ces premières tendances ne disent pas ce qu'il en sera à plus long terme, alors que les investissements peuvent avoir des rendements décroissants et que l'écart de performance (« quality gap ») entre les meilleurs modèles et les alternatives bon marché pourrait se réduire. Deux indicateurs permettent de suivre cet équilibre, la part des laboratoires de pointe dans les revenus, en recul depuis la mi-juillet 2026, et le délai dans lequel les modèles les moins chers rattrapent la qualité des meilleurs, aujourd'hui inférieur à un an. Si ces menaces se concrétisaient, les recettes des laboratoires de pointe ne suffiraient plus à justifier des investissements de cette ampleur, ce qui donnerait du poids aux inquiétudes sur une « bulle » de financement de l'IA.

Références

- Aghion, P., U. Akcigit et P. Howitt (2014). « What Do We Learn from Schumpeterian Growth Theory? ». Dans P. Aghion et S. Durlauf (dir.), *Handbook of Economic Growth*, vol. 2, 515-563. Elsevier.
- Aghion, P., N. Bloom, R. Blundell, R. Griffith et P. Howitt (2005). « Competition and Innovation: An Inverted-U Relationship ». *Quarterly Journal of Economics*, 120(2), 701-728.
- Aghion, P., C. Harris, P. Howitt et J. Vickers (2001). « Competition, Imitation and Growth with Step-by-Step Innovation ». *Review of Economic Studies*, 68(3), 467-492.
- Aghion, P., P. Howitt et S. Prantl (2015). « Patent Rights, Product Market Reforms, and Innovation ». *Journal of Economic Growth*, 20(3), 223-262.
- Artificial Analysis (2026). Évaluations comparatives de la qualité et des prix des modèles d'IA. <https://artificialanalysis.ai>
- Azoulay, P., J. L. Krieger et 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 et E. Grass (2026). « L'augmentation des factures d'IA et la montée de l'open weight ». LIFT Policy Paper 2026-03, HEC Paris.
- Bloom, N., M. Schankerman et 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 et 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 et J. Dean (2015). « Distilling the Knowledge in a Neural Network ». arXiv:1503.02531.
- Mukherjee, S., A. Mitra, G. Jawahar, S. Agarwal, H. Palangi et A. Awadallah (2023). « Orca: Progressive Learning from Complex Explanation Traces of GPT-4 ». arXiv:2306.02707.
- OpenRouter (2026). Données d'usage de la plateforme, janvier 2024 à août 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). Estimation du marché total des modèles de langage à partir de plusieurs sources. <https://tokensperday.com>

Auteurs

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

L'auteur remercie Justin Summerville, d'OpenRouter, pour l'accès aux données, et Pierre Clarke pour sa relecture et son aide à la rédaction.

À propos du LIFT

Le Laboratoire de l'Innovation et du Futur du Travail (LIFT) est un laboratoire de recherche d'HEC Paris consacré à l'analyse de l'impact de l'intelligence artificielle sur les entreprises, l'emploi, l'innovation et la croissance. Il associe des chercheurs académiques et des praticiens autour de six axes de travail, produit des données originales et publie régulièrement notes, études et indicateurs à destination des décideurs publics et privés.

Les publications du LIFT

Les Policy Papers présentent, dans un format court, des résultats empiriques et leurs implications pour la décision publique et privée. Les Working Papers présentent des travaux de recherche en cours. Les trackers du LIFT donnent accès à des indicateurs actualisés sur la diffusion de l'IA. Les opinions exprimées n'engagent que leurs auteurs et ne représentent pas la position de leurs employeurs et institutions, ni celle des partenaires du laboratoire.

Citer ce document

Delbouve, P. (2026), « Innovation et rattrapage sur le marché des modèles d'IA », 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

Document publié sous licence Creative Commons CC BY 4.0. Reproduction autorisée avec mention de la source.

LIFT

LABORATOIRE DE L'INNOVATION
ET DU FUTUR DU TRAVAIL

